EDITORIAL

nature cell biology

Sharing data

Reference datasets should be accessible independently of scientific papers in a citable form, allowing attribution.

Scholarly publication remains essential for describing and contextualizing findings, but it is inadequate as the only document of research activity. Most journals require a significant conceptual advance, and format constraints typically allow only for the presentation of representative qualitative, or statistically processed quantitative data. Consequently, the majority of raw data never emerges from lab hard drives, and a wealth of information, hard work and funding is wasted. High throughput platforms generate reams of data that cannot be captured in traditional papers. Moreover, methods sections fail to adequately describe metadata essential for the comparison and reproduction of experiments. Databases are essential for comprehensively archiving both published and unpublished data, but have only become fully integrated into the scientific process in a few cases, such as DNA sequencing and microarray data. For many types of data, including light microscopy, no databases exist at all. This is partially due to the variety of formats, making it hard to derive universal standards (the open microscopy environment, OME, aims to develop these for light microscopy; *Nature Cell Biol.* **6,** 909). Furthermore, only a few funding agencies have pursued database development and long-term maintenance (notably the NIH and Wellcome Trust; *Nature Cell Biol.* **8,** 425). While having to comply with journal/funder data deposition policies can seem a chore for scientists, systems biology is upon us and it is essential that cell biologists embrace the chance to contribute to, use and develop databases. In turn, to remain relevant, journals must link content systematically to community-endorsed public databases.

Prepublication data deposition on databases and preprint servers, as long practiced in the physical sciences, is relatively new to biology. The concept came of age with the 1996 Bermuda declaration of the Human Genome Project, presenting a community standard for rapid sequence-data deposition. A string of subsequent meetings, most recently the 2009 Toronto Data Release workshop, called for prepublication release of other large datasets that constitute a resource for the whole community, including proteomic, metabolomic, RNAi, chemical and, importantly, clinical data (*Nature* **461,** 168). Prepublication release of large datasets that cannot be efficiently or comprehensively analysed by the generating lab is essential for efficient scientific progress. Notably, the depositors gain from having access to other prepublication datasets and from collaborations deriving from the data. On the other hand, research achievement is measured by publication and much biology research is inherently competitive. It is therefore essential that the source lab is given appropriate protection and recognition. Some databases have therefore rightly developed a 'publication exclusivity' policy that users have to sign for full access. For example, the NCBI dbGaP database, which hosts genome-wide association studies, requires a 12 month publication delay. In fact, for databases such as dbGaP that carry patient data, controlled access is also important to address privacy laws.

Setting embargoes is a reasonable compromise, as long as the rules are policed. Recently, a paper identifying a gene associated with addiction was published in PNAS several weeks before the end of the embargo period of the dbGaP dataset analysed. While this breach slipped by referees and editors, the system worked in that a complaint by Laura Bierut, co-author of the primary data, led to retraction of the paper within days and an investigation by the NIH (*Science* **325,** 1486–1487; *Proc. Natl Acad. Sci. USA* **106,** 16893). Nevertheless, the case demonstrates that journals must incorporate data-use limitations into their policies, author declarations and compliance screens.

Some say embargoes are not enough, as researchers can prepare manuscripts for submission on embargo expiry. In our view this is not a problem, as long as the source data is citable and cited, and the primary data-generating researchers receive appropriate 'academic credit' for generating the data — even if they loose out on derivative papers. In the future, bibliometric assessment must therefore take into account both citations of research papers and database entries. Databases must allow attribution to defined data subsets. Importantly, database entries can evolve. Thus, it is essential to archive uniquely citable versions of a dataset.

Large reference datasets that benefit the wider community and that cannot be analysed efficiently by the data producers should enter the public domain without delay, as long as appropriate attribution and credit can and is given. Scientific culture has to change so that data is valued alongside publications.

Funding pain in Spain

On the eve of budget decisions, the scale of cuts to basic research funding remains ill-defined.

Spain is suffering particularly badly from the global economic crisis and significant fiscal cuts seem unavoidable. On the other hand, Spain remains the poster child of a reinvigorated European research landscape and a remarkable example of how well-directed funding can catalyse the emergence of a world-class research base in a short time.

Even as other equally affected countries, notably the US, inject cash into research to stimulate the economy, many scientists and societies like SEBBM (Spanish Society for Biochemistry and Molecular Biology) are expecting some funding contraction after the 2009 science budget shrivelled from the planned 16% increase to a mere 2.5%. The situation remains confused on the eve of the upcoming budget decisions: on September 24th, the Spanish media reported a 28% overall cut, but a 3% increase for the biotech industry. A day later, science Minister Cristina Garmendia stated that nothing would change. On October 1st, a 14.7% cut (about 300 million Euros) was reported, which even when offset against the biotech spending increase would remain a 3% cut overall (a 13.6% cut was earmarked for the National Research Council, CSIC, which accounts for half the research activity). The following Monday, Spanish premier Rodríguez Zapatero quoted a 0.2% increase for *Becas* (research fellowships), while cuts would be focussed on grants and lab infrastructure at government research centres (los Organismos Públicos de Investigación, encompassing CSIC). As we went to press, Garmendia reaffirmed on breakfast TV her plans to redirect funds to the biotech industry to increase the number of firms fivefold.

Significant real-term cuts may be unavoidable, although Spain would do well to invest stimulus funds into both basic and applied research, as proven long-term wealth generators. A dramatic re-focus from basic science to biotech incubation would belie a short-term income-generating strategy that should be carefully considered. Industry relies on thriving basic research. Spain has painstakingly built an enviable reputation in basic research, but maintaining this requires reliable funding streams.