My data are your data

Encouraging more broad and inclusive data sharing in today's world will involve concerted community efforts to overcome technical barriers and human foibles. Vivien Marx investigates.

In January, over 50 researchers from 30 academic and commercial organizations agreed on a standard for describing data sets. The BioSharing initiative, comprising both researchers and publishers, launched the Investigation-Study-Assay (ISA) Commons, which promises to streamline data sharing among different databases¹. Life scientists have thousands of databases, over 300 terminologies and more than 120 exchange formats at their disposal, says BioSharing co-founder Susanna-Assunta Sansone of the University of Oxford. In this era of collaborative big science, researchers only move forward by “walking together.”

Although increased data sharing is central to scientific progress and is attracting attention from many quarters², standards are only some of the stars that must align to make it possible.

Share and share alike?

Oversharing is embarrassing in social media but sharing is always a virtue for scientists. Although many scientists embrace the idea of sharing data in research, few manage it in practice. The traditional sharing method—the research paper—has made the transition online through HTML and PDF formats, but is becoming outmoded and unwieldy as life science research generates increasingly varied types of big data sets to be associated with a claim. These data sets quickly pile up as high-throughput instruments spray a fire hose of data in giga- or terabyte-sized files. Data are also getting an ever longer tail: the provenance of claims, hypotheses and arguments, along with supporting metadata, methods, software code and tools, multimedia, workflows and models. These challenges are spurring initiatives, such as BioSharing and several other nonprofit and commercial projects, to make sharing easier and more palatable (**Supplementary Table 1**).

Sharing “isn't just a good idea, it's essential,” says John Quackenbush of the Dana-Farber Cancer Institute and Harvard School of Public Health in Boston. “The progress of science has been slowed by failure to share data and methods,” he says. A physicist turned computational biologist, Quackenbush builds analysis tools and platforms, for example, for the Lung Genomics Research Consortium, which let users crunch through high-dimensional data sets (**Box 1**). Besides the risk of duplicating

research, “it's also about discoveries and breakthroughs that would be entirely missed if data and methods aren't shared,” says David De Roure of the University of Oxford, who strategizes on the future of research methods for the UK's e-Science program.

Fair share?

Now that high-throughput devices, such as next-generation sequencers, let laboratories generate mountains of data overnight, biomedical researchers have plenty to feed computers down the hall or in a cloud. Even scientists without technology access can use their colleagues' data, commonly shared through online repositories, such as the US National Institutes of Health (NIH) National Center for Biotechnology Information (NCBI).

In the mid-1990s, NCBI's GenBank, one of the first public repositories to hold DNA sequences, was disseminated on a CD that came in the ‘snail’ mail. Today, researchers download dozens of terabytes of information a day from NCBI's resources, including GenBank, says NCBI director David Lipman. He acknowledges that some scientists might not always quickly deposit all their genome or expression data. But one positive move does follow another in a “virtuous circle,” he says.

Researchers harvest data, analyses, software tools and insight from papers and online portals in hunting-and-gathering activity that is a “hodgepodge of methods” that functions well enough, says Steven Salzberg from the Johns Hopkins University School of Medicine. One problem, he cautions, is that the number of studies providing all the ingredients needed to reproduce a study's findings “is still relatively small.” Thus, the consistent and full sharing of data and methods still isn't the norm^{3,4}.

Scientists' motivation to share has always been mixed, according to Bob Campbell, editor at Wiley-Blackwell in Oxford and pub-

lishing historian. “Newton was famously secretive,” he says. To claim credit, researchers will switch into sharing mode. “Darwin delayed publication until he realized that Wallace was on to the same idea.”

Expanding the outlook

What data sharing in science looks like tomorrow may be utterly unlike today's practices. De Roure promotes abandoning the idea that “knowledge has to be exchanged in paper-sized chunks.” He points to the Force11 group, formed in 2011 by scientists, librarians, research funders and editors at several publishing houses, which strives for “semantically enhanced, media-rich digital publishing.”

The group's ‘manifesto’ published last fall recommends that we “rethink the unit and form of the scholarly publication,” to move beyond the electronic facsimiles of an ink and paper journal publication toward digitally enabled complex entities. A world of “networked knowledge objects” may seem a futuristic scenario, but efforts now underway to give data, methods and metadata their own web-enabled voice already

tug science-sharing in that direction⁵.

Some researchers are pushing publishers for computational access to the scientific literature so machines can harvest and share findings more quickly than human readers. These ventures head into choppy seas. Since beginning their

Relic of the past? Data are trying to jump off shelves. (istockphoto).

Genocoding Project in 2009, Maximilian Haeussler of the University of California at Santa Cruz and his colleague Casey Bergman, genomicist at the University of Manchester, UK, have received a frosty reception from publishers^{6,7}.

Haeussler's software tool scans papers for genomic identifiers and maps them to the human genome. In a log of responses to permission requests (<http://text.soe.ucsc.edu/>), the Genocoding site currently lists only 5 consents from publishers; 12 are thinking about it and 26 have issued outright denials.

Tune into my network

Systems biology researchers already exchange models, sequences and software code about networks of interacting genes and proteins, but “we can't store working hypotheses,” says Trey Ideker at the University of San Diego, who developed Cytoscape software to analyze and visualize biological networks.

Researchers include network models in papers, many of which inevitably get stuffed into the supplementary online sections of publications. But only a “very small set of people” will discover that additional material and “an even smaller set” will write a software script to pull that model into their own, he says.

To explore new ways of exchanging models, Ideker is involved with Seattle-based Sage Bionetworks, a nonprofit that reaches out to the scientific community to encourage sharing. Founded in 2009 by former Merck researchers Stephen Friend and Eric Schadt, Sage is an open repository for different kinds of data and network models, says Ideker. Separately, Ideker's team is working on a Cytoscape-based database for sharing network models.

Sharing a special sauce

Through web portals, computational biolo-

gists share software tools and code in their papers, and they create web services, often installed on clouds, where users can run others' software. Cloud computing standardizes the hardware, remedying one challenge in software sharing, according to Ideker. Scientists using a cloud can no longer say, for example, “I could not get your software to run on my version of Unix,” he says.

Cloud or no cloud, Salzberg says that reproducing some types of *in silico*-based analyses remains a challenge. In a paper comparing leading genome-assembly programs, he and his group shared software and data. He included details about his software methods, “the ‘special sauce’ you need to run these assembly programs and reproduce our results,” and which can be quite complex. The recipes are also on his laboratory's website and in a journal supplement⁸.

In an independent effort called the Assemblathon, a sort of bake-off of genome

assembly programs, details were lacking on how the assemblers were run, “making it impossible to reproduce those computational experiments,” Salzberg says. This absence of pertinent information is the norm, in his view.

Sharing the “special sauce” of methods along with data and metadata is “still an enormous problem,” says Emory University computational biologist James Taylor, who co-developed the Galaxy platform of open-source sequence analysis tools, which can be downloaded or used on the cloud. And although publications lack fully detailed methods, policing that gap is too burdensome for reviewers, he says. Galaxy has an idea, called Pages, to stave off that chore for scientists.

Scientists are starting to use Galaxy Pages, a functionality for sharing data and workflow analysis steps. A scientist reading a published paper can visit the study's Galaxy Page to plug different data into the computational workflow

Box 1 Letting data speak

Web inventor Tim Berners-Lee, who is at the Massachusetts Institute of Technology's computer science and artificial intelligence laboratory in Cambridge, Massachusetts, and also directs the World Wide Web Consortium (W3C), is wont to say that the semantic web—the web structure that links not just documents but also disparate data and other entities—will profoundly change how scientific knowledge is produced and shared.

For the semantic web to help scientists share, computers must parse papers alongside humans¹⁰. Machines struggle with this homework because “free text is a nightmare of ambiguity,” as Barend Mons from Leiden University Medical Center in The Netherlands points out. To promote web-based semantic sharing, he co-founded an initiative called Concept Web Alliance. The organization with scientist members from many countries, aims to leverage semantic web approaches to organize knowledge into uniquely identified assertions, one example of which would be ‘A inhibits B’¹¹.

One idea that amplifies data's voice is one that Mons and colleagues at the Netherlands Bioinformatics Centre are pushing into prototype. The approach equips assertions with a digital object identifier (DOI), a tag that has mainly been reserved for journal articles. This tag converts assertions into nanopublications, units that can be shared and which map back to provenance information contained, for example, in journal articles.

Yield from a nanopublications query could be a stack of research results, a few words each, from scores of published studies about a pathway, or a bundle of hypotheses on a particular gene-protein interaction. Nanopublications are being tested in Open Pharmacological Concepts Triple Store (Open PHACTS), the knowledge management side of a European collaboration between pharma and academia called Innovative Medicines Initiative.

Having computers process research “paves the way to machines giving us greater assistance in our work,” says De Roure. When it comes to nanopublications, however, he is less comfortable codifying research into a common framework, because imposing one style of discourse and reasoning on a research problem could restrict insight. Goble, who also withholds praise, takes practical issue with nanopublications: DOIs cost money to set up and maintain, although

costs vary depending on the DOI-issuing organization. Given the vastness of scientific literature, producing billions of DOIs may well “break the bank,” Goble says.

BGI's Scott Edmunds in Shenzhen, China, likes the concept of offering microattribution to contributors, which appears to increase database submissions, he says. Yet he shares concern about splitting the literature into a huge number of assertions. As an alternative, “data citation is a very easy first step,” he says. Data sets can be converted into citable, sharable entities through DOIs, which is not a major technical or cultural step, as the infrastructure is already in place, says Edmunds, who is also the editor of the new life sciences journal *GigaScience*. Datacite, an international organization of libraries, is making DOIs available for data sets in the journal. It stands to “create incentive for authors to share, as they will get this additional form of credit,” he says.

Although the Protein Data Bank, managed by Rutgers, The State University of New Jersey in New Brunswick, New Jersey, and the University of California at San Diego, issues DOIs for protein structures, these identifiers have not been used much as citable units, Edmunds explains. What is different with data set DOIs is that citation indices have agreed to begin tracking them this year. This arrangement can lessen researchers' fears about sharing.

Self-interest and technical curiosity motivate BGI to explore new ways to share data. There is no point to producing petabytes of data unless people use them, which requires having data in accessible form, says Edmunds. BGI's benefit will come from routinely moving terabytes of sequencer output and other kinds of high-dimensional data sets. *GigaScience* draws on BGI's massive cloud-based storage capacity. The idea of sharing truckloads of data with their papers stands to give researchers a giddy case of petabyte whiplash.

While Quackenbush is intrigued by BGI's approach, he points out necessary limitations to data sharing. Even short snippets of genomic sequences have been shown to identify individuals. He explains “that almost any data with the suffix ‘-seq’ from humans is now defined as identifiable, limiting the ways in which we can provide access.” There is a fine ethical line to walk between sharing and oversharing.

and make high-throughput analysis interactive. “You can play with it, change the parameters, edit the workflows, they are now yours,” says Penn State University biologist Anton Nekrutenko, Galaxy’s co-developer. Through Galaxy, laboratories can find methods that others successfully use, says Taylor. He and Nekrutenko agree that genome assembly is particularly challenging and echo Salzberg’s call for sharing methods detail.

Not sharing workflows can lead research down blind alleys or worse, biomedical wild goose chases, when an experiment yields rare variants, which might or might not be correct, he says. He points to a human resequencing project focused on heteroplasmy, which is when the mitochondrial genomes in one organism’s cells or tissues show several variants. Nekrutenko and his Penn State colleagues showed that misalignment led to erroneous findings about heteroplasmies at certain spots in the genome from nine examined individuals in three families⁹. Galaxy wants to help researchers avoid such pitfalls. Sharing methods and tools can lead to community-based decisions on best practices, says Taylor.

When pipelines creak

To make analysis steps sharable, some laboratories want to embed clickable computational workflows into a paper. To do so, a journal article must be able to link to remote computing and data resources. Workflows have mixed stewardship in their orchestration of local and remote computational resources operated by different groups and companies, says University of Manchester computer scientist Carole Goble, whose team has built many web-based sharing platforms for the life sciences. Dashing through its paces, a workflow pipeline may break: software on a site crashes, a web service changes data format, funding lapses put a web service offline, she says. A sharable, reproducible workflow becomes unsharable and irreproducible.

With their platform Workflow 4Ever, Goble and colleagues across Europe want to lend longevity to methods sharing by preserving workflows. If a pipeline breaks, a record shows the previous iteration that ran smoothly. It avoids time spent on glitch hunting. Researchers know where to tweak the workflow.

The human factor

Better ways are needed to encourage sharing and recognize the effort of those producing data and methods and subsequently sharing them, says Quackenbush, lamenting that these efforts are still seen as the ugly stepchildren of modern science. Portals for data or methods sharing often suffer because they are built by “data geeks” who do not keep users in mind, he says.

Platform builders must “respect people’s

fears,” says Goble. When ready, researchers will share with trusted collaborators, but commands to share can fall on deaf ears. She prefers to credit people for sharing. Not only must these environments be “safe havens,” they should let scientists build “a reputation economy” around the objects being shared, so others see who shares how much, she says.

Quackenbush and his team designed the data coordination and analysis center for the Lung Genomics Research Consortium, in which teams across the United States characterize a biological sample collection for hints about lung disease. Although the LGRC research portal is open, it does not offer sequence-based data, which must go to the NIH database of Genotypes and Phenotypes, where controlled data access policies are in place and individuals are de-identified.

His group chose analytical methods and presentation formats that let users edit genomic content without having to be computer scientists. “It’s not intelligence but rather empathy that drives a good user experience,” he says.

Users who are computational and statistical scientists prefer doing analysis right away. But they like filtered data chunks, so the team created a “shopping cart” to allow this filtering. Nonquantitative scientists looking for specifics about a gene or a pathway are the largest group of users. For them, the team made decisions about the major phenotypes and present “reasonable summaries of the data, based on looking at what people wanted to know,” Quackenbush says. His team expanded a commercial tool, ClinicalSense, from the software and consulting firm ID Business Solutions headquartered in Guildford, UK, to organize clinical data, so users can define phenotypic groups on the fly if they do not like predefined types.

The sample and data tracking system makes transparent to group members who has samples and who produces data, he says. One of the “objective measures of success” is sharing data by depositing it in the portal. Report season in this collaborative effort shows who is meeting production targets, which the project scientists set themselves.

“We’re all human, so as you might expect, data deposition spiked before calls and in-person meetings,” Quackenbush says. The system’s transparency has a number of advantages. The team can quickly see a problem when array data do not match a patient’s exome sequence profile. “So our ‘due diligence’ in measuring output had some very positive side effects,” he says.

Goble adds an ingredient to her platforms that she calls “sharing creep.” In the web-based platform myExperiment, users share workflows or keep them private. First, a scientist must be

ready to deposit information, which is a step reluctantly taken, as it is the “expensive” and “undervalued” task of sharing and describing what is deposited. A follow-on chapter is persuading scientists to share with others. A project will garner few joiners if it begins by proclaiming, in the name of “open science fundamentalism,” that everything must be open, Goble says.

She and her team also expanded on a commercial tool: the spreadsheet. When developing a data-sharing platform called SysMo-SEEK for the European Systems Biology of Microorganisms program, she and her team made the spreadsheet readable by people and machines. The team equipped spreadsheets with objects that adhere to minimum information models for different ‘omics classes. When an experimentalist enters data, the spreadsheet cells offer drop-down menus with terms, standardizing categories and making them easier to share. The tool, called RightField, does “semantic annotation by stealth,” Goble says.

Users need not bother with the fact that beneath the spreadsheet sits a controlled vocabulary that uses the same terms to describe the same things, an important element of sharing. The vocabulary is connected to BioPortal, a repository run by the US National Center for Biomedical Ontology, an NIH-funded consortium that is part of the National Centers for Biomedical Computing.

RightField reflects Goble’s strategy of building “ramps” around tools already in use. That approach gets scientists to “just do that little bit more” to enrich research assets so they can be more readily shared. Rather than ask people to fill out a six-page sharing template, one needs need “spoons full of sugar to make the medicine go down.”

Vivien Marx, New York

Corrected after print 7 December 2012

Note: Supplementary information is available at in the online version of the paper.

1. Sansone, S.-A. *et al. Nat. Genet.* **44**, 121–126 (2012).
2. National Science Board Task Force on Data Policies. *Digital Research Data Sharing and Management* (NSF, December 2011). <http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf>
3. Tenopir, C. *et al. PLoS ONE* **6**, e21101 (2011).
4. Carpenter, J., Tanner, S., Smith, N. & Goodman, M. *Researchers of Tomorrow: a Three Year (BL/JISC) Study Tracking the Research Behaviour of ‘Generation Y’ Doctoral Students*. Annual Report (British Library/Joint Information Systems Committee, May 2011).
5. De Roure, D., Bechhofer, S., Goble, C. & Newman, D. Scientific social objects: the social objects and multidimensional network of the myExperiment website. Presented at 1st International Workshop on Social Object Networks (SocialObjects 2011), October 2011, Boston, MA, US.
6. Van Noorden, R. *Nature* **483**, 134–135 (2012).
7. Anonymous. Gold in the text? *Nature* **483**, 124 (2012).
8. Salzberg, S.L. *et al. Genome Res.* **22**, 557–567 (2012).
9. Goto, H. *et al. Genome Biol.* **12**, R59 (2011).
10. Altman, R. *et al. Genome Biol.* **9** (suppl. 2), S7 (2008).
11. Mons, B. *et al. The value of data. Nat. Genet.* **43**, 281–283 (2011).

Erratum: My data are your data

Vivien Marx

Nat. Biotechnol. 30, 509–511 (2012); doi:10.1038/nbt.2243; published online 7 June 2012; corrected after print 7 December 2012

In the version of this article initially published, the statement on p. 509, column 2, top paragraph, was missing a few words, making it a fragment: “Besides the risk of duplicating research discoveries that would be missed if data and methods aren’t shared,” should have read, “Besides the risk of duplicating research, it’s also about discoveries and breakthroughs that would be entirely missed if data and methods aren’t shared’...” On p. 511, column 2, paragraph 3, it was stated that the Lung Genomics Research Consortium (LGRC) shares data within the consortium but not with a wider community. In fact, that is not the case, and the sentence now reads, “Although the LGRC research portal is open, it does not offer sequence-based data, which must go to the NIH database of Genotypes and Phenotypes, where controlled data access policies are in place and individuals are de-identified.” Also on p. 511, in column 3, paragraph 2, the European Systems Biology of Microorganisms program was incorrectly identified as SysMo-SEEK. It is SysMo. SysMO-Seek is a data-sharing platform developed for the 13 consortia that make up SysMo. The sentence now reads, “When developing a data-sharing platform called SysMO-SEEK for the European Systems Biology of Microorganisms program...”. The errors have been corrected in the PDF and HTML versions of this article.

Erratum: Beyond counting tumor cells

Jim Kling

Nat. Biotechnol. 30, 578–580 (2012); doi:10.1038/nbt.2295; published online 9 July 2012; corrected after print 7 December 2012

In the version of this article originally published, On-Q-It platform was called by its previous and outdated name, C5 (Circulating Cancer and Characterization Chip). It should have been called OnQChip. The error has been corrected in the HTML and PDF versions of the article.

Erratum: Antibody specification beyond the target: claiming a later-generation therapeutic antibody by its target epitope

Colin G Sandercock & Ulrich Storz

Nat. Biotechnol. 30, 615–618 (2012); doi:10.1038/nbt.2291; published online 9 July 2012; corrected after print 7 December 2012

In the version of this article initially published, in Table 1 “Patents claiming anti-TNF α antibodies,” patent US 7012135 should have been assigned to UCB Pharma and not UC Berkeley. The error has been corrected in the PDF and HTML versions of this article.

Erratum: Drug pipeline: Q212

Laura DeFrancesco

Nat. Biotechnol. 30, 817 (2012); doi:10.1038/nbt.2359; published online 10 September 2012; corrected after print 7 December 2012

In the version of this article initially published, the Thrombogenics drug Jetrea (recombinant microplasmin) was listed in “Notable upcoming regulatory decisions” as indicated for thyroid cancer. It should have read vitreomacular adhesion. The error has been corrected in the HTML and PDF versions of the article.

Corrigendum: Pharmacogenomics in clinical practice and drug development

Andrew R Harper & Eric J Topol

Nat. Biotechnol. 30, 1117–1124 (2012); doi:10.1038/nbt.2424; published online 8 November 2012; corrected after print 7 December 2012

In the version of this article initially published, the funders were not acknowledged. The funding statement should have read “E.J.T. was funded by the National Institutes of Health/National Center for Advancing Translational Sciences TR000109.” The error has been corrected in the HTML and PDF versions of the article.