

Let data speak to data

Web tools now allow data sharing and informal debate to take place alongside published papers. But to take full advantage, scientists must embrace a culture of sharing and rethink their vision of databases.

Upload and share your raw data, and have a high impact factor for your blog — or perish? That day has not yet come, but web technologies, from personal publishing tools such as blogs to electronic laboratory notebooks, are pushing the character of the web from that of a large library towards providing a user-driven collaborative workspace (see page 547).

This will in turn expose many fields of research to changes that are already sweeping disciplines such as bioinformatics and high-energy physics. A decade ago, for example, astronomy was still largely about groups keeping observational data proprietary and publishing individual results. Now it is organized around large data sets, with data being shared, coded and made accessible to the whole community. Organized sharing of data within and among smaller and more diverse research communities is more challenging, owing to the plethora of data types and formats.

A key technological shift that could change this is a move away from centralized databases to what are known as 'web services'. These are published interfaces that serve to simplify access to data and software (for an example of such services in action, see www.ebi.ac.uk/xembl/index.html). Until recently the preserve of expert programmers, such interfaces now mean that anyone with even a basic knowledge of programming can automate data processing and analysis.

Various sorts of data are increasingly being stored in formats that computers can understand and manipulate, allowing databases to talk to one another. This enables their users quickly to adapt technologies to extract and interpret data from different sources, and to create entirely new data products and services.

In biodiversity research, for example, rather than creating centralized monolithic databases, scientists could tap into existing databases wherever the data are held, weaving together all the relevant data on a species, from its taxonomy and genetic sequence to its geographical distribution. Such decentralization also helps to solve the problem that databases are often the fruits of individual or

lab research projects that are vulnerable to the vagaries of funding, and to people and labs moving on to pastures new.

Although discipline-specific databases have an indisputable role, science also needs to capitalize on large common repositories for data, whose preservation is guaranteed, and where the data can easily be used by anyone. If that sounds utopian, consider OurMedia, a service created by the Internet Archive and the Creative Commons, which allows anyone to store and share permanently and free of charge any digital work — even their videos and holiday photos. And last month Google launched Google Base, which also allows anyone to upload anything to its massive platform.

Such services will also require new thinking on open data. Web services are dependent on computers being able to freely access data in real time. Although GenBank and many large databases allow unhindered access to their data, many research organizations still cling to obsolete manual data permission policies, which prevent their data being used by web services.

Scientists may be justified in retaining privileged access to data that they have invested heavily in collecting, pending publication — but there are also huge amounts of data that do not need to be kept behind walls. And few organizations seem to be aware that by making their data available under a Creative Commons licence (see <http://creativecommons.org/license>), they can stipulate both rights and credits for the reuse of data, while allowing its uninterrupted access by machines.

As web services empower researchers, the biggest obstacle to fulfilling such visions will be cultural. Scientific competitiveness will always be with us. But developing meaningful credit for those who share their data is essential, to encourage the diversity of means by which researchers can now contribute to the global academy. ■

"By making data available under a Creative Commons licence, scientists can stipulate rights and credits for the reuse of data."

Life at the edge

Successes in structural studies of membrane proteins deserve to be celebrated.

Sealed membrane systems are a defining feature of cellular life. Membranes provide a barrier between the cell and its external environment and, in many organisms, divide the interior of the cell into functionally distinct compartments. The barrier, comprising lipids that are impenetrable to electrically polarized molecules, has proteins inserted within it that allow the selective transport of

ions and molecules. These proteins enable cells to ingest nutrients, excrete metabolic waste, sample the environment for the sake of the immune system, and store energy by means of ion electrochemical gradients. They mediate molecular signalling across the barrier. And they are the very devil to study.

Genome sequencing projects have highlighted the central role of membrane-linked processes in cells. They have revealed that membrane proteins represent about a third of the gene products in most organisms. Unfortunately, our molecular knowledge of these membrane proteins lags far behind that of proteins found in the cell cytoplasm and in external environments. This is primarily due to the difficulty in obtaining high-resolution structural information on