EMBO reports

outlook

Sustainable digital infrastructure

Although databases and other online resources have become a central tool for biological research, their long-term support and maintenance is far from secure

Ruth Bastow & Sabina Leonelli

The past decade has seen an unprecedented explosion of data, tools and databank resources in the biological sciences, most of which can be freely accessed by researchers over the Internet. A 2009 survey listed 1,230 molecular and cell biology databases that were available online (Cochrane & Galperin, 2010). A recent review of the data generated by plant biologists in the past 10 years illustrates how these data sets are useful not only for the individual laboratory that created them, but also for other researchers. Combining data from several laboratories enables scientists to gain valuable insights into biological processes (Brady & Provart, 2009). In most areas of experimental biology, databases and online repositories have become central tools in laboratories; *in silico* experiments now regularly play an important role in planning and analysing experiments leading to scientific discoveries.

In most areas of experimental biology, databases and online repositories have become central tools in laboratories…

Whilst researchers continue to spend more of their time grappling with the growing deluge of data, database administrators and curators face the problem of securing long-term, sustainable funding. Access to online data has become a basic requirement for conducting scientific research, but the growth in data, databases, websites and resources has outpaced the development of mechanisms and models to fund the necessary cyberinfrastructure, curation and long-term stewardship of these resources. Social science research on the sustainability of databases and their funding sources has blossomed (Wouters & Schroeder, 2003; Chandras *et al*, 2009; Maron *et al*, 2009; Leonelli, 2010a,b), but a single, viable framework for sustainable and long-term stewardship of data and resources has not emerged.

Government-funded databases and repositories tend not to be controlled by private interests or concerned with short-term impact

In this article, we review some of the financial models and mechanisms that could be employed to support public repositories, databases and resources in the long term. We provide examples of efforts that have been made so far, and critically discuss the advantages and disadvantages of each model.

A large number of publicly available databases and repositories are supported by funds from government bodies. The Mouse Genome Informatics resource (MGI; http://www.informatics.jax.org), the Saccharomyces Genome Database (SGD; http://www.yeastgenome.org), FlyBase for *Drosophila* research (http://flybase.org), the Zebrafish Information network (ZFIN; http://zfin.org), The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/index.jsp) and Gramene, a data resource for comparative grasses genomics (http://www.gramene.org) are all supported by public money (Bult *et al,* 2008; Dwight *et al,* 2004; Tweedie *et al,* 2009; Sprague *et al,* 2008; Swarbreck *et al,* 2008; Liang *et al,* 2008).

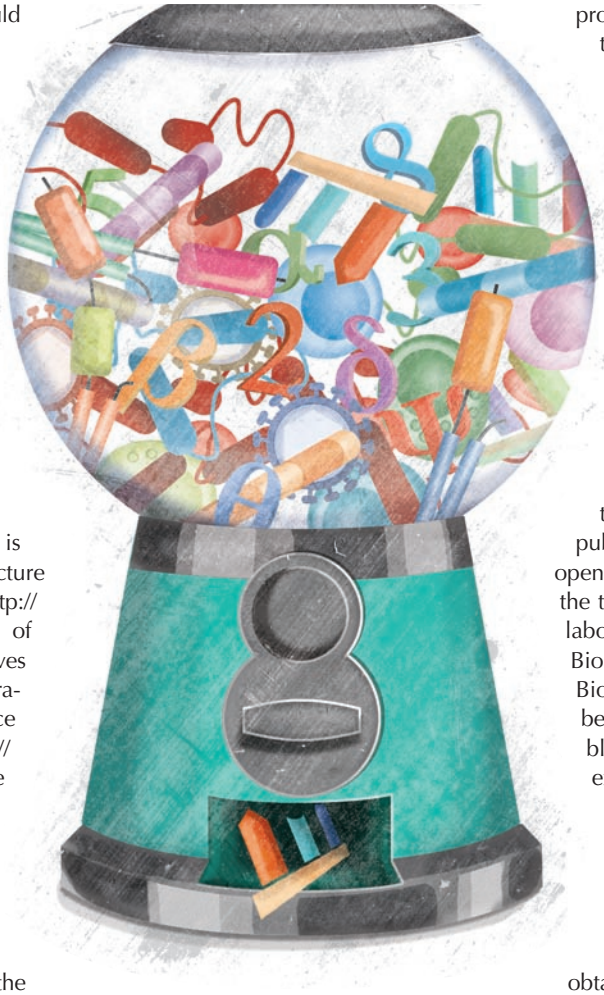
Government-funded databases and repositories tend not to be controlled by private interests or concerned with short-term impact. Instead, they provide researchers with instant, free access to data with no login procedures or payment required. This open access model acts as an underpinning structure that lowers barriers for the sharing of ideas and information. It helps to foster collaboration and ultimately drives the progression of science by facilitating future discoveries (Leonelli, 2010b).

Many funding bodies require databases to compete for funding with hypothesis-driven research that is usually assessed on its novelty and ability to generate publications. However, these measures of success are incompatible with data-intense research (O'Malley *et al,* 2009) and cyberinfrastructure (Leonelli, 2010a) for which the key requirement is not the generation of new resources, but rather the capacity to maintain and improve existing ones. This creates an environment in which it is relatively easy to set up novel databases or resources, but difficult to maintain infrastructure over the long term.

… it is relatively easy to set up novel databases or resources, but difficult to maintain infrastructure over the long term

Exceptions include the UK Biotechnology and Biological Sciences Research Council (BBSRC)'s Bioinformatics and Biological Resources Fund—which allows databases, tools and repositories to apply for funds separate from hypothesis-driven grants—and the US National Science Foundation (NSF)'s directorate of biological infrastructure support for research resources, which includes funding for the development of informatics

tools and resources. These models could be expanded by the BBSRC and the NSF, and would be useful for others to emulate. However, any national funding scheme will not be able to provide a solution to a global problem: data do not adhere to geographical boundaries. An international funding agency that could fund data infrastructure would be an ideal solution, but such an institution does not currently exist.



Several models for collaborative funding of data infrastructure are being investigated in Europe. The European Life Sciences Infrastructure for Biological Sciences is planning to build a sustainable infrastructure for biological information in Europe (http://www.elixir-europe.org). The Council of European Social Science Data Archives has initiated a distributed research infrastructure to integrate 20 social science data archives across Europe (http://www.cessda.org). Beyond Europe, one example of international collaboration is the Protein Data Bank (PDB; www.wwpdb.org). Initially set up in 1971 as a collaboration between the Cambridge Crystallographic Data Centre in the UK and the Brookhaven National Laboratory in the USA, the PDB has changed many times during the past 39 years and had several management structures (Berman *et al*, 2007). Today, it exists as a collaboration between the major protein data banks and repositories in Europe (PDBe), Japan (PDBj) and the USA (RCSBPDB), which are supported by numerous funding agencies.

Regardless of the funding mechanism, the public purse—both nationally and internationally—is not infinite, and the continuing growth of databases and repositories means that it is not financially viable for these to be exclusively public-funded. So, how does one decide what should be funded? On the basis of data from the NSF and the US National Institutes of Health (NIH), the 2009 annual running costs of the major model organism databases (MGI, SGD, FlyBase, Rat Genome Database, WormBase, ZFIN and TAIR) ranged from $1.6 million for TAIR to $6.3 million for MGI.

What are the measures for assessing the effectiveness and efficiency of these databases, and how do we know if they are able to meet the needs of their communities over time? Current assessment methods for hypothesis-driven research are not appropriate for data infrastructure. Therefore, new metrics are needed for assessing the impact, quality and usefulness of databases or repositories. These could include mechanisms for peer review of annotations; explicit academic recognition for donating and sharing data—which would act as an incentive for researchers to invest their time, and value the work of those responsible for the curation and maintenance of data; mechanisms for assessing and monitoring how databases respond to user feedback; and, most importantly, methods to track how widely a database is used, its role in accelerating research within the community it serves, and how data in open-access databases is re-used to generate new discoveries. All of this would help to measure the contribution of databases to research.

Until now, public funding of databases has generated the most effective model for providing free access to data for scientists, which is essential for the progress of science. However, in the future this model will need to support the growth of open science in a globally networked world.

The following six models are different approaches to partly or fully covering the running costs of databases, repositories and other service infrastructures, by charging the end users.

The 'industrial support model' requires commercial users to pay a fee for access to data, tools or resources, whereas publicly funded users gain instant and open access. One example of this model is the two-tier system that Swiss Prot—a collaboration between the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI)—introduced between 1998 and 2002 as a possible solution to the funding crisis it was experiencing at that time. Interestingly, Swiss Prot returned to a free access model in 2003 after SIB, the European Molecular Biology Laboratory (EMBL)—which includes the EBI—and the Protein Information Resource formed the Uni Prot consortium and obtained a grant from the NIH (Bairoch *et al*, 2004).

Requesting a subscription fee from commercial users provides databases and repositories with at least one source of stable income. In some cases, this might also make it easier to request funds from a public funding agency. However, this type of funding can be difficult to police; for example, if a commercial user is working from home. Monitoring commercial use also generates additional costs in comparison to a 'free access for all' model. The database provider would need to monitor which companies are using the database, request subscription fees and ensure that payment is made in full and on time. Relying on industrial support also exposes this model to the vagaries of financial markets, perhaps making it less secure in the long term. This model might also encounter intellectual property issues, with regard to the data-release policy of public funding bodies. An industrial approach to data access and storage is not inherently compatible with the non-commercial goals of academia, and industry is unlikely to

invest in resources that are not guaranteed to bring returns (Leonelli, 2010a).

A public–private consortium is a mixture of funding from government bodies and industry. One of the most successful examples is the Structural Genomics Consortium (SGC), which "provides protein structures of relevance to human health in the public domain free from restriction on use" (Lee *et al,* 2009). The SGC consists of three academic laboratories in Oxford, Toronto and Stockholm and is funded by a consortium of 13 public and private bodies including GlaxoSmithKline, Genome Canada, Merck, Novartis, the Swedish Foundation for Strategic Research and the Wellcome Trust. The three laboratories work to solve structures from a list of target proteins from humans and human parasites that are nominated by the funders. All solved structures are deposited in the data bank and supporting companies do not receive priority access to the structures.

An international funding agency that could fund data infrastructure would be an ideal solution, but such an institution does not currently exist

Public–private partnerships often occur when the project clearly benefits the industry partner; a company will pay for data and services that it needs for its own research and development activities. The industry partners in the SGC, for instance, receive high-quality, pre-competitive data that helps them to facilitate drug discovery. Can such a model be employed for more generic research—such as genome annotation—for which the outcome is less defined? The EBI is one example of a public–private partnership that supports a range of molecular databases and is funded by a mixture of private and public funders. However, its large size and unique position as the fulcrum of European bioinformatics research are likely to have contributed to its success.

Public–private consortia are useful for securing commercial support for generating freely available data that will benefit the wider academic community, provide funds for data infrastructure and reward publicly funded scientists for their expertise and contributions to the market economy. However, this approach is also subject to the vagaries of the market and company policies, and it is not clear how intellectual property issues can be reconciled with the 'sharing' ethos fostered by governmental funding agencies.

In the 'value-added/asymmetrical pricing model', a basic data set within the database is freely available and anyone—individual scientists or companies that are willing and able to pay a higher fee—can buy additional levels of service, better data access or additional tools and resources. One example is Genevestigator, a high-quality, manually curated expression database and meta-analysis system for animals, plants and microorganisms (Zimmermann *et al,* 2004). Genevestigator was originally developed at the Swiss Federal Institute for Technology in Zurich and is now run by a company that licenses the platform to academia and industry. It uses data that have been generated by third parties and adds high-quality curation, data control and meta-analysis tools. Users can access Genevestigator through three routes: 'open access'—with no fees attached and limited use of meta-analysis tools (one gene at a time); 'classic'—free access for academics only and use of meta-analysis tools for up to 50 genes at a time; and 'advanced'—academic and commercial users purchase access to broader meta-analysis tools and several additional tools.

This tiered payment system allows data to be freely accessible to the community and also provides a reliable income for the company. As more data sets become available, this might be a productive approach for generating financial support. However, this model creates inequalities in access which could potentially increase the divide between rich and poor research institutes and organizations.

Researchers commonly pay a set fee for materials, in return for guaranteed quality control, distribution and service. An example of this product supply model is the Nottingham Arabidopsis Stock Centre (NASC) and its sister organization in the USA, the Arabidopsis Biological Resource Centre (ABRC), which provide seeds, stocks, information resources and transcriptomics services to the international *Arabidopsis* community (Scholl *et al*, 2000). These organizations provide a secure archive of genetic material to the community and enable NASC and ABRC to cover the costs associated with growing, harvesting and storing seeds. NASC is partly supported by BBSRC funding, which means that sales do not need to cover the costs of online ordering and maintenance of the data infrastructure. It should be possible to run all NASC services—both seed distribution and informatics—from income generated by sales alone, but this would require NASC to raise prices from £3.50 to around £11 per stock. It is unclear whether this model would provide a stable source of income, or to what extent current sales would be affected.

An industrial approach to data access and storage is not inherently compatible with the non-commercial goals of academia…

A 'cost recovery model' can in theory provide a reliable mechanism by which to recoup fixed costs. However, it seems that biological resource centres using this approach are only able to partly cover the full cost of the service, because fees have to be sustainable for investigators on a fixed grant income and affordable to the majority of users. The end user provides a percentage of the income for the service, with the remainder being provided by public and governmental grants.

A cost recovery model for a service that only provides data would probably involve a compulsory subscription, whereby every user would pay for access. A recent survey of the *Arabidopsis* community indicates that this might not be viable; 50% of those surveyed (147 individuals) indicated that they would be willing to pay up to $50 a year for access to *Arabidopsis* genome data supplied by TAIR, a fee that would only provide one-tenth of their annual running costs (http://www.arabidopsis.org/portals/masc/journal.jsp#Bioinformatics_Survey_18Mar2010). Subscription costs could be increased, but this would create inequalities in data access—only those researchers, laboratories or countries that were able to afford the subscription would have access, while teachers, students, undergraduates and researchers in countries/institutions with low budgets would be locked out. This approach is not compatible with the open access data policies of funding agencies, such as the NSF and BBSRC.

Compulsory subscription would also create major obstacles for data sharing and interoperability between databases. For example, TAIR supplies all the *Arabidopsis* gene structure and function

data available in NCBI's RefSeq and Entrez Gene websites, AtEnsembl, Ensembl Plants, UniProt and the Gene Ontology Consortium, among others. If TAIR were to use a compulsory subscription model, it would not make commercial sense to share data with free repositories, as users would use these instead of TAIR. This would be counter-productive to the interoperability of databases and their utility within integrative biology, systems biology and any other form of data-intensive science. Stopping the flow of data across databases would also damage resources such as the Gene Ontology, which is constructed as a collaborative consortium of model organism databases (Ashburner, 2000).

The wiki approach to generating and curating data is extremely attractive as, in theory, it has extremely low costs

It is difficult to predict the sustainability of a subscription model. Subscription numbers might decrease overtime when users migrate to free databases/repositories that emerge. Researchers might also stop providing data to a compulsory subscription database/repository if it does not fulfil the free access policy requirement of the agency funding their research. This would make the database/repository progressively out-of-date, further reducing subscription numbers over time. It would also increase the risk of 'old data' being lost if subscription levels were to fall to a level that could no longer cover the running costs of the database/repository.

Overall, the compulsory subscription model for databases/repositories has several drawbacks that make it both financially unviable and counter-intuitive to the idea of data sharing and data-driven research.

Online advertising is often a useful way for a business to accrue part of its income. Yet, when advertising was tested on the Bio-Array Resource for the Plant Functional Genomics website (Toufighi *et al*, 2005), less than 1% of the funds needed to employ a bioinformatician were generated, despite more than 50,000 uses of the website in one month. It therefore seems unlikely that this approach could provide enough income to support databases and resources. It might also be incompatible with the educational, not-for-profit status maintained by some universities and research institutes.

Model organism databases such as MGI, FlyBase and TAIR provide vast amounts of information on each gene in a genome via a small team of contributors/curators. In contrast, the online encyclopaedia Wikipedia functions by receiving small amounts of information from a large number of contributors. The Wikipedia approach to sharing knowledge has been so successful that scientific contributions to Wikipedia now rival the online *Encyclopaedia Britannica* for accuracy (Giles, 2005). It has therefore been suggested that a 'wiki' approach to genome annotation—that is a shared resource that anyone can add to and edit—could collate accurate information from experts, without the need for expensive curators (Salzberg, 2007). The GeneWiki portal contains several examples of genes from the human genome that have been annotated online and do not currently exist in the Gene Ontology Annotation database (Huss *et al*, 2010).

The wiki approach to generating and curating data is extremely attractive as, in theory, it has extremely low costs. It depends, however, on community participation and might face difficulties in drawing contributions from busy users. It would therefore need to provide incentives for active participation, such as making data donation and annotation compulsory for submission to high-profile journals and/or receipt of grants. Compliance with these regulations would need to be actively policed, and it is unclear who would be responsible for ensuring that data and annotations had been submitted to the right databases. This would increase the cost and complexity of this model. Further, the 'wikification' of a genome might lead to poor quality control. Incomplete or inaccurate data can easily be placed in a GeneWiki, while crowd-sourcing is likely to generate inconsistencies in annotation methods and formats. Overcoming this would require administrators and/or curators, which would again increase the costs.

Wikis can be more flexible than structured genome annotation databases and allow the addition of free text; this is often more approachable for a wider audience and not just scientific experts. This might encourage non-experts to learn more about the field, but it is unlikely to assist in the curation of a genome, given the highly technical nature of annotation. The lack of skills and expertise to annotate data in a model organism database is a problem within the research community; experimenters are often unaware of how data should be annotated and lack the time and resources to learn these skills (Leonelli, 2010a).

Finally, despite its low overheads, some costs would still be incurred by this model: storage space, software, basic maintenance by IT engineers and curators. Funds would have to be found and sustained from public or private sources, and these would be subject to the same problems experienced by other databases and repositories.

…no current model is able to meet the requirements of cyberinfrastructure and data-intensive research

The drawbacks of the wiki approach to genome annotation might mean that GeneWikis become complementary to, rather than alternatives to, biological databases. This might change as new generations of researchers become more comfortable with online tools and acquire the skills needed to annotate a genome. However, the problem of incentives remains: public and private funders need to add rewards and recognition for contributing to online resources. One example of this could be to make data donations equivalent to publications in assessment exercises, perhaps by assigning them a unique DOI (Digital Object Identifier).

A final possibility for database funding is partnership with the publishing industry, who could support and finance resources as part of their publishing efforts. Many publishers have the expertise and infrastructure to host database curation in-house and could expand their existing services to publish data, alongside papers and books. This model has not yet been tested, although several companies are considering ways of implementing it. This approach would allow curation and ensure high-quality data through an in-house, expert review process; services would be centralized, thus avoiding the uncoordinated proliferation of databases, and long-term maintenance would be guaranteed.

However, it is likely that this model would involve subscription prices, which might be offset by open source publishing. Furthermore, it might encourage competition among publishing houses for the curation of specific data sets, thus making it difficult to integrate data into larger data

sets. Still, the idea holds much promise as it gives new purpose to a well-established set of scientific institutions—publishing houses—who are already exploring the possibilities of digital publishing. It also easily aligns itself with the existing division of labour between researchers, funders, editors and publishers. It will be interesting to see whether and how publishing houses will actively pursue this route.

Two key insights emerge from this review of models for long-term funding of online data, tools and resources. First, no current model is able to meet the requirements of cyberinfrastructure and data-intensive research. These requirements include users' expectation that reliable and ready-to-use data can be found in databases, which in turn implies that data are high quality and up-to-date with the latest advances in the field. Arguably, this state of affairs can only be obtained through sustained funding to maintain the infrastructure, and with professional curators responsible for ensuring that online resources are reliable and trustworthy. In addition, many funding bodies expect that data derived from experimental biology are freely accessible to all.

These expectations are the background to the increasing emphasis on 'data-intensive' science; without interoperable and freely accessible databases there is little chance to build clouds, grids and other smart tools for data analysis (Hey *et al,* 2009). Finally, funding agencies and national governments assume that cyberinfrastructure can be treated either as another branch of the research process—the value and novelty of which needs to be constantly assessed and demonstrated—or as an inexpensive service that can be outsourced to industry or users themselves. This way of treating data is a remnant of the past and is already undergoing extensive revision by funders and researchers.

Second, the current division of labour underlying scientific research is not sustainable. For one or more of the above models to work, current modes of interaction between researchers, funders, publishers, curators and editors, and their respective responsibilities, need to change. This might happen through a change in the role of publishers, who could become central to the management and dissemination of databases and related personnel. It might involve a change in the assessment of researchers' work, clearly distinguishing between their roles as data generators and data users and assigning penalties for failure to upload and maintain data in public repositories. It might involve a global change in funding policies, if science funders recognize that they need to provide targeted, long-term funding for cyberinfrastructure. Most probably, all three of these shifts will need to occur to secure the long-term sustainability of database building and curation.

ACKNOWLEDGEMENTS

We are grateful to E. Huala, S. May, N. Provart and P. Zimmerman for their contributions. I. Lavagi and M. O'Malley provided useful comments on the final draft. R.B. is funded by the BBSRC. S.L. is funded by the Economic and Social Research Council, as part of the Economic and Social Research Council Centre for Genomics in Society.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

Ashburner M *et al* (2000) Gene Ontology: tool for the unification of biology. *Nat Rev Genet* **25:** 25–29

Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: juggling between evolution and stability. *Brief Bioinform* **5:** 39–55

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* **35:** D301–D303

Brady SM, Provart NJ (2009) Web-queryable large-scale data sets for hypothesis generation in plant biology. *Plant Cell* **21:** 1034–1051

Bult CJ, Eppig JT, Kadin JA, Richardson JE, Blake JA (2008) The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res* **36:** D724–D728

Chandras C, Weaver T, Zouberakis M, Smedley D, Schughart K, Rosenthal N, Hancock JM, Kollias G, Schofield PN, Aidinis V (2009) Models for financial sustainability of biological databases and resources. *Database* doi:10.1093/database/bap017

Cochrane GR, Galperin MY (2010) The Nucleic Acid Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acid Res* **38:** D1-D4

Dwight SS *et al* (2004) Saccharomyces genome database: underlying principles and organisation. *Brief Bioinform* **5:** 9–22

Giles J (2005) Internet encyclopaedias go head to head. *Nature* **438:** 900–901

Hey T, Tansley S, Tolle K (2009) *The Fourth Paradigm Data-Intensive Scientific Discovery*. Redmond, WA, USA: Microsoft Research

Huss JW, Lindenbaum P, Martone M, Roberts D, Pizarro A, Valafar F, Hogenesch JB, Su A (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acid Res* **38:** D633–D639

Lee WH, Atienza-Herrero J, Abagyan R, Marsden BD (2009) SGC—structural biology and human health: a new approach to publishing structural biology results. *PLoS ONE* **4:** e7675

Leonelli S (2010a) The commodification of knowledge exchange: governing the circulation of biological data. In Radder H (ed) *The Commodification of Academic Research*. Pittsburgh, PA, USA: Pittsburgh University Press

Leonelli S (2010b) Packaging data for re-use: databases in model organism biology. In Morgan M, Howlett P (eds) *How Well Do Facts Travel?* Cambridge, UK: Cambridge University Press

Liang C *et al* (2008) Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res* **36:** D947–D953

Maron NL, Smith KK, Loy M (2009) *Sustaining Digital Resources: An On-the-Ground View of Projects Today*. Ithaka Case Studies in Sustainability. http://www.ithaka.org/ithaka-s-r/strategy/ithaka-case-studies-in-sustainability

O'Malley MA, Elliot KC, Haufe C, Burian RM (2009) Philosophies of funding. *Cell* **138:** 611–615

Salzberg S (2007) Genome reannotation: a wiki solution? *Genome Biol* **8:** 102

Scholl RL, May ST, Ware DH (2000) Seed and molecular resources for *Arabidopsis*. *Plant Physiol* **124:** 1477–1480

Sprague J *et al* (2008) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res* **34:** D581–D585

Swarbreck D *et al* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* **36:** D1009–D1014

Toufighi K, Brady SM, Austin R, Ly E, Provart NJ (2005) The Botany Array Resource: e-Northerns, Expression Angling, and promoter analyses. *Plant J* **43:** 153–163

Tweedie S *et al* (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* **37:** D555–D559

Wouters P, Schroeder P (2003) *The Public Domain of Digital Research Data*. Amsterdam, The Netherlands: NIKI-KNAW

Zimmermann P, Hennig L, Gruissem W (2004) Gene expression analysis and network discovery using Genevestigator. *Trends Plant Sci* **10:** 407–409

Ruth Bastow [left] is at the University of Warwick, UK.
E-mail: ruth@arabidopsis.info
Sabina Leonelli [right] is at the Economic and Social Research Council Centre for Genomics in Society, University of Exeter, UK.
E-mail: s.leonelli@exeter.ac.uk

Received 26 May 2010; accepted 26 August 2010; published online 17 September 2010

EMBO *reports* (2010) **11,** 730–734.
doi:10.1038/embor.2010.145