CSIRO PUBLISHING

Australian Systematic Botany, 2016, **29**, i–ii
http://dx.doi.org/10.1071/SBv29n1_ED

Editorial

Data archiving – editorial

Endymion D. Cooper

Data archiving

The adoption of data archiving policies by a broad range of research journals will not be news to most readers of *Australian Systematic Botany*. Significantly, in 2010 a group of leading journals in ecology and evolutionary biology introduced the Joint Data Archiving Policy (JDAP) (see http://datadryad.org/pages/jdap) and in a series of editorials explained the benefits they expect will flow from it (Moore *et al*. 2010; Rieseberg et al. 2010; Rausher *et al*. 2010; Whitlock *et al*. 2010). The JDAP and similar policies have been, and continue to be, adopted by an increasing number of research journals. In keeping with best practice in scientific publishing, the editors of *Australian Systematic Botany* have elected to adopt the JDAP, and are updating the author guidelines to include the following policy statement:

> *Australian Systematic Botany* requires, as a condition for publication, that data supporting the results in the paper are archived in an appropriate public archive. Nucleotide and amino acid sequences must be deposited in GenBank (see http://www.ncbi.nlm.nih.gov/genbank/) or partnered database. Other types of data (e.g. sequence alignments and other phylogenetic matrices) should be submitted to a public archive, or provided as supplementary material for publication online. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at the time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to 1 year after publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as the location of endangered species.
>
> *Australian Systematic Botany* recommends that authors finalise data archiving prior to submission of their manuscript, and that accession numbers and confidential reviewer links are made available during the peer-review process. Authors should take care to provide accurate and informative annotations and metadata. Once the manuscript is accepted for publication and bibliographic details are available, these details should be added to archived data records in order to enable correct citation of the data's source.

As some might question the value of data archiving, it is worth reflecting on the benefits it provides. These can be summarised as: (1) preservation, (2) verification, (3) extension, (4) re-use and (5) recognition. If, as systematic botanists, we think of data in the way that we think of herbarium specimens the value of data archiving is obvious.

Preservation

The rapid loss of most scientific data has been written about at length elsewhere (e.g. Whitlock 2011). In large part data are lost because maintaining them takes time and resources that researchers would prefer to use for other things. And yet, the responsibility of maintaining data for future use, whether it be by the originator of the data or by others, often rests with individual researchers. Our science generally requires that voucher herbarium specimens are not privately maintained, but lodged in appropriate herbaria. This reduces the cost and risk of loss or damage by entrusting specimens to the care of reliable institutions; it also makes the specimens available to other researchers. It is logical to treat our other forms of data,

Endymion Cooper is a Marie Skłodowska-Curie Fellow at Queen Mary University of London where he researches genome evolution in plants. He studied for his Ph.D. at the University of Sydney and the Royal Botanic Gardens Sydney where his research focussed on the systematics of leafy liverwort family Lepidoziaceae. Between graduating from his Ph.D. in 2012, and taking up his present position, he spent 3 years as a Faculty Research Associate at the University of Maryland, College Park where his research used transcriptomics to study genome evolution in green algae. His primary research interest is in the evolutionary origin of land plants and his research integrates systematics, genomics and developmental biology. From his Honours year and Ph.D. he retains a particular fondness for solving the challenging taxonomic problems that tiny, cryptic plants present. Endymion joined the editorial board of Australian Systematic Botany in 2015. As an Associate Editor he has a particular interest in analytical rigour and reproducible research.

which are also costly and time consuming to generate, with similar care for their long-term preservation.

Verification

The principle of preserving the evidence upon which our conclusions are made is deeply embedded in systematic botany. Vouchers and type specimens provide a long-term record of the material used to make taxonomic conclusions, and therefore rules regarding their deposition in collections are mandated in our code of nomenclature. The specimens are also essential for deriving data for testing diverse hypotheses. In the same way that depositing voucher specimens and designating types enables conclusions based on those specimens to be verified, archiving data enables verification of conclusions based on those data. With concerns over a 'crisis of reproducibility' and the accuracy of published research (Ioannidis 2005), preserving the means of verification is essential for the advancement and credibility of science.

Extension

The availability of datasets allows researchers to add new data of their own or add to the analyses performed, and thereby generate new knowledge and deeper understanding (while appropriately crediting the authors of the datasets, see below). This is standard practice with herbarium specimens. We identify taxonomic novelties by examining new collections in the context of herbarium material, and we even recognise taxonomic novelties through closer inspection of existing specimens. Similarly, if phylogenetic data matrices are available, the incorporation of new accessions into existing phylogenies can accelerate description and classification of taxonomic novelties, and even lead to better resolution of the existing phylogeny. It is not hard to think of other examples where the availability of archived data makes scientific progress possible, where the cost of starting from scratch would be prohibitive.

Re-use

Archived datasets can be useful for answering questions not thought of by the originator of the data, and questions that had not yet arisen when the data were generated. Again, the specimens in our herbaria provide an excellent example. Recently the value of herbarium specimens as temporal and spatial records has become clear. Now, these records are increasingly being used to understand phenomena like air pollution (e.g. Shotbolt *et al*. 2007) and climate change (e.g. Primack *et al*. 2004), that were not issues until long after many herbaria were established. Sadly, questions that we might hope to answer based on the published literature, might be out of reach due to a lack of data archiving. For example, a comprehensive tree of life would be an incredibly powerful tool for testing hypotheses in ecology and evolution, and yet attempts to build large scale phylogenies from published work have been stymied by a lack of useable phylogenetic data (e.g. Hinchliff *et al*. 2015).

Recognition

There are various reasons why authors may be reluctant to publicly archive their data. One common view is that data archiving is an imposition on authors that provides benefit to others but returns little benefit to the authors themselves. However, this is not the case. When authors make data available with their publications it increases the impact of their work. Because data archiving enables verification, extension, and re-use, it increases interest in the research and invites citation and collaboration. Not only are papers accompanied by publicly archived data cited more frequently (Piwowar *et al*. 2007), but the datasets themselves are often highly cited.

References

Hinchliff CE, Smith SA, Allman JF, Burleigh JG, Chaudhary R, Coghill LM, Crandall KA, Deng J, Drew BT, Gazis R, Gude K, Hibbett DS, Katz LA, Laughinghouse HD IV, McTavish EJ, Midford PE, Owen CL, Reed RH, Reesk JA, Soltis DE, Williamsm T, Cranstonk KA (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 12764–12769. doi:10.1073/pnas.1423041112

Ioannidis JPA (2005) Why most published research findings are false. *Chance* **18**, 40–47. doi:10.1080/09332480.2005.10722754

Moore AJ, McPeek MA, Raucher MD, Riesberg L, Whitlock MC (2010) The need for archiving data in evolutionary biology. *Journal of Evolutionary Biology* **23**, 659–660. doi:10.1111/j.1420-9101.2010.01937.x

Piwowar HA, Day RS, Fridsma DB (2007) Sharing detailed research data is associated with increased citation rate. *PLoS One* **2**, e308. doi:10.1371/journal.pone.0000308

Primack D, Imbres C, Primack RB, Miller-Rushing AJ, Del Tredici P (2004) Herbarium specimens demonstrate earlier flowering times in response to warming in Boston. *American Journal of Botany* **91**, 1260–1264. doi:10.3732/ajb.91.8.1260

Rausher MD, McPeek MA, Moore AJ, Rieseberg L, Whitlock MC (2010) Data archiving. *Evolution* **64**, 603–604. doi:10.1111/j.1558-5646.2009.00940.x

Rieseberg L, Vines T, Kane K (2010) Editorial and retrospective 2010. *Molecular Ecology* **19**, 1–22. doi:10.1111/j.1365-294X.2009.04450.x

Shotbolt L, Büker P, Ashmore MR (2007) Reconstructing temporal trends in heavy metal deposition: assessing the value of herbarium moss samples. *Environmental Pollution* **147**, 120–130. doi:10.1016/j.envpol.2006.08.031

Whitlock MC (2011) Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution* **26**, 61–65. doi:10.1016/j.tree.2010.11.006

Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ (2010) Data archiving. *American Naturalist* **175**, 145–146. doi:10.1086/650340