GLOBAL CHANGE

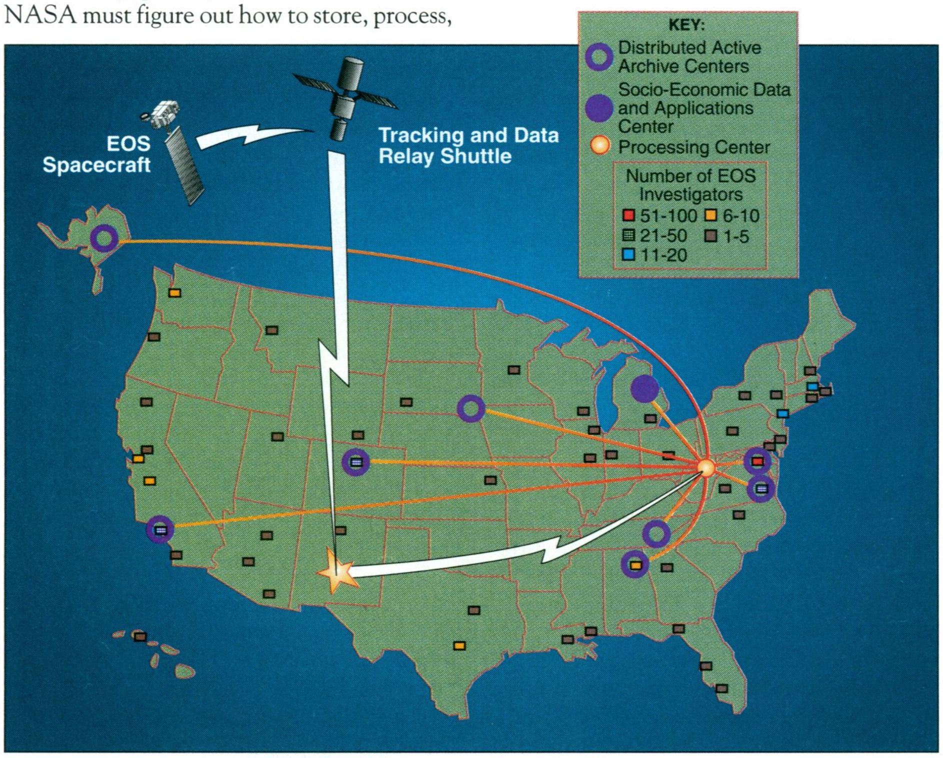
Fitting Planet Earth Into a User-Friendly Database

Shortly after the turn of the century, an array of satellites will be observing Earth, gathering data on everything from ocean color and ozone concentration to ice flows and vegetation. Each day, this fleet will beam down more than a terabyte (trillion bytes) of data—the equivalent of an entire Library of Congress every 10 days. To handle this torrent of information, the National Aeronautics and Space Administration (NASA) is planning to spend $2 billion to $3 billion for the largest and most complex scientific database ever constructed, the Earth Observing System Data and Information System (EOSDIS). "Only the hypothetical Star Wars was bigger, in terms of a large-scale distributed computing system," says Ethan Schreier, an x-ray astronomer and associate director for operations of the Space Telescope Science Institute in Baltimore.

The Star Wars computer system was never built, of course, but EOSDIS appears to have a brighter future—if it can overcome a series of hurdles. Some are political, as NASA struggles against strict budgetary limits that have already shrunk the original $17 billion plan for its system of Earth-observing satellites to less than half that size (*Science*, 12 February, p. 912). Others are technical. NASA must figure out how to store, process, and distribute data in amounts larger than anything attempted before, to as wide a community of users as possible. And that technical challenge is forcing NASA into a balancing act. To meet it, NASA researchers are scrambling to develop a user-friendly system to find and retrieve data quickly, but other researchers warn against getting locked into hardware and software that could become obsolete before they are built.

Assuming the problems can be solved, the data available through EOSDIS promise to revolutionize global environmental research. And the data system itself could have a lasting impact on fields outside Earth science. Computer wizards hope the project will develop new tools for manipulating scientific imagery, and they foresee spinoffs in fields such as crystallography and computer-aided design, which rely on databases of three-dimensional structures. Film and news video archives could benefit, as well.

NASA has little time to get all this going, as the first of eight planned satellites in its Earth Observing System (EOS) is scheduled for launch in 1998, with others spread over the following 5 years. The first steps have already been taken: NASA has signed a contract with Hughes Applied Information Systems to build a core system, installed a prototype data management network that will go public next year (version 0), and drafted a list of system requirements to be made public in September.


KEY:
○ Distributed Active Archive Centers
● Socio-Economic Data and Applications Center
○ Processing Center
Number of EOS Investigators
🟥 51-100
🟧 6-10
▨ 21-50
⬛ 1-5
🟦 11-20
EOS Spacecraft
Tracking and Data Relay Shuttle

SOURCE: NASA ILLUSTRATION: DAN REBEIZ

Shower of data. Beginning in 1998, sensors will relay images and measurements through a ground station to a processing center and archives around the country.

While it will take years for NASA to fill in the details, it has already laid out a broad plan for channeling the data from orbit to archive. Signals from the EOS sensors will travel first to a NASA relay satellite and then to a ground station at White Sands, New Mexico. From there, the raw data will flow to a processing center in West Virginia (a site favored by Robert Byrd (D–WV), the chairman of the Senate appropriations committee), where they will be stored, copied, and processed to remove obvious errors.

Next the data will go into a third element of the system, shunted along internal EOSDIS networks to at least eight "distributed active archive centers," or DAACs, around the United States. NASA will establish a single format for the data, but each DAAC will enjoy some autonomy and scientific specialization. For example, the DAAC located at the University of Alaska in Fairbanks will take the lead on synthetic aperture radar imagery of ice, snow, and sea surfaces. NASA's Goddard Space Flight Center in Maryland will house a big DAAC concentrating on climate, meteorology, the stratosphere, ocean biology, and geophysics. And the Earth Resources Observation System center, located in Sioux Falls, South Dakota and run by the U.S. Geological Survey, will take the lead on managing EOS land-based information.

All of these libraries will also include major existing collections of Earth science information. The Sioux Falls center already houses Landsat images, and the DAACs will also include data ranging from air surveillance images gathered in 1939 all the way to data from existing satellites, including the European Community's ERS-1 (see box). More than 250 databases are tagged for inclusion. EOSDIS managers are now processing these files into a format compatible with EOSDIS.

Customer service. A final link in the data chain will carry information from the DAACs to the work stations of individual scientists. Some researchers will receive funding to help process and distribute the information in "standard data products"—a time series of carefully documented polar ozone observations, for example. This part of the network, most likely connected to the Internet, will also be accessible to researchers not in the EOS core group.

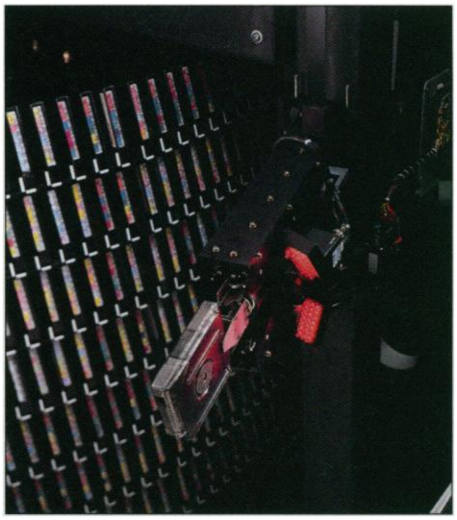
The physical task of processing and shipping all this information will be enormous. By 2003, the flow is expected to reach a rate of 1 to 2 terabytes a day, says Robert Price, chief of EOS projects at Goddard. "This is a massive amount of data. We can't fall behind or we'll never catch up," he says. "Our requirement is to process a day's worth of data in a day," and release it in 24 hours.

But Price doesn't see this as the toughest job. "Having the ability to search through all that data and find what the requester is asking for, and bringing it back for him to view, that's going to be the biggest challenge," says Price. According to NASA's plan, a scientist should be able to tap into the entire EOSDIS database by 1998 at any of the eight DAACs and call up complex data sets using plain English commands. NASA is planning for a client base of about 10,000, and its goal is to have the system respond to a query from any one of 100 simultaneous users within seconds. That's a tall order, and to understand why, it helps to know something about current methods of storing and retrieving information.

Most big data systems in use today were built for business users, says Michael Stonebraker, a computer scientist at the University of California, Berkeley. That limits their value for Earth science data, as Stonebraker is learning in an experimental Earth science data project called Sequoia 2000, which he co-directs with Jeffrey Dozier of the University of California, Santa Barbara. On a budget of $14 million from a consortium led by the Digital Electronics Corp., Sequoia 2000 is tackling problems of the kind EOSDIS faces, including management of complex data from Earth observing satellites.

Stonebraker refers to the smallest bit of retrievable Earth science data as a "grain." It differs from a business record in having more dimensions—latitude, longitude, time, and spectral value, among others. Business data records are simpler and more discrete, and the programs devised to handle such records (relational systems) are limited in their ability to target and fetch multidimensional information. When such software is used on complex data systems, Stonebraker says, it tends to "lay an egg," failing to target useful information or clogging the system with redundant search processes. On the other hand, relational systems are the most thoroughly developed, most widely available, and the safest buy for an institution that doesn't want to take a risk. Nonetheless, Stonebraker argues that it would be far better for EOSDIS to shop for novel types of software capable of identifying the content of multidimensional images.

In the past, Earth scientists have dealt with the weaknesses of search software by educating themselves on how the system stores grains of data and preparing detailed queries. If carefully targeted, a request will make a hit—say, finding images of Mt. Pinatubo during an eruption. But this hit-or-miss approach will not work well with EOSDIS, because the volume of data to be searched is enormous. Besides, the old technique is hostile to newcomers who aren't intimately familiar with the system.



STORAGE TECHNOLOGY

Cool hand. Robots like this Storage Technology device, which manipulates a library of 6000 tape cartridges, are likely to be used to retrieve data in EOSDIS archives.

NASA wants software that will allow any researcher to submit plain-language requests such as: "Find Mt. Pinatubo during an eruption." One way to make that possible, says Price, would be to label each bit of data by content as it enters the system. "We would like to make machines smart enough...to identify the features within the data on the fly" and automatically create a catalogue of descriptive labels, he says. Price dreams of a robot skimming along the rivers of EOS data, tagging each grain with a note: "This scene contains a volcano, or this scene contains a cornfield." But at present this is only a dream, as Price readily concedes: "The technology does not exist today."

How, then, will NASA devise a user-friendly EOSDIS? The answer isn't clear. The software now being used to link up the DAACs, version 0, will not solve the problem. It provides abstract descriptions of all data sets at the DAACs and already can retrieve imagery on demand from some of them. It's working "pretty well," says Dixon Butler, who runs Earth science programs at NASA headquarters. H.K. Ramapriyan, an EOSDIS manager at Goddard, thinks it's "better than we expected." But it is essentially, Ramapriyan says, "an early prototyping effort...a good sociological exercise," designed to get the system launched while Hughes comes up with a permanent solution.

A question of balance. A key issue in that search is philosophical. Should EOSDIS play it safe, investing only in "mature" technologies and designs, or should it stress innovation and try to promote entirely new ways of handling data—even at the risk of making mistakes?

So far, NASA's managers have opted for the cautious approach. The agency has been criticized in the past for launching grandiose projects it couldn't finish. And NASA administrator Dan Goldin has given EOSDIS a "yellow light," meaning it must keep a tight lid on costs and meet hard deadlines. For example, EOSDIS chiefs have promised to have the version 0 prototype of the data network ready for general use by June 1994. They plan to overlay a more sophisticated version 1 in a "seamless" fashion by 1997. Such requirements tend to put a damper on experimentation. "It's hard in government procurement to allow things to change, to evolve," notes Schreier, who sits on a 13-member National Research Council panel, chaired by Charles Zraket of the Mitre Corporation. Adds Dozier, "We've got a design that's fundamentally 3 years old already, and we're sort of proceeding with it."

Yet Earth and computer scientists warn that a strategy that's too safe could be just as risky as one that's too adventurous. Unless EOSDIS creates and adopts new ideas, they foresee it becoming a white elephant. By the time the influx of new Earth science data reaches a peak in 2002-2005, the technology of 1993 will be obsolete. The computer experts say that the best way to avoid sclerosis in the system is to plan a system architecture that encourages continuous innovation.

NASA says it is trying. In March, the agency signed a $766 million contract with Hughes through the year 2002 to develop EOSDIS based on an "open architecture" concept. Unlike some earlier projects, it won't lock into a single, proprietary format. Ramapriyan says the plan is to "build a little, test a little."

This approach reflects the advice of NASA's technical advisers—including the Zraket panel. Like Dozier and Stonebraker, the Zraket panel has urged NASA to devote a chunk of its budget to R&D on computer science and data handling. NASA has taken the advice to heart, and Butler at NASA headquarters says the agency is committed to spending about $20 million on this kind of research over the next few years. Recently, NASA solicited proposals for advanced computer technology work, and another announcement is pending. This will go on "in parallel" to the routine operations, says Price, and new ideas "will be tested...and brought into the operation."

Several panels will be reviewing the EOSDIS strategy this fall—including the Zraket group, which expects to give its judgment in October. In December, NASA and Hughes will review the design and solidify the system architecture. But the final test of the concept won't come until 5 years from now, when the new data begin to flow and researchers start making demands on the system. And by then, of course, it could be too late for second thoughts.

–Eliot Marshall