Making good databanks better

Like motherhood before Green considerations made it unfashionable, the scientific community's databanks enjoy the presumption of virtue. But there remain important problems to be solved.

It is splendid, and entirely consonant with the doctrine that the scientific enterprise is a communal enterprise, that data arising in the course of discovery should be generally available. Access is especially important when a claim can be checked only by reference to the original data, but it must often be that data gathered from several investigators is more valuable in a common databank than if separate parcels remain where they arise.

So why does *Nature* not make it a precondition of publication that experimental data should be submitted to the appropriate databank? The question is more often asked as databanks proliferate, and as compliance with the pleadings of the databanks becomes more common. What follows is an explanation both of this journal's continued caution and of why that diffidence does not imply dissent from the objectives of the databanks.

The sheer number of databanks now extant is part of any explanation. The days have gone when the Crystallographic Data Centre at the University of Cambridge, created in the 1930s by the International Union of Crystallography, was the chief organization of its kind, but that has been the model for several more recent creations, notably the nucelotide sequence databanks operated in concert at the European Molecular Biology Laboratory (EMBL), Heidelberg, the Los Alamos National Laboratory and at the Riken Laboratory at Tokyo in Japan. Now, largely on the initiative of Dr Fred Roberts at Yale University, an international network (centred on the Brookhaven National Laboratory) is also being formed to collect three-dimensional structure data on proteins and other structures.

Such databanks function on common principles. The submission of data is voluntary, but access is general. Organization is often a formidable undertaking, as the nucleotide sequence databanks have discovered in the past five years. Compilations of data are available to all, usually at the modest cost of the magnetic tape or compact disk that carries them. They are a considerable public service.

Other databanks function differently. Part of the legacy of international projects such as the International Geophysical Year is the network of data repositories scattered about the globe, while particular Earth satellites have led to the creation of their own data repositories whose value persists — data from the Einstein X-ray satellite, launched more than a decade ago, are still being used, for example (see page 309, this issue). Cores drilled in the ocean floor are similarly available, while it is intended that data collected by the Hubble Space Telescope will be open to all once the principal investigators have had a crack at them. These are sensible ways of making information generally available, but because these databanks are created by the voluntary acts of single projects, they pose no difficulties for researchers or for journals.

Problems of principle do arise when success rests on the willingness of individual researchers to submit their data. This is the case with the internationally integrated nucleotide sequence databank — but nothing in what follows implies that the objectives of that enterprise are anything but excellent.

One difficulty is geographical. As things are, the three collaborating centres have divided the world between them, preferably taking in data electronically from individual researchers, but otherwise undertaking to type in manually data appearing in journals. Access to this databank will soon be possible by computer. One obvious difficulty is that access is patchy. People in, say, India have less easy access than those elsewhere, and may feel doubly injured by demands that the contribution of their data to the common pool from which others will benefit most should be a precondition of publication.

The Soviet Union is in worse shape; the Soviet Academy of Sciences buys an updating tape from Heidelberg each month, but has been told that it cannot be a full partner in the collaboration for many different reasons — that full membership would entail information about the operation of computers covered by the strategic embargo, that three partners are a sufficient headache for the time being and that, in any case, the Soviet Union does not produce many nucleotide sequences. Even allowing that the task of accumulating sequences is herculean, an enterprise that seeks to be international and comprehensive could have spent more time worrying about equality of access.

Commercial difficulties also arise. Academic researchers may regard submission as the rule, but those working for commercial organizations will be more restrained. Sequences published *in extenso* will be available, but what about the rest? Commercial organizations are notoriously unwilling to let their competitors know what interests them, and may legitimately claim the right to secrecy while enjoying free access to what academics publish. That is why the databanks have been reduced to pleading with commercial companies at least to keep their sequences securely.

Still more worrying is the commercial use of nucleotide sequence data. Already there are several small companies selling packages of proprietary interpretive software together with the contents of the databanks. It is not difficult to think that there will soon be consultancy firms offering to interpret the contents of the databanks for the benefit of commercial clients. The development of such services shows a need for them, but their equitability as it affects the providers of the data and even the databanks themselves has been inadequately explored.

The three-dimensional structure databank raises another difficulty. The case for it is strong: the publication of a molecule's structure may be useless to others without more detail, atomic coordinates for example. But it may be easier (and quicker) to reach broad conclusions than to refine the atomic coordinates, making them unambiguous. So authors will enjoy a moratorium between publication and the submission of data that will no doubt also ensure that no "sharpshooting theorist" (one researcher's phrase) is the first to interpret the structure.

The moral for journals such as this is plain. Their first duty is to speed the process of publication but also to ensure (as far as possible) the integrity of what they publish, which gives them a right to ask for supporting data (sequences, atomic coordinates) even when they have no space to publish them. They should do that more often, and should also arrange to provide access to them. But journals have no right to adjudicate upon a contributor's subsequent conduct — and have few sanctions, anyway. If there must be policemen, grant-making agencies are better placed. But there is a prior need — that the unanswered questions of access and equity should be tackled energetically by some body such as the International Council of Scientific Unions. Meanwhile, *Nature* will continue to urge on its contributors the importance of submitting their data to the databanks, but will not exact unenforceable promises to do so.

John Maddox