Catalogue of life could become reality

Taxonomists have long dreamt of creating a master 'catalogue of life'. For various reasons — lack of money and competing schemes for going about the job being the two most prominent — it has not yet happened. But the 1992 global biodiversity treaty may be a spur to further action.

Frank Bisby of the Centre for Plant Diversity and Systematics at the University of Reading, England, hopes that the Species 2000 project to federate as many as 200 databases into a single searchable archive of all the world's 1.7 million known species "is about to turn from a plan into a reality".

Other groups with similar ambitions have signed on as partners, including the US–Canadian Integrated Taxonomic Information System and the Global Plant Checklist based in Australia and Europe.

Today, the prototype Species 2000 'dynamic checklist' searches just three databases. But up to 30 links are expected by the end of the year. Bisby is also discussing links to the geospatial database created by the University of California at Santa Barbara's Alexandria Digital Library, so that species data could be combined with geographic information.

Funding so far has been "ridiculously fragile," he says. But he is optimistic that the Global Environment Facility and the European Union will help pay some of the estimated $140 million cost of the basic (non-georeferenced) system.

Bisby and other taxonomists have been envious of the ample funding bestowed on molecular biologists, when "we think of ourselves as being of equal stature". Chris Thompson, a US Department of Agriculture researcher and former vice-chairman of Species 2000, says: "We've just been ineffectual at selling our vision."

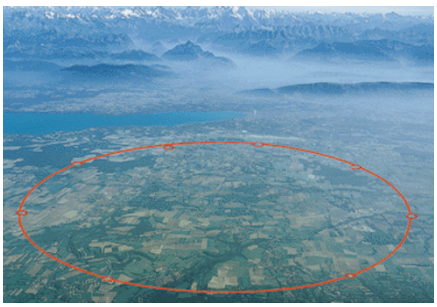
Species 2000 is at www.sp2000.org

A data workshop sponsored by the US National Science Foundation last year produced an even more ambitious vision: digital journals that would link not only to the data used in an experiment, but to the programs that created or analysed the data, so that readers could verify the results of an experiment or run their own variations. The workshop participants called this 'deep citation'.

"That scares me a bit," admits Bunn. He understands the appeal to scientists who want to recompute some controversial result for themselves. But who would be allowed access to the data and the computational resources, and on what terms? It's an idle worry today, he says, but "I guess it will come".

Uncertainty about what information to keep — a particular problem in young fields such as genomics — will be a big contributor to database bloat. Are all expressed sequence tags sent to GenBank worth keeping? And all single nucleotide polymorphisms? Before scientists have thoroughly analysed the data, no one can say. Until then, says Spengler, "you have to be a pack rat" to avoid throwing away something important.

CERN



Data factory: handling experimental results from CERN's Large Hadron Collider (above), currently under construction, has presented a daunting challenge to physicists.

It's unwise to cater to every scientist's whim about what data should be archived, says Graham Cameron, joint head of the European Bioinformatics Institute (EBI) outside Cambridge, England, which maintains, among others, the SwissProt and EMBL nucleotide sequence databases. "You could soak up any amount of money" trying to store everything, he says. Database managers have to judge what data are used most often by scientists, and what might be used in the future.

This is not easy. "Complexity fights its way in," says Cameron. The EBI recently decided to establish a public domain repository for DNA microarray-based gene expression data, despite concerns that such an archive might be premature (see *Nature* **398,** 646; 1999). Cameron calls this a "strategic" commitment, even though "technically we may not be at the stage yet to do it right".

Maynard Olson, a geneticist at the University of Washington who has been deeply involved in the Human Genome Project, thinks that the rush to produce a 'quick-and-dirty' draft of the human genome may lead to headaches later, as a mountain of low-quality data is harder to analyse than a smaller, more refined dataset.

Who will pay?

With the web firmly established as a primary avenue of scientific communication, the notion of a database as a large repository in a single location has become *passé*. Grossman, of the National Center for Data Mining at Chicago, says "the tide is shifting pretty dramatically to distributed systems," which can be loose federations of independently operated databases using common data standards and transfer protocols.

The federations may not be as efficient as centrally managed archives — "every link you build sets up a dependency," says Cameron — but they have real advantages, such as allowing specialists to keep and curate their own data.

Concern about ownership remains an obstacle to database sharing. Researchers at the University of Kansas Natural History Museum hope they have found one solution in a data retrieval protocol called Z39.50, which has proven successful in the bibliographic community. A Z39.50 query retrieves and pools data from multiple sources — perhaps museums in different locations that hold specimens from the same taxon or region. Each museum retains control of its own database, but the pooled results add up to something no single collection could offer — enough data points to allow detailed analysis of biodiversity patterns.

But the issue of data ownership will not go away easily, especially for information with perceived commercial value, and this could prevent many scientists from making the most of the current data bonanza. At least some new information on the human genome — particularly products derived from raw sequence data — will be off limits to those who do not pay private companies such as Celera for access rights (Novartis, Upjohn and other large companies have already done so).

Scientists worry that proposed changes to US intellectual property laws could push researchers to view their data as commodities to be sold rather than as information to be shared (see *Nature* **394,** 410; 1998).

Cameron at the EBI places some of the blame on stingy governments. SwissProt reluctantly began charging commercial users for access to its database only after government funding dried up. The present situation is "not ideal," he says, but it reflects the "inability of the public funding mechanism in Europe" to recognize the importance of free genomic data.

The commercialization of research databases could also shut poorer developing nations out of the scientific mainstream. But the ramifications go beyond North–South politics. So agitated did European nations become over US companies' practice of gathering free European meteorological data, then repackaging them into commercial databases sold back to Europeans, that the World Meteorological Organization passed a resolution several years ago allowing countries to restrict access to certain kinds of commercially valuable weather data.

Until that time, the information had always been shared freely among nations. "No question about it, it was a step back," says Michael Crowe, science planning officer at the US National Climatic Data Center. Data exchange and intellectual property rights are "becoming a thornier and thornier issue," he says.

The good news in today's data explosion