Neuroinform (2011) 9:317–320
DOI 10.1007/s12021-011-9131-0

EDITORIAL

Next Steps in Data Publishing

David N. Kennedy · Giorgio A. Ascoli · Erik De Schutter

Published online: 20 September 2011

Many different stylistic types of publication are necessary to support the various types of information that this journal is designed to serve. Specifically, *Neuroinformatics* currently supports publication of the following article types: Original Article, News Item, Editorial, Commentary, and Review. In general, editorials and commentaries are invited submissions, whereas original articles, news items and reviews are non-solicited. While the majority of published papers in this journal are 'original articles', it is felt that the 'news item' is under-utilized by this community as an important form of rapid communication.

Stylistically, the guidelines for a *news item* are that it be short (up to 4 printed pages), does not contain an abstract or section headings, is footnoted as opposed to referenced, is limited to two figures/tables, and must include an Information Sharing Statement (see below). Editorially, they are not necessarily reviewed by outside reviewers, and can be thought of as an 'advertisement' for a resource or concept. There is typically no specific 'hypothesis' tested and the implications or significance are not required to be stringently defended, as long as the topic is within the scope of the journal. Deviations from these guidelines will only be granted by the editorial staff in exceptional circumstances.

An *original article*, on the other hand, is stylistically full-length, fully referenced, sectioned and is required to include an abstract. Original Articles are usually peer reviewed by three outside reviewers, and must make a significant contribution to the scientific or technological advancement of the field. In most cases, there should be a hypothesis, method of approach, results with analysis of data acquired to support or refute the hypothesis, and discussion of the significance and implications of the findings in the context of the rest of the field. Exceptions to these presentation criteria are permissible for submissions presenting a methodological approach. While there is no specific limit on length or number of figures/tables, page costs mandate a premium on brevity and conciseness of presentation to the extent possible. All 'original articles' must include an Information Sharing Statement (see below). An optional Acknowledgements section is permitted immediately prior to the References.

Information Sharing Statement—Original Articles and News Items must (and all other articles should if relevant) include an Information Sharing Statement immediately prior to the Acknowledgments and References section. The purpose of this statement is to disclose the practical sharing details for all information sources utilized in each article. Examples include public data repositories and databases for structured data and models, websites for

D. N. Kennedy (✉)
Division of Neuroinformatics, Department of Psychiatry,
University of Massachusetts Medical School,
Worcester, MA, USA
e-mail: David.Kennedy@umassmed.edu

G. A. Ascoli
The Krasnow Institute for Advanced Study,
George Mason University,
Fairfax, VA, USA

E. De Schutter
Theoretical Neurobiology, University of Antwerp,
Antwerp, Belgium

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

software distribution and database schemas, etc.[1] The statement should make an explicit statement about if/how all resources utilized in this work can be accessed by the general public.

Enhanced Data Publication—A recent editorial by De Schutter in the pages of this journal discussed the topic of data publishing in scientific journals.[2] In the current publication scenario, a data release does not necessarily fit well in the traditional original article format. This limitation has been well known in the neuroinformatics field and indeed is one of the premises on which this journal was founded; to specifically support description, dissemination, and advancement of the resources essential to the infrastructure of neuroscience. When articles are received for review that principally describe a data release, we, as editors, and our selected reviewers tacitly suspend our 'traditional' manuscript concepts and review the submission based upon the perceived impact and significance of the data release and the completeness and accuracy of its description.

Thus, in order to realize the vision[2] that promotes the primacy of the 'data' in the scientific endeavor (in contrast to the current primacy of the interpretation of data), we want to encourage and strengthen the abilities of this journal to support the publication of high quality, richly reusable, fully described data through consideration of a new 'Data Original Article' publication format. We invite feedback from the community on this format as we move towards adoption of this new publication type.

One of the issues underlying a data-based publication, and the premise that the data described is publicly available, is where and how to host the data. Journals are not, of course, databases or likely to evolve into large-scale data storehouses. Thus such data publications should represent a collaborative connection between a data repository and a journal publication. While this journal has always promoted data sharing for data reported in its publications, the Data Original Article for a data release is a chance to turn the emphasis more in the sharing direction. In other words, the current sharing policy could be roughly translated as: "If you publish an article in this journal, we strongly recommend that you share the data somewhere". The rough policy statement behind the 'Data Original Article' concept could be stated as: "If you are sharing your data, we can reduce the barriers to publishing it". This is not only beneficial to the data collectors, but also beneficial to the data hosting operation in that it promotes both contribution to and use of its data resources.

There are myriad types of data that can, and should, be shared; and there are numerous existing data stores, tailored to specific subsets of these data.[3] Currently, data sharing tends to be retrospective and haphazard. Of the entire set of 'data' that is collected in support of neuroscience research, only a miniscule fraction is captured, in its raw from, in any sort of data store. Advancing the publication value of data collection itself, and promoting data sharing through data publication, we can exert a 'positive pressure' towards even more active data sharing and dissemination. Data store maintainers should embrace, adopt, promote and facilitate the active coupling of data deposition in a specific store with data publication, and their recommendations for publication will be actively sought. Indeed, the domain specificity of the data store provides provides a first level of curation and identification of datasets that are likely to be ready for the publication process as well as the minimum data requirements for publication.

The next substantial issue is related to how data descriptions should be reviewed. The criteria for the publication of public sharing and reuse of data should focus on the following domains:

- ***Data Description***

– Why were the data collected? Data are not collected in a vacuum. In all cases, there is some overarching reason for collection of a specific set of data, and hence some hypothesis that was envisioned that the data would shed light upon. A prior hypothesis is essential in order to assess the utility and completeness of a data collection for specific purpose, but the initial hypothesis does not necessarily limit the future questions that can be asked and answered with the data. As an example, data that are collected to assess the relationship between the structure of the planum temporale and language would be expected to have a structural and language component to the assessment[4]; lacking one or the other would render the data incomplete to answer the proposed question. This does not, however, mean that a data set supporting only the structural assessment of the planum temporale would not be of value to the community, only that the data support a different

[1] Kennedy, D. N. (2004). Barriers to the socialization of information. *Neuroinformatics*, *2*(4), 367–368.

[2] De Schutter, E. (2010). Data publishing and scientific journals: the future of the scientific paper in a world of shared data. *Neuroinformatics*, *8*(3), 151–153.

[3] Examples include: Halavi, M., et al. (2008). NeuroMorpho.Org implementation of digital neuroscience: dense coverage and integration with the NIF. *Neuroinformatics*, *6*(3), 241–252; Laird, A. R., Lancaster, J. L., & Fox, P. T. (2005). BrainMap: the social evolution of a human brain mapping database. *Neuroinformatics*, *3*(1), 65–78; Marcus, D. S., et al. (2007). The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data. *Neuroinformatics*, *5*(1), 11–34; Biswal, B. B., Mennes, M., Zuo, X. N., et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4734–4739.; etc.

[4] Zheng, Z. Z. (2009). The functional specialization of the planum temporale. *Journal of Neurophysiology*, *102*(6), 3079–3081.

question, and the presentation of the data to the community should be couched in a different fashion.

- How were the ‘raw’ data collected? In principle, a data release may contain both ‘raw’ and ‘derived’ sets of measures. An argument can be made, however, that the scope of a data release should be as fine-grained as possible since the potential independent uses of the data may require different starting points. This implies that a set of brain MRI data acquired, for example, to assess brain volume over the course of development may warrant an independent release of the ‘raw’ image data, as well as a release of the derived brain volume segmentation data, since the later may be subject to numerous interpretations and techniques related to its derivation. If bundled together in one release, the citation trail becomes muddled with discussion about what part of the release (the raw data or derived data) was the source material for a subsequent study. In the case of numerous derived data releases of similar scope that originate from a single ‘raw’ data release, the relative future impact of the different derivations is easily assessed through the development of the citation pattern of the resultant data articles. Collection details must include acquisition and sample/subject details. Sufficient details must be provided that enable the reader to believe that they could, given the time, resources and inclination, collect a comparable data set. An assessment of ‘completeness’ is predicated on the intent of the hypothesis, as introduced above, and the acquisition must be described in a fashion that supports such an assessment.
- How are derived measures created? When a release includes derived measures, a detailed methodological description of how the ‘raw’ data is transformed to generate the ‘derived’ data is required. This is a great opportunity to expand upon detailed analysis procedures, including provenance, in a fashion that is almost completely ignored in conventional articles.[5] Exact software versioning, parameter sets, pipelines and workflow descriptions, execution environment, etc. are examples of important factors that ultimately impact the reproducibility, and hence validity, of the analysis. The question reviewers should ask themselves is, “Given the data source and the processing description, can I get to the same derived result as presented here?”
- How to access the data. Part of the data description must include details regarding how to access the data. As the data itself is hosted remotely from the data article, a description of the access process, any permissions, license requirement, costs, other barriers, etc. should be fully disclosed. There may be valid reasons that access to the data is not completely free and unrestricted, and that should not preclude its publication; however, the barriers and restrictions must be disclosed and justified[1].
- Repository sustainability. What is the prognosis that the data will remain available in the future? Assessment should include the stability and robustness of the repository.

- ***Data Assessment***

- Quality—Disclosure of the data and its acquisition details are not, however, sufficient. The intent of data publication and sharing is, of course, data reuse. Data reuse is predicated on data quality. So, in order to facilitate reuse, it is important that data ‘quality’ metrics are provided for each data release. Note that this is not so much to act as the data quality police and set arbitrary thresholds for publication, but rather to make sure that future end users have the ability to elect to use specific data, or not, based upon relevant quality metrics. Of course, data with better quality metrics in general will be preferred, but it is impossible to state at time of publication what minimum quality may be of use to future processing and questions. Another confound is that there is typically no singular quality metric that is sufficient for all anticipated applications of the data. The authors, and the reviewers, must consider carefully what set of metrics are most appropriate for the specific domain and class of data. The most important ‘impact’ for a data article is the extent of future use of the data, and this will be related to the matching of the data quality with the needs of potential reuse opportunities.

- ***Data Significance***

- Discussion should review anticipated uses for the data. Data that has no conceivable use, of course, is of little value to document and release. Conversely, data that is amenable to many and diverse use cases will be highly valued, cited and rewarded.
- Discussion should review the relationship of the current release to other similar or related data. As most current scientific discovery is under-powered and under-replicated, there is a substantial value in providing a mechanism that supports the addition of comparable data to existing data reserves, as well as releases that establish new areas. The ability to augment existing data collections with new material keeps the releases vital, enhances the original investment in time and money expended for collection, and maximizes the value of new investment in data collection.

While one could imagine entire new journals devoted to data,[6] an alternative is to harness the capacity of the

[5] Mackenzie-Graham, A. J., et al. (2008). Provenance in neuroimaging. *Neuroimage*, *42*(1), 178–195.

[6] http://sites.google.com/site/beyondthepdf/, http://projects.iq.harvard.edu/datacitation_workshop/

existing journals to capture the data that are relevant to their readership and review expertise. Doing this in a standard fashion across the neuroscience journals would result in a greatly expanded set of available and fully described data.

In summary, all the available publication types should be actively utilized to support the culture and mandate to share data and resources. This mandate is particularly important as financial pressures magnify the scrutiny and required value of return on the research investment in the neurosciences. As the field evolves, the journal, its authors, reviewers, editors and readers are also required to evolve with it. Promoting a Data Original Article type will only be successful if it is adopted and embraced by the community, so now is the time to provide feedback on such an endeavor. In whatever form it eventually takes, it is clear that through facilitating data publication we can play a lead role in advancing science through transparent and reproducible methods.