Data rescue fills in the climate record

Scientists around the world have moved aggressively in recent years to identify and 'rescue' valuable stores of data — particularly atmospheric and oceanographic records extending back a century or more — before they are lost to the ravages of time.

The records range from handwritten nineteenth-century ships' logs, to weather observations from colonial Africa, to decaying magnetic tape from early weather satellites. The treasures still turn up regularly on dusty shelves and in basements, and even more could be salvaged if a greater public investment were made.

"Every few months we hear about a major archive," says Sydney Levitus of the US National Oceanographic Data Center in Maryland. As an example he cites the Scripps Institution of Oceanography's discovery of 180,000 ocean temperature profiles taken by the US Navy during the Second World War.

No one knew these data existed. But they neatly filled a gap in the North Pacific for which there had been no contemporary record. Altogether, some two million ocean temperature profiles have been added to the centre's World Ocean Database in the past five years, nearly doubling the previous amount.

Part of the credit goes to a campaign to digitize archives already known to exist. But some credit goes to the six-year-old Global Ocean Data Archaeology and Rescue (GODAR) programme, which Levitus heads and which is conducted in association with the Intergovernmental Oceanographic Commission.

GODAR partners meet annually to report on significant data archives in their countries that might be worth saving and converting to digital form for easy electronic sharing. "There's been incredible international cooperation," says Levitus. The Russian Navy has released declassified ocean data, as have the US and British navies.

Most of this historical information, which was gathered for operational purposes, "has never been touched" by scientists, says Thomas Karl, chief of the US National Climatic Data Center. More than 100,000 chlorophyll profiles and 600,000 plankton observations have been added to the database, along with measurements of salinity and ocean chemistry.

The addition of the temperature profiles has already had an important scientific impact, says Levitus. "These data taken together allow us for the first time to compute the interannual variability of the heat storage in the upper ocean," a key factor in climate-change research.

The US National Oceanographic and Atmospheric Administration has its own data rescue programme. It aims to preserve and eventually digitize millions of old paper, film and tape records stored at three data centres. The task is enormous, and the waiting list is long.

With current funding of $5 million a year, the agency reckons it will take 18 years to rescue 137 million paper records, and 42 years to salvage 450 terabytes of environmental data stored on outdated computer cartridges and disks. The programme has made great strides, says Karl, "but a lot more could be done".

Ironically, records from the computer age are among the most perishable, with storage media fast becoming outdated. Geostationary Operational Environmental Satellite (GOES) weather data from the 1970s are housed at the University of Wisconsin, where the late meteorologist Verner Suomi had the foresight to transfer them to tape.

But the tapes are deteriorating, and the number of machines that can read them is dwindling. The high-resolution GOES data could yield "potential treasure," says Karl. But the tapes remain in storage because of a lack of money to convert them.

Spurred in part by the signing of the Kyoto treaty on carbon dioxide reductions, scientific organizations around the world have grasped the importance of tracking down and preserving as much climate-related data as possible, says Levitus. "The consciousness has really been raised."

The Belgian government, working with the World Meteorological Organization, has helped African nations to digitize weather observations from the colonial era. Similar work is under way in Caribbean countries, and the European Commission is funding an effort called MEDAR to retrieve data on the Mediterranean Sea.

Funding is sparse for these ad hoc programmes. But enthusiasm is high. The Australian Oceanographic Data Centre recently advertised in a marine science newsletter in an attempt to root out hidden treasure held by individual scientists. Centre head Ben Searle estimates that 70 to 80 per cent of the marine and coastal data that has been collected in Australia "resides in filing cabinets and on personal computers, and its existence is unknown" to all but the owners.

Karl says the tremendous effort that has gone into data rescue should be a sobering lesson for designers of modern electronic databases, who can expect to have to "migrate" terabytes of data periodically to updated storage media, or risk finding themselves some day with a "dead archive". It won't be glamorous work, he says. But "we'll all have to pay some tax to do this".

data management task that they started working on it in 1995, ten years ahead of the accelerator's coming online.

The LHC's data system will push the state of the art in several areas, including high-performance storage for computer data, where capacity may be less of a problem than speed of transfer between tapes and other media. Fortunately, the commercial world is working on the same problem. LHC data managers therefore hope to buy the storage system and database software essentially off the shelf, then modify them.

The four individual LHC experiments have already made a decision to forsake tried-and-true Fortran code and switch to more versatile object-oriented programming (of which C++ and Java are common varieties). This will require "a different mindset" for CERN programmers, says Julian Bunn, a CERN physicist working at the California Institute of Technology, which is collaborating on the LHC data system.

It was a bold step, not taken lightly. NASA's Earth Observing System data system (EOSDIS) crashed on the rocks of object-oriented programming in the early 1990s, when the tools were newer. The project suffered delays and cost overruns as a result.

The LHC data system will be distributed, with a central archive at CERN and regional centres serving users nearest to them. Scientists tapping into the database, though, will have the impression of a single repository. Bunn worries most about networking — moving around enormous volumes of data, quickly and seamlessly, among ten regional centres, the central archive at CERN, and up to 2,000 individual users around the world.

A new, higher-capacity Internet should be available by then. But millions of users worldwide will be soaking up the bandwidth by downloading movies and shopping online. Even dedicated scientific networks are likely to fill up quickly, says Bunn.

The life expectancy of the archive is equally problematic. "We've never been faced with maintaining a 20- or 25-year database," says Bunn. As the LHC matures, the calibration of its instruments will become more refined, and all those petabytes of data will have to be regularly reprocessed, adding to the volume of data the system has to churn. Project managers only expect 100 petabytes from the collider itself. But they are designing their system to handle ten times more data.

What's worth keeping?

Computational demands will become greater as scientists build more linkages and capabilities into their databases. Services such as the US National Center for Biotechnology Information have already begun to merge scientific journals and databases into unified searchable libraries (see Briefing, *Nature* **397,** 195–200; 1999).