J Comput Aided Mol Des (2014) 28:973–974
DOI 10.1007/s10822-014-9796-1

Data sharing as an issue

Wendy A. Warr

Received: 9 September 2014 / Accepted: 12 September 2014 / Published online: 24 September 2014

In one of my recent discussion pieces [1] I suggested a special issue on data and here it is. I invited about 40 scientists to contribute and eventually nine teams produced the articles in this issue of the journal. Fortunately the nine contributions do cover most of the issues addressed in my "taster" article [1] and we hope that at least one of the authors who could not make my deadline might contribute to a future special issue on data.

Two of the articles might be considered as "position papers". Leah McEwen and Ye Li highlight opportunities for librarians as curators across the entire research life cycle, beyond ELNs and data sharing. Jeremy Frey and Colin Bird present a viewpoint on data sharing, addressing the reluctance of researchers to share information with their peers, and examining the processes of data exchange from the perspective of a trading environment. Another contribution on data sharing, by Sean Ekins and colleagues, on bigger data and collaborative tools, is noteworthy for its extensive reference list and its discussion on mobile apps, new technologies and the future of predictive drug discovery.

Other articles cover discipline-specific data repositories and databases. Helen Berman and co-authors give a history of the Protein Data Bank (PDB) and provide some insights into how this resource has evolved into one of the most widely used open access data resources in biology (and chemistry). Ian Bruno and Colin Groom of the Cambridge Crystallographic Data Center discuss the informatics tools that have been developed to allow expert human curation of the Cambridge Structural Database (CSD), and the opportunities afforded by recent technological developments and changing attitudes. Tony Williams and Valery Tkachenko describe the chemical structure centric hub ChemSpider, and discuss how changes in database technologies and the growing importance of the Semantic Web have motivated the Royal Society of Chemistry to re-architect ChemSpider and create a more generic data repository. Tony Slater describes a newer database of accurate and well-curated experimental pKa data: a computer-readable form of the IUPAC pKa data compilations. It adds value in the form of ionization assignments and tautomer enumeration.

Finally a couple of papers discuss "secondary services". Janna Neumann and Jan Brase report on the global consortium DataCite, which aims to establish easy access to data, to increase the acceptance of data publication, and to support data archiving. Digital Object Identifiers (DOIs) assigned to datasets facilitate access to data, and citation and reuse of data. Megan Force and Nigel Robinson give an overview of Data Citation Index which aims to link published research articles to their underlying datasets and track the citation of the data as well as to encourage bibliographic citation of the data. Data Citation Index also evaluates data repositories with respect to various selection criteria.

My initial article [1] discussed stakeholders, the culture of our scientific discipline, and barriers to data sharing. These issues are covered in particular by the articles in this issue by Leah McEwen and Jeremy Frey, and their co-authors. I introduced data on the Web and data repositories. Sean Ekins addresses data in "the cloud". The domain-specific repositories ChemSpider, PDB and CSD are well covered here.

Sadly, no major publisher volunteered to write about the Supporting Information to the research articles that they

W. A. Warr (✉)
Wendy Warr & Associates, Holmes Chapel, Crewe, Cheshire, UK
e-mail: wendy@warr.com

publish, but Thomson Reuters is represented as a secondary publisher in the article on Data Citation Index. Institutional repositories, as opposed to domain-specific ones, and generic data curation centers are also not well covered, and there is perhaps not as much in here about metadata and curation as I might have liked. Tony Williams does discuss data validation. Cost and sustainability are challenges that no one can ignore.

I am aware that the articles in this special issue are very different from the sort of articles that the Journal of Computer-Aided Molecular Design would usually carry. I hope that you, as a practicing researcher, will take the time to read them because sharing valuable data is so important to the future of science.

Reference

1. Warr WA (2014) Data sharing matters. J Comput-Aided Mol Des 28:1–4