J Comput Aided Mol Des (2014) 28:1–4
DOI 10.1007/s10822-013-9705-z

WARR'S PIECE

Data sharing matters

Wendy A. Warr

Received: 18 December 2013 / Accepted: 26 December 2013 / Published online: 17 January 2014

Introduction

It is 10 years since NIH issued their 2003 statement [1] requiring grant applicants to include details of their plan for sharing data. Numerous other mandates have followed [2–7], but the data that want to be freed still do not show many signs of escaping. Many would agree with the Royal Society [8] in the United Kingdom that:

- scientists need to be more open among themselves and with the public and media
- greater recognition needs to be given to the value of data gathering, analysis and communication
- common standards for sharing information are required to make it widely usable
- publishing data in a reusable form to support findings must be mandatory
- more experts in managing and supporting the use of digital data are required, and
- new software tools need to be developed to analyze the growing amount of data being gathered.

Unfortunately there has been less agreement about who should carry out and fund all these laudable objectives. There has been much publicity about "big data", but in the world of cheminformatics we cannot effectively share even "small data".

Why does this matter? Olga Kennard said many years ago [9] that the foundation of the Cambridge Crystallographic Data Center (CCDC) fulfilled a dream she shared with J.D. Bernal. They "had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments". Only by building on the results of others do we discover new knowledge, but sharing does not only encourage new discoveries in old datasets, it also fosters new collaborations, opens up observations to independent scrutiny, and may help detect fraud. Interestingly, it has also been shown that sharing detailed research data is associated with an increased citation rate for the scientist who produced the original dataset [10]. In this era of data-intensive science [11] access to data is of paramount importance. Data sharing matters and this article outlines some of the matters involved.

Stakeholders

The stakeholders are many: grant-making agencies; national scientific and technical information centers and major libraries who have a mission to preserve the digital record of science; data curators and data center organizers; publishers; scientists themselves, and others who use the data; commercial and academic organizations who employ and pay scientists; and the missionaries and visionaries who dream of a world where all data are open to all. There needs to be cooperation among all the stakeholders in the scholarly communication cycle: the Research Data Alliance [12] is one sign of progress toward that goal.

Definitions

Research data management is the organization of data, from its entry to the research cycle, through to the dissemination and archiving of results. Data curation is about the organization, management, and long-term preservation

W. A. Warr (✉)
Cheshire, UK
e-mail: wendy@warr.com

of research data. "Open science" means optimal sharing of research results and tools such as publications, research data, software, educational resources and infrastructures across institutional, disciplinary and national boundaries. "Open data" is the growing movement to disseminate datasets along with their published articles. Open notebook science takes this concept further by putting laboratory notebooks online in real time [13]. The open infrastructures needed are being addressed by a number of initiatives [14–18].

Barriers to data sharing

Many of the technological problems of data sharing are being addressed, but what is harder to change is the culture. There are differences between disciplines. Useful data lifetime is different for different disciplines and attitudes to raw and processed data differ. Anecdotal accounts estimate that 75 % of raw "small" data sit inaccessible in drawers and filing cabinets [19]. A cultural shift is needed if data are to be seen as equal in value to a publication.

Most researchers pay lip service to the ideal of data sharing, but in practice the advantages often fail to outweigh scientists' concerns. Many have a sense of "ownership", some of it bound up with the fear of relinquishing their rights, or having their data scooped, poached or misused. Finding the time and money to find, format and submit the data is another barrier to open communication and sharing. Other barriers are lack of organizational structures in academia, lack of professional training in data management, lack of priority among researchers, and lack of institutional mandates. Currently, researchers get little credit for data sharing; in an open science world, the academic and research career systems should support and reward scholars who participate in the culture of sharing.

Other barriers to sharing are overestimation of the commercial value of data and a belief that commercial value means closed data. It is worth noting, though, that openness is not unqualified; commercial activity may be a legitimate boundary, as may personal privacy, safety and national security. Unfortunately the complexities of intellectual property rights for data can cause problems. In many instances a dataset will not have been created entirely by the authors themselves but rather will have made use of pre-existing datasets, and the publishing of a dataset may require the approval of many associated data publishers.

Data on the web and in repositories

Nowadays there are many free data sharing resources on the Web: databases of chemical structures, collaborative environments, and portals that allow chemical and biological data to be mined and modeled. There are so many that I can cite only some representative examples. Some gather chemical structures and data [20–25] while others allow modeling of data related to structures [26–30].

It goes without saying that data must be of high quality, and that in itself raises fundamental issues. The growing number of data sources on the Internet raises questions about quality [31], and the laudable aim of linking all the sources can also affect quality by proliferating errors. Quality is likely to be higher in a data repository managed by a professional data center. RSC has started to address the issue of quality with its Chemistry Validation and Standardization Program to normalize the chemistry present in ChemSpider [20]; Open PHACTS [23] uses the same validation system.

Few of these resources are repositories run by a data center. Data repositories may be institutionally-based or discipline-based. A data center does not only store and share data: it manages them, improves them through appropriate data curation, protects them and encourages their creative re-use [32]. For the data to be searchable and usable, they must be formatted in a standard way, conform to standard structure and semantics, and have appropriate metadata attached. Metadata management tools streamline the process of annotating data with a description of what the items mean, which instrument collected them, which algorithms have been used to process them, and so on. Also necessary is software that can keep track of which pieces of data came from whom. This is essential if tenure and promotion committees are to give credit to scientists for their track-record of data deposition, but provenance is clearly of value for other reasons as well.

Databib [33], a tool for helping people to identify and locate online repositories of research data, lists no fewer than 602 repositories of one sort or another. Some discipline-specific successes well known to readers of this journal are arXiv.org, Protein Data Bank, GenBank, and the Cambridge Structural Database (CSD).

Supporting information

Some researchers archive their data on their own websites. In disciplines with community-supported data archives, researchers can deposit their data in a safe and reliable way, and publishers can also ensure persistent links between the data and related publications. Many authors add their underlying research data as supporting information (SI) to journal articles and publishers can do the linking [19]. DataCite [34] was founded in 2009 to help researchers find, access, and reuse data. It aims to establish easier access to scientific research data on the Internet,

increase acceptance of research data as legitimate, citable contributions to the scientific record, and support data archiving that will permit results to be verified and re-purposed for future study. Working with about 80 data centers worldwide, it assigns persistent identifiers (Digital Object Identifiers, or DOIs) to datasets. DataCite metadata are exposed using the Open Archives Initiative Protocol for Metadata Harvesting [35].

Unfortunately, the data in SI are often formatted in a non-standard way, are not readily searchable, and in the long term not guaranteed to persist. Some years ago, Anderson et al. [36] found that on average only 83 % of SI sets were still accessible a year after publication and about 10 % of all data supposedly available through an SI site was seemingly never available at all. Rzepa [37] insists that data should be presented in a form that allows humans to browse them, and machines to find and act on them. His solution is a two-component model, in which the article and the data are separately published. His group has published a recent article [38] in *Nature Chemistry* and the authors used figshare [39] for a citable DOI for the data [40]. The article is the "narrative" into which the data are woven and the data themselves are citable. The two components are individually citable and symbiotic. Rzepa encourages other authors to share their data, either by using a digital data repository at their institution, or by using a site such as figshare.

PLoS journals have made publication contingent on making the data "freely available without restriction, provided that appropriate attribution is given and that suitable mechanisms exist for sharing the data used in a manuscript".

Real "data journals" are now appearing. For example, Nature Publishing Group will launch *Scientific Data* in spring 2014. *Scientific Data* is a new open-access, online-only platform for the publication of descriptions of scientifically valuable datasets. It will publish a new type of content called Data Descriptors: peer-reviewed, scientific publications that provide detailed descriptions of experimental datasets. *Scientific Data* will allow for the formal peer review, publication and citation of datasets. It will give credit, through a citable publication, for depositing and sharing research data. The actual data files will be stored in one or more public, community-recognized systems. Where a community-recognized repository does not exist, *Scientific Data* supports the deposit of the data into a more general repository such as Dryad [41] or figshare.

Cost and sustainability

Significant sums of money are needed for sustainability of any funding model, and sustainability is essential. Many of the major public repositories have no stable underlying funding, and there are data types, particularly new ones, without appropriate public data repositories. The PDB, which researchers consider "free", has public funding in multiple countries and seems to be sustainable for several years, but will it be viable in 20 years' time, if governments and funding agencies change their strategies? CCDC is a not-for-profit company. In 2012, over 80 % of the income needed for the database was provided by industrial users and by the activities of a separate software company the CCDC operates [42]. When the Open PHACTS project ends in 2014, there are plans for a successor organization that will offer membership and paid-for services to encourage application builders. ChemSpider is now owned by the Royal Society of Chemistry, which runs it as a free service, financed by the sales of Web Services. Data may need to be "open" but they are certainly not "free"!

Conclusion

As a reader of this journal you might well be an avid user of data who does not give much thought to the problem of curating those data, or to ensuring that your own data will be usable by all and sundry in 10 years' time. In 2014 we plan to publish at least one special issue on data and some of the matters outlined in this article will be addressed in much more detail.

References

1. Final NIH statement on sharing research data. http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html. Accessed 16 Dec 2013
2. NSF Data Management Plan. http://www.nsf.gov/eng/general/dmp.jsp. Accessed 16 Dec 2013
3. NSF Data Sharing Policy. http://www.nsf.gov/pubs/policydocs/pappguide/nsf13001/aag_6.jsp. Accessed 16 Dec 2013
4. UK Medical Research Council data sharing policy. http://www.mrc.ac.uk/Ourresearch/ethicsresearchguidance/datasharing. Accessed 16 Dec 2013
5. Wellcome Trust policy on data management and sharing. http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm. Accessed 16 Dec 2013
6. UK Biotechnology and Biological Sciences Research Council data sharing policy. http://www.bbsrc.ac.uk/publications/policy/data_sharing_policy.html. Accessed 16 Dec 2013
7. EPSRC policy framework on research data: expectations. http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx. Accessed 16 Dec 2013
8. Science as an open enterprise. http://royalsociety.org/policy/projects/science-public-enterprise/report/. Accessed 16 Dec 2013
9. Kennard O (1996) Bernal's vision: from data to insight JD Bernal Lecture 1995. Birkbeck College, London
10. Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. PLoS One 2(3):e308

11. Hey T, Tansley S, Tolle K (2009) The fourth paradigm: data-intensive scientific discovery. Microsoft Research, Redmond, WA
12. Research Data Alliance. https://rd-alliance.org/. Accessed 16 Dec 2013
13. Interview with Kristin Briney. http://bulletin.acscinf.org/node/540A. Accessed 16 Dec 2013
14. EUDAT. http://www.eudat.eu/. Accessed 16 Dec 2013
15. Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission. Oct 2010. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf. Accessed 16 Dec 2013
16. Permanent Access to the Records of Science in Europe. http://www.parse-insight.eu/. Accessed 16 Dec 2013
17. U.S. National Science Foundation Office of Cyberinfrastructure Sustainable Digital Data Preservation and Access Network Partner (DataNet) http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141. Accessed 16 Dec 2013
18. Opportunities for Data Exchange. http://www.allianceperma nentaccess.org/index.php/community/current-projects/ode/. Accessed 16 Dec 2013
19. Integration of data and publications. http://www.stm-assoc.org/integration-of-data-and-publications/. Accessed 16 Dec 2013
20. ChemSpider. http://www.chemspider.com/. Accessed 16 Dec 2013
21. PubChem. http://pubchem.ncbi.nlm.nih.gov/. Accessed 16 Dec 2013
22. ChEMBL. https://www.ebi.ac.uk/chembl/. Accessed 16 Dec 2013
23. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B (2012) Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today 17(21–22):1188–1198
24. Collaborative Drug Discovery. https://www.collaborativedrug.com/. Accessed 16 Dec 2013
25. Toxnet. http://toxnet.nlm.nih.gov/. Accessed 16 Dec 2013
26. DSSTox. http://www.epa.gov/ncct/dsstox/. Accessed 16 Dec 2013
27. OCHEM. https://ochem.eu/home/show.do. Accessed 16 Dec 2013
28. ChemBench. https://chembench.mml.unc.edu/. Accessed 16 Dec 2013
29. OpenTox. http://www.opentox.org/. Accessed 16 Dec 2013
30. eTOX. http://www.etoxproject.eu/. Accessed 16 Dec 2013
31. Williams AJ, Ekins S (2011) A quality alert and call for improved curation of public chemistry databases. Drug Discov Today 16(17/18):747–750
32. Ashley K (2012) Data centers enable sharing and research recognition. Res Inf April/May 2012:26
33. Databib. http://databib.org. Accessed 16 Dec 2013
34. DataCite. http://datacite.org/. Accessed 16 Dec 2013
35. Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/pmh/. Accessed 16 Dec 2013
36. Anderson NR, Tarczy-Hornoch P, Bumgarner RE (2006) On the persistence of supplementary resources in biomedical publications. BMC Bioinform 7:260
37. Rzepa HS (2013) Emancipate your data. Chem World 10(11):39
38. Cowley MJ, Huch V, Rzepa HS, Scheschkewitz D (2013) Equilibrium between a cyclotrisilene and an isolable base adduct of a disilenyl silylene. Nat Chem 5(10):876–879
39. figshare. http://figshare.com/. Accessed 16 Dec 2013
40. Cowley MJ, Huch V, Rzepa HS, Scheschkewitz D (2013) Data for Nature Chemistry article "Equilibrium between a cyclotrisilene and an isolable base adduct of a disilenyl silylene". 10.6084/m9.figshare.744825. Accessed 16 Dec 2013
41. Dryad. http://datadryad.org. Accessed 16 Dec 2013
42. Sansom C (2012) Climbing the data mountain. Chem World 9(1):58–61