D-Lib Magazine

January/February 2011
Volume 17, Number 1/2

"Earth System Science Data" (ESSD) — A Peer Reviewed Journal for Publication of Data

Hans Pfeiffenberger
Alfred Wegener Institut, Germany
Hans.Pfeiffenberger@awi.de

David Carlson
UNAVCO, USA
ipy.djc@gmail.com

doi:10.1045/january2011-pfeiffenberger

Abstract

In 2008, ESSD was established to provide a venue for publishing highly important research data, with two main aims: To provide reward for data "authors" through fully qualified citation of research data, classically aligned with the certification of quality of a peer reviewed journal. A major step towards this goal was the definition and rationale of article structure and review criteria for articles about datasets.

Introduction

Much has been said about the need to make research data available [1]. Meanwhile, this insight has found its way into policies set by research governing [2] and funding bodies at the highest levels, and even into high-level papers commissioned and authored [3] by the European Commission. This short paper cannot but briefly, in the first section, discuss how scientists can be brought to actually publish their data in a meaningful way. In short, the point is "reward", or actually: Recognition of data publishing as an academic achievement.

Lately, additional emphasis has been placed on quality assured data [4] (as well as reliable data repositories / data libraries). This topic is addressed in the second section on peer review of data as *one* means of making sure that other scientists can re-use published data reliably and "economically", i.e. without duplicating effort.

Rewarding Scientists who Publish Data

Hesitation still prevails among data creators, which feel themselves to be owners of the data: What do they have to gain, what would they loose by publishing their hard won data, *now*? Which is, of course, the dilemma faced by early scientists, who tended to accumulate their new findings until they were worth a book — until the 17th century.

As Mabe [5] richly illustrates, this dilemma was solved brilliantly in one sweep: In 1665 Oldenburg, the first editor of the Philosophical Transactions of the Royal Society, introduced all the essential elements of a modern scientific journal, including peer review and establishing priority. It is Mabe's theme, and in fact

most telling, that despite all those imaginable technical freedoms and added value possible in online publishing, the essential form and elements of recognized scientific publication has not changed at all.

Making data — technically — citable has been a theme for a number of years. To name a few, the CLADDIER [6] project and Green [7] explored how a citation of a dataset should be derived from the parties involved and from the (de-)composition of data elements actually used in an article citing the data. DataCite, the agency to assign DOI® names to datasets, derives from the project STD-DOI, which explored this agenda as early as 2004 [8].

Although technical and formal means to cite data thus exist, a wave of data publishing has not happened. This is frequently attributed to a restrictive "culture" or tradition in science: *What* to cite, not just *how* to cite. One must not look down on this conservative behaviour; it is a matter of survival for an author and even of credibility of science to cite *reliable* sources, only. Regarding reliability, the projects and methods mentioned provide stability and precision of reference to (specific parts of) datasets — but not certification of its content.

The journal Earth System Science Data (ESSD) aspires to provide datasets with this missing element (from the Phil. Trans. set). It will not replace data repositories, since data will not be stored or made accessible by the publisher (not to speak of preservation); rather, the editors require the dataset to reside in a reliable repository *and* to have a stable way of linking to it. That is, for all practical purposes: we wish to see DOIs for datasets (but reserve the right to lower this barrier, for the time being).

A future place of ESSD or other data publishing journals in the workflow and infrastructure for publishing a scientist's new results is depicted in Figure 1.

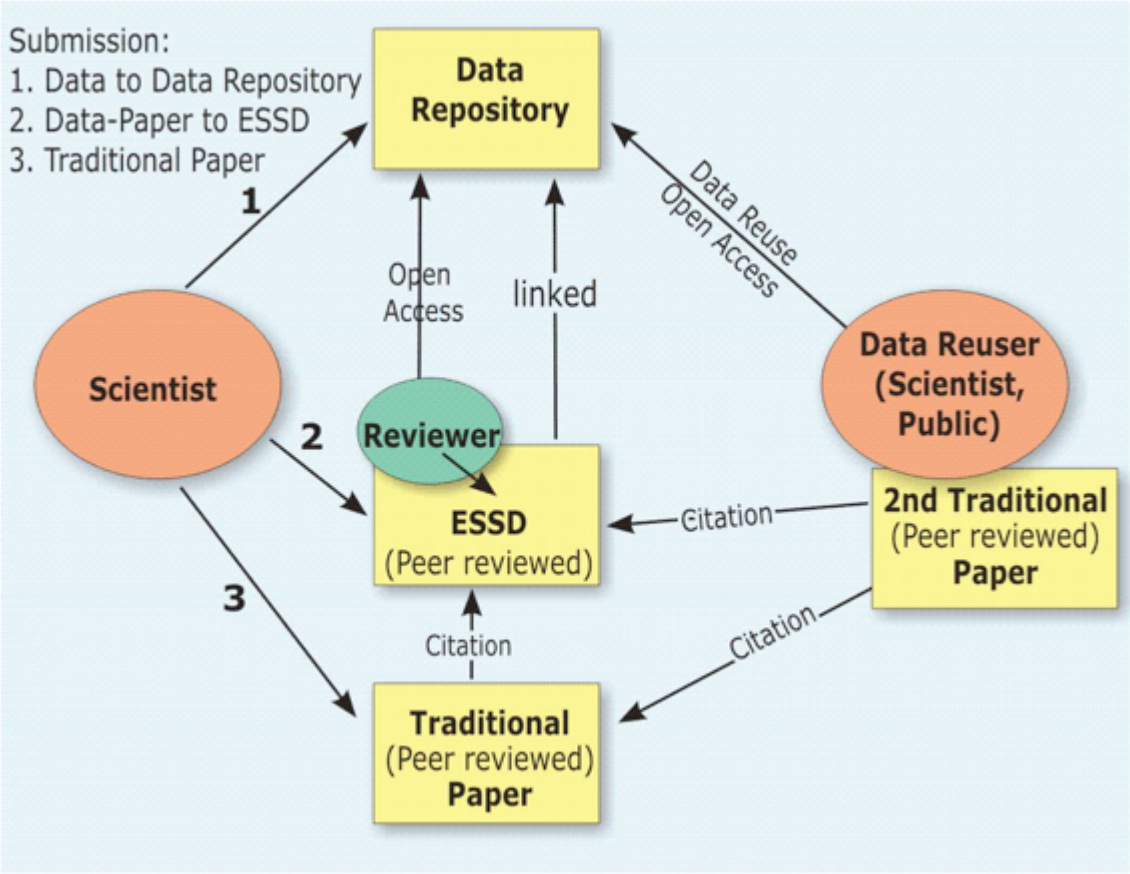

Figure 1: Ideal sequence and venues of submission and citation of scientific information, with ESSD. (For details, see text.)

It would of course be ideal if those data on which an ("traditional") article — with interpretation of the data — relies were published before submission of the article (so that anonymous peer review can take place).

ESSD enables the scientist to do this without giving up on receiving a reward when someone else publishes his or her own conclusions from this dataset. If the author, in her or his "traditional" article, cites the dataset as a reference to a journal article in ESSD, any other scientist ("Data Reuser") would practically and morally be forced to do likewise.

Publishing data in ESSD requires additional effort over submitting it to a reliable repository. However, the potential impact should in many cases be much higher than that from traditional papers, neutralizing the perceived need to keep the data under lock, for personal use only.

Peer Review of Data

In recent years and even months there has been discussion whether to ask reviewers of classical journals to "look at" the data. When, not so long ago [9], data were presumed to be stored on CDs in cardboard boxes, it was definitively not practical to ask reviewers to do this. Beyond this problem of easy (and anonymous) accessibility, it is obvious that a review of many datasets and data types will require skills different from those of the reviewer of interpretations and probably even more effort than consumed by the review of the article itself.

If a journal "publishes" an author's data supplement, what is a reviewer supposed to do with it? There are considerable differences answering this question, leading to editorial policies from requiring supplementary data as is now usual in genomics, to a thoughtful rejection of supplementary material as part of the material published under editorial/reviewers scrutiny. In his announcement of the later policy, Maunsell [10] finds this move necessary "because supplemental material has begun to undermine the peer review process in important ways." and that "Supplemental material also undermines the concept of a self-contained research report..." Authors of this journal "will be allowed to include a footnote with a URL that points to supplemental material on a site they support and maintain, together with a brief description..."

This radical position, obviously oriented at first principles of scientific communication, may actually make it easier to understand that we need to distinguish between two reasons — and associated modalities — to provide data: the first and most prevalent reason is to underpin an article with additional evidence. This would more typically be a dataset as small as possible, an excerpt or derived values only. It is this kind of data to which the "radical" position refers.

If, however, a potentially large number of articles are expected to rely on a (comprehensive or exceptional) dataset — also known as re-use of data — there is no way around the need to make sure, as far as possible, that this dataset itself is reliable. There may be communities of practise, e.g., in remote sensing or monitoring of environmental data, which work by established practises of documenting, testing and calibration of instruments, complemented by methods of (semi-)automatic validation of results. As long as those instruments and methods are operated by experienced, professional staff it may suffice for quality assurance to affirm just that, and by making all necessary documentation available. One might think of *a priori* quality assurance, here.

However, especially in pure research, there are many innovative and evolving and therefore less thoroughly documented and tested methods, which nevertheless produce substantial results, i.e., valuable data. It is this subset, which needs to be subject to quality assurance *a posteriori*. How can this, to put it loosely, "quality assessment with somewhat incomplete and/or ingenious documentation/proof" be done? One "obvious" answer is: Peer review, a method already practised and reasonably well understood by the parties involved.

Review Criteria and Article Structure

The editors of ESSD therefore believe that authors as well as all topical editors and reviewers will be quite confident in interpreting and applying the review criteria available from the journals' website [11]. Deliberately, the structure and wording has been adapted or even copied from traditional journals' criteria, so that just a mapping is needed of what these words could mean in the context of data. Some of these

mappings are straightforward ("uniqueness" becomes "originality") - others, possibly not.

As a speciality in data publishing, it would be much too easy to mass-produce articles, based on data series. We tried to capture a criterion to counter this under the slightly awkward label "Completeness". This, together with other criteria should result in the better-known "significance".

Many of these well-known criteria are simply an instrument to remove redundancy or other "noise" from the body of articles published, thus reducing the effort of reading and digesting the information received. Possibly even more important, because this is *the* bottleneck in current scientific communication, it also cuts the workload of reviewers.

Of course, the paragraphs on "data quality" are the most important part of the criteria. First, there is a seemingly trivial requirement for the data published to be "easily accessible in a useful format" — which, in real life, is neither trivial nor typical. This requirement is not only important for later re-use, but, again, even more for the work of the reviewers: They must be able to concentrate on the core of scientific data quality.

Within the article "claimed accuracy, the instrumentation employed and methods of processing should reflect the 'state of the art' or 'best practises' ". The reviewers will employ their best tools and competence to make sure that the claims are, at least, plausible and consistent. To facilitate this, there is a manuscript template, which requires from the authors sections on instruments and provenance, which should present and support claims of accuracy and circumstances of data acquisition or processing which might be significant in order to judge quality and reliability of the data published. (For a nice example, see the very first article in ESSD [12].)

Conclusion

Today, quality-related information is frequently not present in data or metadata, rendering both practically useless. ESSD requires it and adds reliability through the scrutiny of peers. Both are needed so that future users of these data can build on them, sceptically as always, but confidently.

The criteria and methods of ESSD will not be directly applicable to all types of data — for practical as well as intellectual reasons. However, it is the vision of founders and editors that the added value it provides to datasets will help foster substantial data publishing, re-use and *mandatory* citation of data in a subset of this journal's domain, Earth System Sciences.

Acknowledgements

Hannes Grobe and Manfred Reinke from AWI contributed to the initial discussions about complementing DOI-referenced datasets in PANGAEA with peer reviewed articles. Arne Richter and Martin Rasmussen from Copernicus Publications encouraged and helped establish the journal. The review criteria themselves were drawn up together with Sünje Dallmeier-Tiessen at AWI.

References

[1] Data's shameful neglect, Nature 461, 145, 2009. doi:10.1038/461145a

[2]Alliance of German Science Organizations, Principles for the Handling of Research Data. http://www.allianzinitiative.de/en/core_activities/research_data/data_policy/

[3] High-Level Expert Group on Scientific Data (web page), http://cordis.europa.eu/fp7/ict/e-infrastructure/high-level-group_en.html, and references therein

[4] EUROHORCs-ESF Task Force, EUROHORCs and ESF Vision on a Globally Competitive ERA and their Road Map for Actions, 2009. http://www.eurohorcs.org/SiteCollectionDocuments/ESF_Road%20Map_long_0907.pdf

[5] Michael A. Mabe, The more things change, the more they stay the same... – Why digital journals differ so little from paper. Third Bloomsbury Conference on E-Publishing and E-Publications, 2009, http://www.ucl.ac.uk/infostudies/e-publishing/e-publishing2009/1b-mabe.ppt

[6] Brian Matthews, Katherine Bouton, Jessie Hey, Catherine Jones, Sue Latham, Bryan Lawrence, Alistair Miles, Sam Pepler, Katherine Portwin, Cross-linking and referencing data and publications in Claddier, Proc. UK e-Science 2007 All Hands Meeting, 10-13 Sep 2007. http://epubs.cclrc.ac.uk/work-details?w=37696

[7] Toby Green, We Need Publishing Standards for Datasets and Data Tables, OECD Publishing White Paper, OECD Publishing, 2009. doi:10.1787/603233448430

[8] Jan Brase, Using Digital Library Techniques – Registration of Scientific Primary Data, Lecture Notes in Computer Science 3232, 488-494, Springer, 2004. doi:10.1007/b100389

[9] Emma Marris, Should journals police scientific fraud?, Nature, 439, 520-521, 2006. doi:10.1038/439520a

[10] John Maunsell, Announcement Regarding Supplemental Material, Journal of Neuroscience, 30(32):10599-10600, 2010. http://www.jneurosci.org/cgi/content/full/30/32/10599

[11] ESSD review criteria, http://www.earth-system-science-data.net/review/ms_evaluation_criteria.html

[12] Gert König-Langlo, Hartwig Gernandt, Compilation of ozonesonde profiles from the Antarctic Georg-Forster-Station from 1985 to 1992, Earth Syst. Sci. Data, 1, 1-5, 2009, doi:10.5194/essd-1-1-2009

About the Authors

Hans Pfeiffenberger is head of IT infrastructure at the Alfred Wegener Institut for Polar and Marine Research (AWI) and speaker of the Helmholtz Association's Open Access working group, where he specializes in access to data. Dr. Pfeiffenberger represents Helmholtz' interest in access to data in various bodies, such as the Priority Initiative "Digital Information" by the Alliance of German Science Organizations and the Alliance for permanent Access (APA). In 2008, Dave Carlson and he established ESSD. He holds a PhD in physics.

David Carlson directed the International Programme Office for the International Polar Year. IPY, with more than 50,000 participants from 60 nations, covered a wide range of science topics at a critical time for polar regions. Dr. Carlson has devoted more than 15 years to guiding and managing large international science programmes, starting from the very large Tropical Ocean Global Atmosphere programme in 1992 and 1993. He holds a PhD in Oceanography and led successful research teams focused on upper ocean physics and chemistry, oceanic microbiology and carbon cycling, and marine chemical ecology. Dr. Carlson now serves as Science Communication Director for the non-profit geodesy consortium UNAVCO in Boulder Colorado.

PRINTER-FRIENDLY FORMAT

Return to Article