Published in final edited form as:
Science. 2009 October 9; 326(5950): 234–236. doi:10.1126/science.1180598.

'Omics Data Sharing

Dawn Field[1,*,†,‡], **Susanna-Assunta Sansone**[1,2,†], **Amanda Collis**[3,†], **Tim Booth**[1], **Peter Dukes**[4], **Susan K. Gregurick**[5], **Karen Kennedy**[6], **Patrik Kolar**[7], **Eugene Kolker**[8], **Mary Maxon**[9], **Siân Millard**[10], **Alexis-Michel Mugabushaka**[11], **Nicola Perrin**[12], **Jacques E. Remacle**[7], **Karin Remington**[13], **Philippe Rocca-Serra**[12], **Chris F. Taylor**[12], **Mark Thorley**[14], **Bela Tiwari**[1], and **John Wilbanks**[15]

[1]U.K. Natural Environment Research Council (NERC), Environmental Bioinformatics Centre.

[2]European Molecular Biology Laboratory (EMBL) Outstation, The European Bioinformatics Institute (EBI).

[3]U.K. Biotechnology and Biological Sciences Research Council.

[4]U.K. Medical Research Council.

[5]U.S. Department of Energy.

[6]Genome Canada and Wellcome Trust Sanger Institute.

[7]Unit for Genomics and Systems Biology, European Commission.

[8]Seattle Childrens Hospital.

[9]Marine Microbiology Initiative, Gordon and Betty Moore Foundation.

[10]U.K. Economic and Social Research Council.

[11]European Science Foundation.

[12]The Wellcome Trust.

[13]U.S. National Institute of General Medical Science, NIH.

[14]NERC.

[15]Science Commons.

Development of high-throughput genomic and postgenomic technologies has caused a change in approaches to data handling and processing (1). One biological sample might be used to generate many kinds of "big" data in parallel, such as genome sequence (genomics), patterns of gene and protein expression (transcriptomics and proteomics), and metabolite concentrations and fluxes (metabolomics). Extensive computer manipulations are required for even basic analyses of such data; the challenges mount further when two or more studies' outputs must be compared or integrated.

Grassroots movements (2-5), efforts including the Science Commons, which is initiating an open-access data protocol (6), as well as top-down (funder-led) efforts (see table, page 235), have led to a range of policies for data management and sharing. A recent European Science

[‡]Author for correspondence. dfield@ceh.ac.uk.
[*]Full author affiliations are available on *Science* Online.
[†]These authors contributed equally to this article.

Supporting Online Material
www.sciencemag.org/cgi/content/full/326/5950/234/DC1

Foundation consultation exercise confirmed a lack of explicit, well-documented data-sharing policies for most funding agencies in European countries (7). If we are to avoid squandering the immediate and extended value of big data, a focused strategy will be pivotal.

Early policies were driven by the need to manage long-term data sets (those accrued over 30 or more years), such as those in the social and environmental sciences. More recently, policies have emerged in response to increased funding for high-throughput approaches in major 'omics fields. The European Commission has invited the member states to develop policies to implement access, dissemination, and preservation for scientific knowledge and data (8).

Beyond public and private funding agencies, regulatory agencies such as the U.S. Food and Drug Administration (FDA) (9), European Medicines Agency (EMEA) (10), and U.S. Environmental Protection Agency (EPA) (11) are also working to define guidelines to facilitate electronic submission of traditional and 'omics data types. These, as well as industry guidelines, are beyond the scope of this document, but much could be learned from an exchange of ideas and practices (12).

The policies listed here share common principles. They aim to protect cumulative data outputs. All recognize data as a public good and data sharing as a way to accelerate subsequent exploitation. On a practical level, all acknowledge the right of first use for data providers and the right to appropriate accreditation. Likewise, these policies have been generated through the same basic process (table S1) (13).

Despite these commonalities, there is still room for heterogeneity, as expected, given the different types of communities served by each funder and the data types they generate. Care must be taken, though, that these differences do not impede seamless interoperability. The path a funding agency takes in supporting its data policy largely reflects the relative emphasis placed on managing versus sharing data. A focus on managing is often accompanied by an institutional infrastructure. Such centralization provides economy of scale, institutional memory, and reusable capability, but it also incurs a substantial direct cost that may compete with research funding (14). The UK Natural Environment Research Council (NERC) sustains a system of national data centers and has invested in the NERC Environmental Bioinformatics Centre (NEBC) to cover 'omics data (15, 16). Similarly, the UK Economic and Social Research Council provides a central data service for social scientists (17). Policies that focus on sharing tend to place more responsibility on researchers. For example, the UK Biotechnology and Biological Sciences Research Council (BBSRC) is supporting its data-sharing policy through funds that allow researchers to develop their own solutions from the bottom up.

Massive-scale raw data must be highly structured to be useful to downstream users. Standardized solutions are increasingly available for describing, formatting, submitting, and exchanging data (18, 19). These reporting standards include minimum information checklists, ontologies, and file formats. Minimum information checklists are simple, structured documents that reflect the consensus view of a community on the information to report about particular kinds of biological studies or instrument-based assays. Ontologies provide terms needed to describe the minimal information requirements. File formats define a shared syntax to transmit and exchange standardized information.

> Data sharing, and the good annotation practices it depends on, must become part of the fabric of daily research for researchers and funders.

There are now an escalating number of community-developed checklists, ontologies, and file-format projects, a positive sign of community engagement. But this proliferation brings with it new sociological and technological challenges—creating interoperability and avoiding unnecessary overlaps and duplication of efforts. These projects largely focus on a particular technology or a specific biological knowledge domain (e.g., ontologies for anatomy, gene functions, or the environment) and are by nature fragmented and not designed to be interoperable. A range of activities are fostering harmonization and consolidation of these standards for checklists (5), ontologies (4), and representation of information in electronic formats (2, 3).

Many large coordinative initiatives (20-23) are working to address the problem of archiving and integrating data. The ELIXIR project (22) aims to construct and operate a common, sustainable bioinformatics research infrastructure to support the life sciences across Europe. The Infrastructure for Spatial Information in the European Community (INSPIRE) directive requires that Europe binds together its geospatial data into portals (23). Widely useful are initiatives like the Digital Curation Centre (DCC), which tracks data standards, documents best practice, and has published a data life-cycle model to underpin long-term data-preservation policies (24).

Achieving Adherence

Community adherence would be automatic if guidelines aligned with prevailing scientific culture and (emergent) practice. However, there is often a gulf or even outright resistance (25, 26).

Policies that stipulate public data release, especially of prepublication data, raise researchers' concerns about loss of intellectual ownership—for example, by compromising chances to publish, to commercialize aspects of funded work, or to collaborate with industry. Public release of 'omics data has also been complicated by the increasing use of human subjects (27) in medical-related studies and the resulting ethical issues. Funding agencies must allay fears that data could be reused without permission or due recognition by clarifying the agency's expectations. There is currently no large-scale infrastructure ready to support data citations, but interest in this issue is growing (28).

Researchers may be limited in their ability to comply by inadequate resourcing; time-inefficient data management at the local or community level; or a lack of tools, databases or informatics expertise. Researchers must now incorporate the cost of this type of essential work into research grants effectively and consistently, and an expert pool of scientists with the requisite skills must be developed, as well as a community of biocurators (29, 30). Mechanisms for crediting data generators when their data sets are published or reused would help justify making the data public in the mind of the researcher, especially if funding decisions took into account prior good practice.

Collecting, holding, and disseminating electronic data are substantial undertakings, if considered at the global level. If policies are to be successful, information superhighway infrastructure must be built. This must involve the creation and adoption of appropriate standards that enable electronic data to be shuttled around, tools for doing the actual task, and world-class database infrastructure to hold the collective submissions. Journals, for example, will only require compliance with reporting standards when appropriate standards-compliant software tools and public repositories become available (31). An exemplar project already exists, the Investigation/Study/Assay (ISA) Infrastructure, which is developing standards to enable freely available tools that encompass several 'omics technologies and facilitate curation and reporting at the community level (3, 32). Lack of funding for these activities has already been highlighted (33, 34), and new ways of balancing streams of funding for the generation of novel data versus the protection of existing data must be found.

The Future

We recommend that a single, brief, high-level consensus guideline serve as a template for policy documents at the funder, community, and project levels. At its heart should be the public and timely release of data. It should be based on the principle that funders and the research community must work together to develop best practice. On enforcement of policy, we suggest that, in addition to mandating the inclusion of data-sharing plans in grant applications, deposition of supporting (or ideally, all) data in appropriate databases be the rule within a specified time period in accordance with international standards. This would uphold and extend the model of “accession number for publication” that has worked well for DNA sequence data (27). “Appropriate” databases, by definition, should be secure, should be publicly accessible, and ought to have a long-term funding horizon. This allows reviewers to focus on the science, while creating a simple way to check compliance via a URL. When funders do not have a suitable database or repository to endorse, they should attempt to find or fund one (14).

We created the BioSharing Web site to centralize and to give a higher profile to bioscience data policies and standards (35). It offers a focal point for stakeholders in data policy (i) by providing a “one-stop shop” for those seeking data policy documents and information (including information about the standards and technologies that support them) and (ii) by encouraging exchange of ideas and policy components among funders, and between funders and potential fundees. For example, a recent post covers the “Toronto” (36) and “Rome” data-sharing meetings (37) that aimed to build upon the highly influential Bermuda Principles (38) and the Fort Lauderdale report (39). Ideally, this hub could spark the formation of a Bio-Sharing Consortium that would work at the global level to build essential linkages between funders and awardees and among the main research groups.

References and Notes

1. Big Data special issue. Nature. 2008; 455:1. [PubMed: 18769385]
2. Jones AR, et al. Nat. Biotechnol. 2007; 25:1127. [PubMed: 17921998]
3. Sansone SA, et al. OMICS. 2008; 12:143. [PubMed: 18447634]
4. Smith B, et al. Nat. Biotechnol. 2007; 25:1251. [PubMed: 17989687]
5. Taylor CF, et al. Nat. Biotechnol. 2008; 26:889. [PubMed: 18688244]
6. Protocol for implementing open access data. http://sciencecommons.org/projects/publishing/open-access-data-protocol
7. European Science Foundation (ESF). Shared Responsibilities in Sharing Research Data: Policies and Partnerships. ESF; Strasbourg, France: 2008. Report of an ESF–Deutsch Forschungsgemeinschaft workshop, Padua, Italy, 21 September 2007
8. European Commission (EC). On scientific information in the digital age: Access, dissemination and preservation. http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf
9. FDA. Genomic data submission. www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083641.htm
10. EMEA. Guideline on Pharmacogenetics Briefing Meetings. www.emea.europa.eu/pdfs/human/pharmacogenetics/2022704en.pdf
11. EPA. Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA. www.epa.gov/osa/genomics.htm
12. Pistoia vision. www.pistoiaalliance.org/

13. OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD; Paris: 2007. Organisation for Economic Co-operation and Development. www.oecd.org/dataoecd/9/61/38500813.pdf

14. Tiwari B, Field D, Snape J. Nature. 2006; 439:912. [PubMed: 16495972]

15. Field D, Tiwari B, Snape J. PLoS Biol. 2005; 3:e297. [PubMed: 16089508]

16. Field D, et al. Nat. Biotechnol. 2006; 24:801. [PubMed: 16841067]

17. Economic and Social Data Service. www.esds.ac.uk/

18. Field D, Sansone SA. OMICS. 2006; 10:84.

19. Standardizing data. Nat. Cell Biol. 2008; 10:1123. [PubMed: 18830215]

20. Cancer Biomedical Informatics Grid, (caBIG). National Cancer Institute, NIH; http://cabig.cancer.gov

21. Biomedical Informatics Research Network. www.nbirn.net/

22. ELIXIR. http://www.elixir-europe.org

23. EC. INSPIRE Directive. http://inspire.jrc.ec.europa.eu/index.cfm

24. DCC. www.dcc.ac.uk/

25. Thomas C. Science. 2009; 324:1632. [PubMed: 19556479]

26. Wiley S. Scientist. 2009; 23:33.

27. Pennisi E. Science. 2009; 324:1000. [PubMed: 19460974]

28. Earth System Science Data. www.earth-system-science-data.net/

29. Howe D, et al. Nature. 2008; 455:47. [PubMed: 18769432]

30. International Society for Biocuration. www.biocurator.org

31. Barsnes H, et al. Nat. Biotechnol. 2009; 27:598. [PubMed: 19587657]

32. Investigation/Study/Assay (ISA). Infrastructure for Managing Experimental Metadata. http://isatab.sf.net

33. Brooksbank C, Quackenbush J. OMICS. 2006; 10:94. [PubMed: 16901212]

34. Merali Z, Giles J. Nature. 2005; 435:1010. [PubMed: 15973369]

35. Biosharing. http://biosharing.org/

36. Toronto International Data Release Workshop Authors. Nature. 2009; 461:168. [PubMed: 19741685]

37. Schofield PN, et al. Nature. 2009; 461:171. [PubMed: 19741686]

38. Summary of principles agreed at the First International Strategy Meeting on Human Genome Sequencing; Bermuda. 25 to 28 February 1996; Singapore: Human Genome Organisation; 1996. available at www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml

39. Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility; 14 and 15 January 2003; Fort Lauderdale, FL: Wellcome Trust; 2003. 2003available at www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf

Examples of data policies from major funding agencies in the United States and United Kingdom

Funders are listed by the first year in which they made their policy public (in parentheses if a newer version of the policy exists). The NSF document is the Grant General Award document, rather than a formal policy. The DOE example is a program-level policy, as an agency-level policy does not yet exist.

Funding body	Country	Year	Policy information
Economic and Social Research Council (ESRC)	UK	(1994) 2000	www.esrcsocietytoday.ac.uk/ESRCInfoCentre /Images/DataPolicy2000_tcm6-12051.pdf
Natural Environment Research Council (NERC)	UK	(1996) 2008	www.nerc.ac.uk/research/sites /data/policy.asp
National Science Foundation (NSF)	US	2001	www.nsf.gov/pubs/2001/gc101 /gc101rev1.pdf
National Institute of Health (NIH)	US	2003	http://grants.nih.gov/grants/policy /data_sharing/
Gordon and Betty Moore Foundation (GBMF)	US	(2005) 2008	www.moore.org/docs/GBMF_Data%20Sharing%20 Philosophy%20and%20Plan.pdf
Genome Canada	Canada	(2005) 2008	www.genomecanada.ca/medias/PDF/EN/DataRelease andResourceSharingPolicy.pdf
Medical Research Council Data Sharing and Preservation Policy (MRC)	UK	2006	www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/ Datasharinginitiative/Policy/index.htm
Biotechnology and Biological Sciences Research Council (BBSRC)	UK	2007	www.bbsrc.ac.uk/publications/policy/ data_sharing_policy.html
Wellcome Trust	UK	2007	www.wellcome.ac.uk/About-us/Policy/Policy-and -position-statements/WTX035043.htm
Department of Energy (DOE)	US	2008	http://genomicsgtl.energy.gov /datasharing
European Commission	Europe	NA	Issued Communication calling for uniform policies across Member Nations http://ec.europa.eu/research/science-society /document_library/pdf_06/communication-022007_en.pdf
European Science Foundation	Europe	NA	Researchers are expected to follow the policies of the national agencies that directly provide research funding.

NA, not applicable.