

Science's neglected legacy

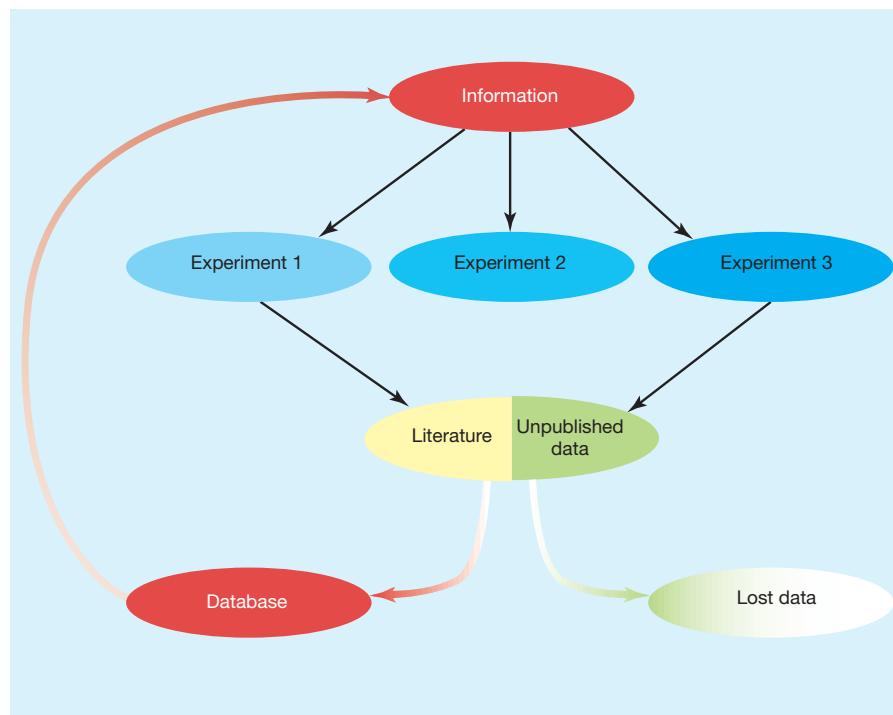
Large, sophisticated databases cannot be left to chance and improvisation.

**Stephen M. Maurer,
Richard B. Firestone
and Charles R. Scriver**

Science is an assault on ignorance. Its legacies are concepts, technologies and databases¹. As with many walks of life, the most glamorous legacies tend to get the most attention and the least are neglected. Databases have never been regarded as glamorous, only useful, and their neglect has not been catastrophic to date because the scientific community is full of individuals and small groups who know how to volunteer, improvise and get things done. Somehow, the most pressing database requirements have always seemed to be met.

But scientific research has outgrown this system. Not long ago, most databases were small (tens of kilobytes) and published as typeset tables or simple online documents. Today, far larger and more complex databases are urgently needed in many fields at a level well beyond the reach of the traditional model of solitary workers or small groups.

For many disciplines, commercial forces are also becoming a problem. This is especially true when the participants in collaborations and government-to-government exchanges are from countries that offer radically different 'property rights' in data (see Box 1). Individuals and small groups are unlikely to bridge such differences on their own. For better or worse, professional societies, government agencies



Databases benefit next-generation experiments by intercepting data that would otherwise be lost.

or other large organizations will have to get involved.

Consequences of failure

If future databases are at risk, what are the consequences of failure? The answer depends on the functions of the databases and how they are expected to perform. Acquiring data is pointless unless society

actually captures what has been learned and uses the information. How well are scientists doing in this regard?

Consider 'big science'. When the Lawrence Berkeley National Laboratory in California shut down its Bevelac accelerator in 1993, most of the facility's unique heavy-ion data had never been published in any form. With better planning, scientists could

Box 1: Commercial turmoil

It is not hard to imagine how the rise of commercial forces in certain fields — notably biotechnology — could interfere with databases' ability to collect and publish data. But will it? So far, the main problem posed by commercialization is not so much greed as uncertainty.

Academic nervousness. Very few academic scientists know what their data would be worth on the open market. Despite this, scientists who operate databases often hear colleagues complain that data-sharing agreements could give valuable information away 'by accident'. It is difficult to say how often such statements translate into actual behaviour. Existing empirical evidence suggests that the effect is still fairly modest⁹.

New ways of doing business. The scientific-database business is so unfamiliar that entrepreneurs are still experimenting with new kinds of contracts. For example, Swiss-Prot often claims 'pass-through' rights in its data even after the information has been re-worked and merged with other material to create new databases (see http://www.expasy.ch/announce/sp_98sum.html for Swiss-Prot's licensing policy). Similarly, Celera has announced that users will not

be allowed to redistribute its data¹⁰. Such practices could limit scientists' ability to create new databases that build on and extend existing resources.

Legal uncertainty. In 1996, the European Union created strong property rights in data; similar, but weaker, legislation is pending in the United States.

Compromise seems inevitable, but may not happen for years. In the meantime, some US scientists have reported that their European collaborators are reluctant to share data. Government-to-government weather-data exchanges have also been affected.

It has been claimed that an analogous mix of inexperience and uncertainty has already interfered with biotechnology's ability to license and exploit patented inventions¹¹. Commercial forces may be acting as a similar brake on the creation of databases. The good news is that such problems should disappear over time. According to economists, excessive secrecy is irrational because owners can always make more money by entering into profit-sharing agreements and then pooling their data¹². In the long run, commercial forces should actually reinforce the traditional ethos of sharing information.

still be using this information to supply critical measurements in hot areas such as solar neutrinos, nucleosynthesis and cosmic rays. Instead, they will probably have to wait decades before these data are remeasured.

Now, history looks set to repeat itself. The US Department of Energy is currently developing advanced information systems for two new \$600 million accelerators — the Relativistic Heavy-Ion Collider at Brookhaven National Laboratory, New York, and the Thomas Jefferson National Accelerator Facility, Virginia. Unfortunately, neither system is designed to archive and disseminate data over the long run, and the energy

department, although aware of the problem, does not have the money to address it.

In the life sciences, millions of observations about location, interpersonal variation and function within the human genome are produced but not published. In principle, these data contain important clues for evolutionary biologists, biochemists, neuroscientists and health-care personnel, to name a few. Highly automated central depositories could make these data available to everyone.

Serving new users

Society cannot get full value for its investment in science unless anyone desiring

existing data actually gets them. So far, success has been uneven. In many fields, the prevailing attitude seems to be that academic databases are 'good enough' for everyone. That misses the point. When non-specialists can't use a particular database, the data might just as well be lost.

Some fields, such as chemistry (see Table 1), have long traditions of re-tailoring data so that they can be used elsewhere. Unfortunately, the experience in physics is more typical. Extensive government funding has allowed nuclear physicists to produce arguably the best scientific databases anywhere. Despite these advantages, most

Table 1 **Mainstays of science**

Database	Data stored	Maintenance
Review of Particle Physics/ Particle Data Group (1957–present)	Mass, spin and other properties for several hundred elementary particles, plus limits on many hypothetical particles.	Government-supported online database; a printed version is regularly published in various journals. http://pdg.lbl.gov
Evaluated Nuclear Structure Data File (ENSDF) (1946–present)	Fundamental nuclear structure and decay information for 3,100 nuclei and isomers.	Database maintained by Brookhaven National Laboratory's National Nuclear Data Center. Published online and as printed Nuclear Data Sheets. Government-supported with supplementary funding by Academic Press. http://www.nndc.bnl.gov/nndc/ensdf/
Table of Isotopes (1940–present)	Advanced, user-friendly listing of ENSDF, including flexible research tools	Government-supported online database maintained by Lawrence Berkeley Laboratory. http://isotopes.lbl.gov/isotopes/toi.html A version with extended content is published commercially by John Wiley.
Swiss-Prot (1986–present)	Protein sequence data, functions, domain structures, variants and related information.	Publicly supported database until 1996. Currently supported by commercial licensing fees plus additional funds from the Swiss government, the European Bioinformatics Institute and other public bodies. http://www.expasy.ch/sprot/sprot-top.html
Chemical Abstracts (1907–present)	Experimental data describing more than 22 million substances.	Non-profit entity supported by the sale of printed, electronic and online databases to academic researchers, industrial scientists, engineers and patent lawyers. http://info.cas.org/casdb.html
GenBank (1982–present)	Community-wide depository listing approximately 4.6 million sequences (more than 3 billion base pairs) found in the human genome.	Government-supported database available free over the Internet. http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
Nuclear Astrophysics Data Center (proposed)	Regularly updated, electronically searchable nuclear-reaction-rate database tailored to meet the needs of finite-element modellers.	Government-supported extension of existing databases and unpublished data. A limited version of the proposed website can be found at http://ie.lbl.gov/astro.html
Human Mutations Data	Advanced architecture database unifying and extending more than 150 separate websites.	Community-wide database to be launched as a private/public consortium.
DbSNP (1998–present)	Central depository containing more than 2.6 million single-nucleotide polymorphisms (SNPs). Advanced software lets scientists deposit thousands of SNPs at a time.	Government-supported database maintained by the US National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/SNP/
PathoSeq	Gene-sequence data focusing on disease-causing bacteria and fungi. Advanced relational database supports powerful search tools.	Proprietary database sold to industrial subscribers; primarily used in pharmaceutical research. http://www.incyte.com/products/pathoseq/pathoseq.html
SenseLab (1993–present)	Advanced data warehouse combining four previously independent neuroscience databases into a uniform, seamlessly searchable resource for neuron properties and models.	Government-supported academic research tool and technology development project. http://ycmi.med.yale.edu/senselab/
Distributed Active Archive Centers (DAACs)	Archived data sets obtained by Earth-observing satellites, aircraft, field campaigns and other experiments.	NASA-led federation of 10 government data centres, each of which publishes data on a particular topic (for example, "snow and ice", "land processes") relevant to global-change research. Data sets are available online and/or on CD-ROM.
Internets (1996–present)	Search engine providing links to hundreds of scientific databases.	Commercial website. http://www.internets.com/sciencedata.htm
Online Mendelian Inheritance in Man (OMIM)	Annotated catalogue of (largely) Mendelian diseases and polymorphisms; mitochondrial disorders are also recorded.	Community-driven database curated and maintained at the US National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/Omim

Large, electronically searchable databases have become essential to progress in many fields. Increasingly, scientists are experimenting with non-profit and commercial models to supplement traditional funding sources.

nuclear data used in medicine and industry are badly outdated. At least one non-destructive chemical-analysis technology, that of neutron-activation analysis, has already been delayed 30 years by this problem.

If anything, the problem is worse in biology. During the 1990s, private-sector bioinformatics companies spent much time turning academic databases into commercial research tools. Despite this investment, many hot research areas are still poorly served. For example, manufacturers of silicon microarray devices urgently need a community-wide database of reported observations to understand and improve their technology. Similarly, researchers studying single-nucleotide polymorphisms (SNPs) would like to compare their data against a comprehensive database describing all known human mutations. But the required databases simply don't exist^{2,3}.

Discovery does not end with publication, because the literature often contains overlapping and contradictory results. Far from being a nuisance, this discrepant information can be exploited to identify errors and recommend best values.

Modern technology has revolutionized this process in two crucial respects. First, on-line databases are never fixed in stone. When questionable new data appear in the *Table of Isotopes*, for instance, affected authors are quick to send e-mails and unpublished information. This practice provides an important safeguard against editorial errors. The same is true for biological databases.

Second, computers have systematized the traditional editor's trick of doing 'sanity check' calculations to see if a particular result looks reasonable. For example, *Table of Isotopes* evaluators routinely combine existing results to calculate additional variables. If the new variable conflicts with known data, the result provides a powerful tool for detecting errors in one (or more) of the input measurements. Otherwise, the new quantity can be added to the literature as a result in its own right. Over the past five years, *Table of Isotopes* has discovered thousands of discrepancies and many new data in this way. Similar opportunities exist in biology, where creators of many databases (for example, for genetic mutations) already have a wide variety of consistency-checking software products to choose from.

An obvious extension of this strategy is to make observations 'in silico'. Several US and European initiatives are working on enormous (up to 40 terabyte) 'virtual sky' archives for astronomers to explore. Early versions of these projects have already found a rich harvest of quasars and other objects^{4,5}. Beyond virtual observations lie virtual experiments. Here, too, accurate data are essential. For example, supercomputer

Discrepant information can be exploited to identify errors and recommend best values.

simulations of controlled fusion, stellar evolution and supernovae are no better than the physics that goes into them. Unfortunately, there is still no up-to-date, curated database that delivers this information in the formats used by modellers.

Purists may object that inferences from the literature and supercomputer simulations are never as accurate as a good experiment, but that is rarely the choice. Many

experiments are unaffordable; others are beyond existing technology. Under these circumstances, using large databases as if they were virtual particle accelerators (or observatories, or gene-sequencing machines) may be the only available option.

Data mining

Automated surveys for faint signals have always depended on advanced databases. Until fairly recently, most of these 'rare event' searches were limited to high-energy physics experiments. By the late 1990s, however, similar searches had become 'grand challenges' in such diverse fields as astronomy, climatology, meteorology, Earth imaging and biology.

Thirty years ago, constructing a computer-searchable database big enough to hold 1.5 million events (several hundred megabytes of data) per year was enough to win Luis Alvarez of the Lawrence Berkeley National

Box 2: E-solutions?

Modern scientific databases are labour-intensive. Despite aggressive computerization, database creators still do much of their work by hand, one entry at a time. Because automation has been under way for nearly half a century, conventional techniques are approaching their limits. This leaves new technology, especially on the Internet, as the most likely way forward. Could future web tools strengthen — or even replace — current methods for producing databases?

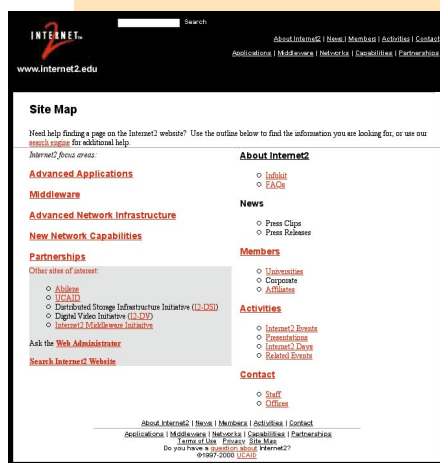
New web tools will make database creation easier in at least three ways. First, scientists need web crawlers to find online data. However, today's crawlers are still far less complete than old-style paper bibliographies. Web developers are working hard to close this gap. Second, some large databases (notably GenBank) are starting to outgrow the web's ability to store and transmit information. CERN's GIOD and The US Grand Challenge collaborations will ease this threat by developing technologies that distribute data across multiple servers (ref. 4 and <http://chep2000.pd.infn.it/paper/pap-c292.pdf>).

Finally, 'middleware' tools are making it easier for cooperating researchers to link the data on their servers into a single, overarching network. One of the best current examples is the US Geological Survey's National Spatial Data Infrastructure Clearing House, which allows users to find and retrieve climate data stored on more than 200 participating servers (J. Restivo, personal communication). Better middleware tools are a top priority for the US Internet 2 and CERN's Grid initiatives (see <http://www.internet2.edu/middleware/overview/areas-of-activity.html>).

Good as they are, none of these tools replaces human editors when it comes to combining and evaluating independent databases. How likely is this to happen? In one promising strategy, human editors write detailed instructions for translating each database's unique computing

conventions, data fields and scientific nomenclatures into a standard format. Computers then execute the instructions to build a 'data warehouse'. Unfortunately, current warehouses are rarely able to unify more than a dozen databases. After that, the amount of human intervention required to create and update the instructions becomes unworkable (T. Slezak, personal communication). In principle, advanced web crawlers could break this impasse by learning to translate data on their own. For now, though, crawlers still have trouble retrieving even simple objects such as telephone listings. Most Internet researchers agree that human editors are not likely to be replaced soon.

New web tools could help put large, sophisticated databases back within the reach of individuals and small groups. At the same time, technology is not enough. For the foreseeable future, even the most powerful web tools will depend on humans to organize networks, adopt standard nomenclatures and combine independent data sets. New tools are only part of the solution. The biggest challenges are social and economic.



Laboratory the Nobel prize for physics. Today, experiments such as Lawrence Berkeley's STAR detector can record 54 terabytes ($\times 10^6$ megabytes) per year. Experiments at the Large Hadron Collider at the European Organization for Nuclear Research (CERN) are expected to reach 100 petabytes ($\times 10^9$ megabytes) by roughly 2010 (ref. 4).

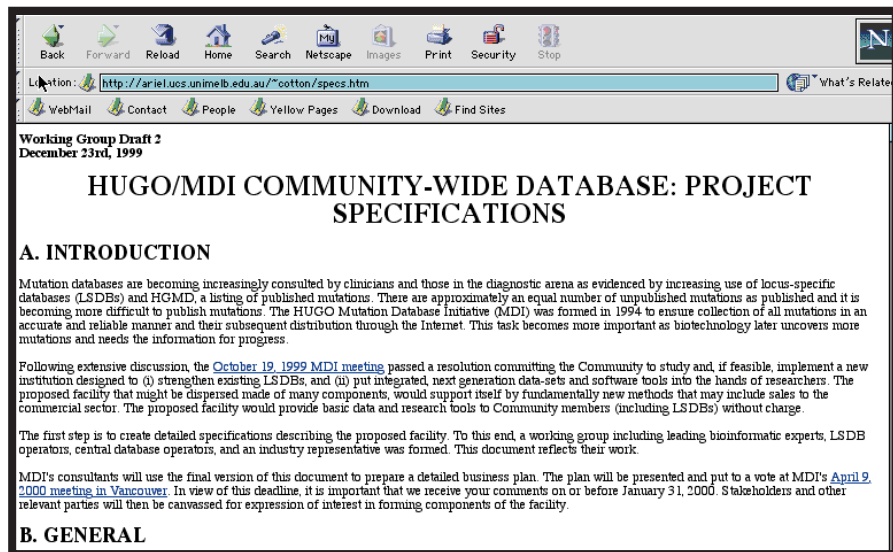
Meanwhile, biologists are already planning for the day when silicon microarray detectors can simultaneously measure the 'expression level' for each of the body's estimated 100,000 genes. Exploiting these data will mean finding the handfuls-to-hundreds of genes responsible for particular biological processes and diseases. Scientists are actively discussing the statistics and computing methods needed to separate these tiny, complicated signals from the background noise of several thousand genes performing unrelated functions. Whichever strategy is chosen, however, one fact is already obvious: biologists will have to combine observations from hundreds and perhaps thousands of different researchers into a unified, seamlessly searchable database. This challenge is far from trivial in a world where bioinformatics specialists still have trouble combining more than a dozen sites into a unified 'data warehouse'. Commercial pressures to hoard data could complicate the task even further (see Box 1).

Taming the web

Ten years ago, many scientists thought that the World-Wide Web would encourage individuals and small groups to build more databases. In a way, this prophecy came true, and some scientists have even built substantial research agendas around data spotted on the web. Unfortunately, the web's strengths are also its weaknesses. Most researchers need to combine and unify data from as many sources as possible. By contrast, the web usually fragments information by encouraging each scientist to create his or her own home page. In genetics, for example, more than 150 sites are currently devoted to mutations in particular human genes (see <http://ariel.ucs.unimelb.edu.au:80/~cotton/mdi.htm>). In many fields, non-standard computing conventions, nomenclatures and quality-control practices create additional difficulties.

New web tools may ease, but will not eliminate, these problems (see Box 2). For example, many scientists want to combine

Society can't get full value for its investment in science unless anyone desiring existing data actually gets them.



Members of the Mutation Database Initiative want to build a single depository for their data.

existing climate, biodiversity and environmental databases into large networks. Unfortunately, advanced search engines and data mining are no better than the information they operate on — and existing databases are riddled with errors. Human editors will have to clear the way before these tools can succeed⁶.

We believe that the web's problems are more social than technological. Over the past six years, the Human Genome Organization's Mutation Database Initiative (MDI) has helped to build a broad community consensus in favour of standardized nomenclatures and — to a lesser extent — computing conventions and systems for cross-referencing human mutation data with other types of databases and/or different species^{7,8}. Many MDI members believe that the next step is to build a single, community-wide depository for its members' data. Private companies — which have their own reasons for wanting larger, more unified mutation databases — will be asked to join the project^{2,7,8} (the depository proposal is described at <http://ariel.ucs.unimelb.edu.au/~cotton/specs.htm>).

Changing the system

Most of the needs discussed here are not subtle. They require money. Those involved need to plan for databases in the same way that we think about new buildings and hardware. Several US federal agencies, including the space agency NASA and parts of the Environmental Protection Agency, have taken an important first step by requiring would-be grant recipients to explain how they plan to store and disseminate their data. In biology, similar foresight has been displayed by the launching of several high-visibility efforts to recruit and train more bioinformatics experts. Such initiatives are encouraging, but they are only a start.

Meanwhile, grants from private and public agencies are approaching their limits.

Sales to the commercial sector offer one way out, although this raises the thorny issue of public access. One of the most appealing strategies is for scientists to keep their general research databases in the public domain while selling customized spin-offs to the private sector. The main advantages of such a strategy are that it would avoid commercializing and/or monopolizing basic data, reward science for at least some of its contributions to the economy, and encourage academic scientists to find users who are currently underserved².

Stephen M. Maurer is an intellectual-property attorney at 2632 Hilgard Street, Berkeley, California 94709, USA, and a consultant to the MDI (e-mail: maurer@econ.berkeley.edu). Richard B. Firestone is a staff scientist at the Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA, and author of Table of Isotopes. Charles R. Scriver is Alva Professor of Human Genetics and Professor of Pediatrics and Biology at McGill University-Montreal Children's Hospital Research Institute, Montreal, Quebec H3H 1P3, Canada. He is a co-director of the MDI and curator of the PAHdb knowledge base.

- Ridley, M. *The Economist* **318**, 7644 (16 Feb. 1991).
- Maurer, S. *Hum. Mutat.* **15**, 1 (2000).
- Brazma, A. et al. *Nature* **403**, 699–700 (2000).
- Szalay, A. et al. *Accessing Large Distributed Archives in Astronomy and Particle Physics* http://pcbunn.cithec.caltech.edu/aldap/kdi_proposal.htm (1999).
- Murtagh, F. *The Virtual Observatory: Methodologies for Data Handling* <http://newb6.u-strasbg.fr/~ccma/vol/virtual-observatory.html>
- Winker, K. *Nature* **401**, 524 (1999).
- Cotton, R. G. H., McKusick, V. M. & Scriver, C. R. *Science* **279**, 10 (1998).
- Cotton, R. G. H. & Kazazian, H. H. Jr (eds) *Hum. Mutat.* **15**, no. 1, spec. issue (2000).
- Campbell, E. et al. *Res. Policy* **29**, 303–312 (2000).
- Butler, D. & Smaglik, P. *Nature* **403**, 231 (2000).
- Heller, M. A. & Eisenberg, R. S. *Science* **280**, 698 (1998).
- Scotchmer, S. J. *Econ. Perspectives* **5**, 29 (1991).

Acknowledgements. We thank Suzanne Scotchmer, University of California at Berkeley, Tom Slezak of the Joint Genome Institute/Lawrence Livermore National Laboratory and Richard Cotton, Mutation Research Centre, St Vincent's Hospital, Melbourne, Australia, for suggestions and support.