Neuroinform (2010) 8:151–153
DOI 10.1007/s12021-010-9084-8

EDITORIAL

Data Publishing and Scientific Journals: The Future of the Scientific Paper in a World of Shared Data

Erik De Schutter

Published online: 11 September 2010

The rapid growth of the internet and related technologies has already had a tremendous impact on scientific publishing. This journal has given attention to open access publishing (Ascoli 2005; Bug 2005; Merkel-Sobotta 2005; Velterop 2005), to reforming the review process (De Schutter 2007; Saper and Maunsell 2009) and to the problems with getting authors to share their data (Ascoli 2006; Kennedy 2006; Teeters et al. 2008; Van Horn and Ball 2008) and how to enhance the use of shared data (Gardner et al. 2008; Kennedy 2010).

But the impact of the internet and data warehousing on science will be much larger and there is a growing interest in how these technologies can be leveraged to improve the scientific process (Hey et al. 2009). Let's travel towards the future and imagine that not only the tools and infrastructure are available to share scientific data at any time after it is generated, but that it has also become standard practice for the community to do so. How this can be achieved is not the focus of this editorial, instead I want to speculate on the relationship between scientific papers and data repositories (Bourne 2005, 2010; Cinkosky et al. 1991) in such an environment. It is important for the scientific community to discuss these issues now because, while these technologies are expected to radically improve the scientific process, they will also change the way in which our work is evaluated.

E. De Schutter
Computational Neuroscience Unit,
Okinawa Institute of Science and Technology,
Onna, Japan

E. De Schutter (✉)
Theoretical Neurobiology, University of Antwerp,
Antwerp, Belgium
e-mail: erik@tnb.ua.ac.be

I propose that we should distinguish data publishing from paper publishing (Callaghan et al. 2009; Cinkosky et al. 1991) and, when established for specific scientific fields, promote data publishing as the primary outlet for much of the scientific output.

A good metaphor for data publishing is to look at how complete organism genomic sequences are published in high impact journals now (Srivastava et al. 2010; Warren et al. 2010). Such papers serve really two goals: to announce the availability of the genome sequence in GenBank and to describe some scientific conclusions based on the analysis of the genome. The perceived importance of the latter determines whether a high impact journal will accept the paper and therefore the authors spend a lot of effort in hyping this part. But are these two components irrevocably intertwined? Couldn't one just publish the data, in this case by depositing the complete sequence in a database, and announce this fact through a form of publication? The analysis can then be published separately at a later time or distributed over different papers, etc. This is not done because at present the publication of the paper in the high impact journal is considered to be the optimal reward for the researchers, both for career advancement and for success in obtaining new grants (Bourne 2005). I call data publication a method where the data providers, who may be different from the people who analyze the data, receive credit for their work when they deposit the sequence in the database and where subsequent access to the data is tracked and considered equivalent to paper citation.

There are a number of advantages to considering data publication as a separate process. First, credit assignment becomes more explicitly defined among the authors. Several journals (like Nature, Science, the PLoS series, etc.) have taken steps towards a more granular credit assignment by asking authors to explicitly list their

contribution to the work, but this information gets usually lost in a curriculum vitae. Data publishing also makes it possible to give better credit to scientists who focus less on paper publication, like for example people working at the Allen Institute for Brain Science (Jones et al. 2009). Second, the data quality can be reviewed separately from the analysis. Quality of review is an important issue in modern science as it becomes increasingly difficult for journals to ensure that the reviewers selected are knowledgeable about all the technical aspects of complex papers using a diversity of methodological approaches (De Schutter 2008; De Schutter et al. 2009). By separating out the review of the quality of data acquisition and data curation one cannot only ensure that appropriate experts are selected, but also make sure that this part of the review is done thoroughly. Finally, and most important, data publishing by definition encourages the sharing of data and can promote its accessibility in ways that paper publication cannot.

There are many ideas floating around about what can be done to annotate and comment data repositories (Gupta et al. 2008; Kennedy 2010), but it is also interesting to think about new ways of publishing data, like for example publishing workflows (Bourne 2010). Here I want to argue that science would be much better served if at least half of the papers being published now would be replaced by data publishing. I refer to the large amount of papers that describe measurements and observations that increment our knowledge but do not provide fundamental new insights. Some examples are measuring rate constants of enzymes, characterizing subunit composition of ion channels in a particular type of neuron, counting cell densities in different brain regions, mapping anatomical connections, etc. These can all be useful data, but are usually not considered to be of general interest and end up being published in low impact journals. While these publications are valuable for the authors and for the grant agencies that supported the science, few people actually read the papers and it is not a very effective way to communicate the data. As a modeler I know how difficult it can be to find such data in papers (De Schutter and Van Geit 2010). Even if one finds relevant papers through a literature search, which is not guaranteed because it is difficult to formulate adequate searches, the paper may be in a low impact journal that is not accessible through the local library. Worse, to increase the perceived impact of these type of data they are often bundled together to obtain a larger paper, making it difficult to find specific data sets like, for example, unique normal control data in a study on pathological cases. When the proper databases and data publication methods exist, scientists will be able to publish their findings as an addition to the database record for, respectively, the enzyme, neuron type or brain region investigated. This submission could be reviewed before publication. Such mechanisms will also encourage scientists to make the data they normally do not publish, like negative results, available. If set up properly, such databases allow merging of high throughput data generated by specialized centers with that generated by the single lab, small-scale approach that is still so common in neuroscience. It is obvious that this will make it much easier to find the data than through PubMed. Data publishing is also a very cost effective model as it makes the authors do the work of entering the data (Cinkosky et al. 1991), presumably in the correct place. Without data publishing one can also create such databases but unless journals force authors to submit the data, which till now has not been successful in neuroscience, it is likely that many would not feel motivated to do so.

Having argued that a focus on data publishing would improve overall data accessibility, and therefore also enhance its effective use and increase credit to the authors, I want to end by considering how this will change the scientific journals. Even if data publishing becomes common, scientists will still want to communicate about their work. I believe, however, that in many cases this may be more effectively done through the well-established mechanism of scientific conferences together with abstract publication in open access journals. These procedures are sufficient to show the work that has been done, to discuss it with colleagues and to announce the availability of new data to the rest of the scientific world for the large amount of research that is of interest only to specialized communities. The same argument applies to analysis or modeling that does not provide insights that are of general interest, provided that these forms of work can also be published in databases.

When the scientific world embraces data publishing, scientific journals will become less important and should focus on papers that give added value to the data instead of just describing it. For sure supplementary information under the form of figures or text will disappear. I believe that four types of papers will dominate in this version of the future: highlights, methods, reviews and opinions and discussions. High impact journals will largely transform into journals that highlight exciting new findings, publishing short summaries emphasizing the implications of the findings and, in some cases, how these findings were obtained and why they are justified. These papers will be short and written for a broader scientific audience, with links to the data and analysis in databases which is where the specialists will go for deeper understanding. Additional layers, like the workflows (Bourne 2010) or technical documentation, could be provided between the highlight paper and the core databases to enhance complete presentation of the work. These journals will have to compete with social media based websites providing similar functionality.

Methods are at present a challenge in databasing because it is very difficult to fully characterize them using ontologies (Mackenzie-Graham et al. 2008; Zhao et al. 2009). Therefore, although eventually methods databases will arise, I expect that the typical methods paper will continue to be useful in the future. Review papers can be seen as literary form of databasing and the increasing number of review journals demonstrates the huge demand for good review papers. They are probably the most efficient way to inform oneself in the presence of data overload. This will not change in a world of data publishing, the difference will be in how review papers refer to findings: instead of citing papers most of the references will be to databases.

The final category of paper may be more a wish than reality, but I hope that the increasing use of social media combined with a decreasing need for scientists to spend time on writing and editing elaborate scientific papers, will encourage the return of the scientific discussion. In the present publishing environment, frank and opinionated discussion of controversial findings is discouraged. It is not wise to do it in your own papers, because it may alienate the reviewers and therefore inhibit the publication of your important data and findings. Journals also discourage long discussion sections because they are concerned about page limits, though one can wonder whether this is still relevant as most journals are moving steadily towards electronic publishing only. Very few journals have opinion sections and, even if they allow for opinion papers, personal experience shows that it is very difficult to get these through the review process (De Schutter 1995). But as it will become much more common to reanalyze data (Van Horn and Ishai 2007; Wan and Pavlidis 2007) in a world of data publishing than is the case now, I predict that there will also be more need to discuss controversies in an accessible way. New journals that publish short opinion papers combined with online discussion forums may have a nice future in such an environment.

References

Ascoli, G. A. (2005). Looking forward to open access. *Neuroinformatics, 3*, 1–4.

Ascoli, G. A. (2006). The ups and downs of neuroscience shares. *Neuroinformatics, 4*, 213–216.

Bourne, P. (2005). Will a biological database be different from a biological journal? *PLoS Computational Biology, 1*, 179–181.

Bourne, P. E. (2010). What do I want from the publisher of the future? *PLoS Computational Biology, 6*, e1000787.

Bug, W. (2005). The impact of the NIH public access policy on literature informatics: what role can the neuroinformaticists play? *Neuroinformatics, 3*, 81–91.

Callaghan, S., Hewer, F., Pepler, S., Hardaker, P., & Gadian, A. (2009). Overlay journals and data publishing in the meteorological sciences. *Ariadne*, *60*. http://ie-repository.jisc.ac.uk/349/.

Cinkosky, M. J., Fickett, J. W., Gilna, P., & Burks, C. (1991). Electronic data publishing and GenBank. *Science, 252*, 1273–1277.

De Schutter, E. (1995). Cerebellar long-term depression might normalize excitation of Purkinje cells: a hypothesis. *Trends in Neurosciences, 18*, 291–295.

De Schutter, E. (2007). Neuroscience leading the way: reviews cascade by the INCF. *Neuroinformatics, 5*, 205–206.

De Schutter, E. (2008). Reviewing multi-disciplinary papers: a challenge in neuroscience? *Neuroinformatics, 6*, 253–255.

De Schutter, E., & Van Geit, W. (2010). Modeling complex neurons. In E. De Schutter (Ed.), *Computational modeling methods for neuroscientists* (pp. 259–283). Cambridge: MIT Press.

De Schutter, E., Ascoli, G. A., & Kennedy, D. N. (2009). Review of papers describing neuroinformatics software. *Neuroinformatics, 7*, 211–212.

Gardner, D., Akil, H., Ascoli, G. A., Bowden, D. M., Bug, W., Donohue, D. E., et al. (2008). The neuroscience information framework: a data and knowledge environment for neuroscience. *Neuroinformatics, 6*, 149–160.

Gupta, A., Bug, W., Marenco, L., Qian, X., Condit, C., Rangarajan, A., et al. (2008). Federated access to heterogeneous information resources in the Neuroscience Information Framework (NIF). *Neuroinformatics, 6*, 205–217.

Hey, T., Tansley, S., & Tolle, K. (Eds.) (2009). The fourth paradigm: Data-intensive scientific discovery. Seattle: Microsoft Research. http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx.

Jones, A. R., Overly, C. C., & Sunkin, S. M. (2009). The Allen Brain Atlas: 5 years and beyond. *Nature Reviews. Neuroscience, 10*, 821–828.

Kennedy, D. (2006). Where's the beef? Missing data in the information age. *Neuroinformatics, 4*, 271–273.

Kennedy, D. N. (2010). Making connections in the connectome era. *Neuroinformatics, 8*, 61–62.

Mackenzie-Graham, A. J., Van Horn, J. D., Woods, R. P., Crawford, K. L., & Toga, A. W. (2008). Provenance in neuroimaging. *Neuroimage, 42*, 178–195.

Merkel-Sobotta, E. (2005). Elsevier and open access. *Neuroinformatics, 3*, 5–10.

Saper, C. B., & Maunsell, J. H. (2009). The neuroscience peer review consortium. *Neuroinformatics, 7*, 89–91.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M. E., Mitros, T., et al. (2010). The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature, 466*, 720–726.

Teeters, J. L., Harris, K. D., Millman, K. J., Olshausen, B. A., & Sommer, F. T. (2008). Data sharing for computational neuroscience. *Neuroinformatics, 6*, 47–55.

Van Horn, J. D., & Ball, C. A. (2008). Domain-specific data sharing in neuroscience: what do we have to learn from each other? *Neuroinformatics, 6*, 117–121.

Van Horn, J. D., & Ishai, A. (2007). Mapping the human brain: new insights from FMRI data sharing. *Neuroinformatics, 5*, 146–153.

Velterop, J. (2005). Necessity is the mother of innovation. *Neuroinformatics, 3*, 11–14.

Wan, X., & Pavlidis, P. (2007). Sharing and reusing gene expression profiling data in neuroscience. *Neuroinformatics, 5*, 161–175.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Kunstner, A., et al. (2010). The genome of a songbird. *Nature, 464*, 757–762.

Zhao, J., Miles, A., Klyne, G., & Shotton, D. (2009). Linked data and provenance in biological data webs. *Briefings in Bioinformatics, 10*, 139–152.