

*Banking on biodata: Primary data sharing in the life sciences

Science in the social networking era

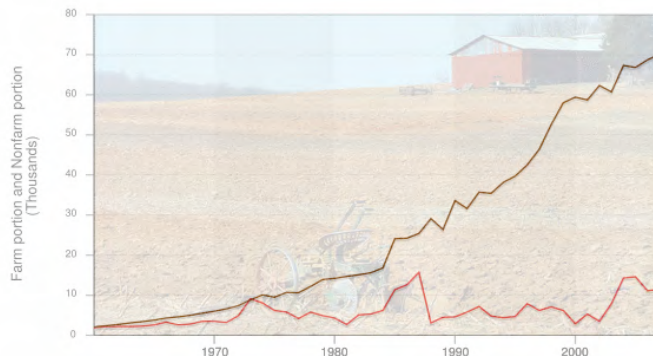
As Web 2.0 technologies continue to transform the Internet from a static library into a dynamic interactive workspace, online social networking tools are offering people around the world the chance to collaborate to share, compare and transform data. At the forefront of this movement are hugely popular websites such as [Facebook](#), [MySpace](#), and [YouTube](#), which allow members to form online communities that can share music, photos and video clips. These sites are overtaking even email as the preferred method of communication for a growing body of young Internet users, dubbed the 'MySpace generation'. Yet these technologies have much more to offer than merely helping teenagers conduct their virtual social lives, and clued-up scientists are now beginning to explore the opportunities they offer for open data sharing.

Sociable scientists

Online social networking services have the potential to make enormous amounts of scientific data freely accessible in large repositories, whether institutional or subject-based rather than being locked away in countless smaller databases. Sharing data can enable the collaboration of scientists across disciplines. For example, in the life sciences cooperation between biologists, mathematicians, and computer scientists can enable mapping of the protein and metabolic networks of organisms, leading to the creation of biological models and theoretical models of bacterial speciation¹.

Inevitably, there are pitfalls. Results can easily be misinterpreted if detailed protocols are not available alongside the raw figures. Added to which, the very concept of sharing can generate anxiety, particularly in disciplines with no previous tradition of free access. That the investigators who generate the data receive proper credit for their efforts is also a concern². Yet, provided issues such as subject confidentiality and misuse are taken seriously, making the primary data behind publications transparent could promote new insights and improve communi-

cation among scientists, benefiting everyone from the initial researchers to society at large.



'Future of the family farm' data from USDA
(<http://www.ers.usda.gov/Briefing/FarmStructure/Data/historic.htm>)
visualized by Swivel user [cindy47452](#)

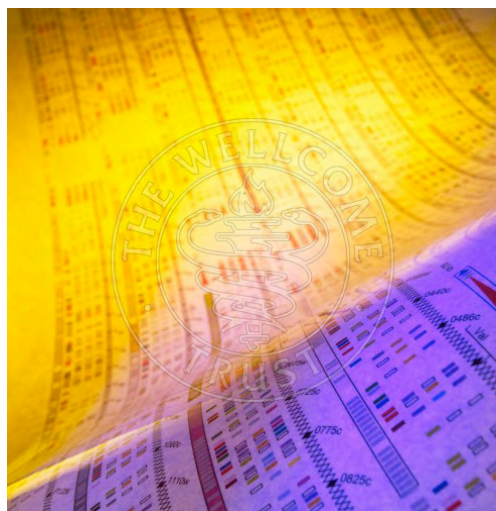
Sharing a vision

Two of the new wave of websites dedicated to making data analysis and interpretation more communal are [Swivel](#) and [Many Eyes](#). Both sites allow users to upload and visualize their own data, as well as downloading and reusing the data of others.

In pursuit of its mission "to make exploring all the world's data fun and insightful for everyone", Swivel covers a wide range of data, hosting groups as diverse as 'Water', 'Viva México', and 'Communication'. It allows uploaded data sets to be reanalysed, recombined and visualized in various forms by its users³. Since the launch of its public version in December 2006, Swivel claims that more than one million data sets have already been uploaded to the site. Other users are able to rate, comment on, email, and bookmark this data, and even add graphs or charts generated from it to their own blogs. Data sets are automatically compared to others to reveal correlations that might otherwise have been missed and the user is presented with a choice of possible comparisons and related material. Not all of the data sets are freely accessible, however, as users can pay to ensure that their results are not made publicly available on the site.

[Many Eyes](#), which was launched by IBM's Visual Communication Lab in January 2007, plans "to democratize visualization and to enable a new social kind of data analysis". To this end, it offers a large range of free visualization tools, many previously available only to experts, ranging from simple line graphs to complex 'zoomable' maps, all of which are interactive. Like Swivel, Many Eyes contains a diverse range of data sets that can be viewed, annotated, and stored in various ways by users.

Bonding over biobanks



More and more research organizations worldwide are also recognizing the value of sharing and reusing data, and are making their findings available in digital repositories⁴.

Map of the Malaria gene [Wellcome Library, London](#) C0014533

Examples include the recent initiative, part of the long-term national [UK Biobank](#) project, to monitor half a million members of the population both in sickness and in health⁴. Its findings will be held in a database, linked to records collected by the National Health Service, and requests for access to the data will be closely vetted⁴. This is just one part of a full-scale global movement, with similar projects already underway or under consideration in countries including Canada, China, Estonia, Iceland, Mexico, Norway, Singapore, and the USA^{4,5}. Pooling the data generated by these projects could create a formidable resource for medical researchers⁵. Along similar lines, an ongoing project in Japan is building a biorepository of blood and tissue samples from 300,000 citizens suffering from common diseases⁶. Once complete, academic and public-sector researchers will have access to the samples, although it is unclear whether private companies will be granted the right to use them⁶.

Another ambitious digital repository project is the US National Institutes of Health free access [Database of Genotype and Phenotype](#) (dbGaP), launched in December 2006. This web-based portal intends to archive and distribute data about the

genes, health and lifestyles of thousands of subjects studied over many years, allowing investigators to search for new associations⁷. So far, dbGaP contains data on age-related eye diseases and Parkinsonism, and findings from the Framingham Heart Study, the Genetic Association Information Network (GAIN), and numerous other sources will be added in the future⁷. In addition, the US Government was recently reported to be considering a plan to store the majority of the scientific data generated by federal agencies in publicly accessible online repositories⁸.

Patient records: Data sharing can be invaluable in medicine but will raise ethical issues of patient privacy and consent.

[Justine Desmond Wellcome Library, London](#) N0030690.



Projects have also begun to emerge that promote data sharing beyond the national level. [Tubafrost](#); The European Human Tumor Frozen Tissue Bank is a collaboration between hospitals across Europe to share data on frozen tissue specimens. A still wider collaboration is envisaged by the [Global Diversity Information Facility](#), which asks countries and international institutions to cooperate to promoting the collection and sharing of data related to biodiversity to inform policy on the subject worldwide. GDIF also employs innovative methods of displaying its data; including the provision of biological occurrence data in the required format for the Google Earth Client. This allows the mapping of one or more taxa, enhanced with web links to the GBIF home page, the species page for that taxon from the GBIF data portal and relevant Google images.

As noted at a recent Wellcome Trust conference, [From Biobanks to Biomarkers](#), for data to be reliable, extensive collaboration between different biobanks as well as the standardisation of protocols for data collection, sample storage, analysis

and access would be vital. Legal questions of intellectual property and patent law, as well as ethical questions of the extent of patient consent, and practical considerations of the applicability of data sets to issues in health care, would also need to be addressed.

Closing the generation gap

Like so many offshoots of Web 2.0, large-scale online data sharing is still in its infancy, and mainstream science has barely scratched the surface of its capabilities, yet alone begun to address the associated legal and ethical issues. The outcomes of current endeavours, such as the biobank projects, will be crucial in determining public confidence in sharing personal data as well as evolving standardised protocols for collecting, storing, and accessing such information. As the MySpace generation rise through the ranks of scientists and policy makers, sharing primary data on an international basis is likely to become an accepted part of research in life sciences and beyond.



Generation games: sharing data online will come naturally to the next generation of scientists. Anthea Sieveking Wellcome Library, London
AS0007501F04AS0007501F04

References

1. Dedeurwaerdere, T. The Institutional Economics of Sharing Biological Information . <http://www.cpdr.ucl.ac.be/docs/tom/TDInstitutionalEconSharingBiIn.pdf> (accessed 7 May 2007).
2. Editorial (2006) A fair share. *Nature* 444, 653–654.
3. Butler, D. (2007) Data sharing: the next generation. *Nature* 446, 10–11.
4. Giles, J. (2006) Huge biobank project launches despite critics. *Nature* 440, 263.
5. Pincock, S. (2006) Biobank project finally underway. *The Scientist* [<http://www.the-scientist.com/news/display/23228/>].
6. Triendl, R (2003) Japan launches controversial biobank project. *Nature Medicine* 9, 982.
7. Russo, G. (2006) NIH offers free access to wealth of disease data. *Nature* 444, 982.
8. Butler, D. (2007) Agencies join forces to share data. *Nature* 446, 354.

First published 17/05/07