AGU PUBLICATIONS

Earth and Space Science

COMMENTARY
10.1002/2016EA000201

Special Section:
Geoscience Papers of the Future

Key Points:
- Arguments toward further reproducible science are increasingly being made
- Enhancing the quality of geoscience papers motivates open data and software
- Best practices for digital scholarship in geoscience are summarized

Correspondence to:
C. H. David,
cedric.david@jpl.nasa.gov

Citation:
David, C. H., Y. Gil, C. J. Duffy, S. D. Peckham, and S. K. Venayagamoorthy (2016), An introduction to the special issue on geoscience papers of the future, *Earth and Space Science*, *3*, doi:10.1002/2016EA000201.

Received 3 AUG 2016
Accepted 3 OCT 2016
Accepted article online 8 OCT 2016

An introduction to the special issue on geoscience papers of the future

Cédric H. David[1], Yolanda Gil[2], Christopher J. Duffy[3], Scott D. Peckham[4], and S. Karan Venayagamoorthy[5]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA, [2]Information Sciences Institute and Department of Computer Science, University of Southern California, Los Angeles, California, USA, [3]Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, Pennsylvania, USA, [4]Institute of Arctic and Alpine Research, University of Colorado Boulder, Boulder, Colorado, USA, [5]Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, Colorado, USA

Abstract Advocates of enhanced quality for published scientific results are increasingly voicing the need for further transparency of data and software for scientific reproducibility. However, such advanced digital scholarship can appear perplexing to geoscientists that are seduced by the concept of open science yet wonder about the exact mechanics and implications of the associated efforts. This special issue of *Earth and Space Science* entitled "Geoscience Papers of the Future" includes a review of existing best practices for digital scholarship and bundles a set of example articles that share their digital research products and reflect on the process of opening their scientific approach in a common quest for reproducible science.

1. Introduction

A wide body of literature has been published in high-impact scientific journals over the past few years on the general subject of promoting transparent and trustworthy science results. Much of the discussion has revolved around the importance of scientific reproducibility [*Ioannidis et al*., 2009; *Hutson*, 2010; *Mesirov*, 2010; *Jasny et al*., 2011; *Russell*, 2013; *Collins and Tabak*, 2014; *Buck*, 2015] and on the need for data sharing [*Kattge et al*., 2014; *Scientific Data*, 2014] and software sharing [*Barnes*, 2010; *Ince et al*., 2012; *Easterbrook*, 2014; *Nature*, 2014; *Nature Geoscience*, 2014; *Shen*, 2014; *Hey and Payne*, 2015]. One of the outcomes of these discussions—as well as a contributor to its liveliness—has been new government mandates on increased sharing of digital products resulting from publicly funded research [e.g., *Holdren*, 2013].

Reproducibility, i.e., the confirmation of results and conclusions from one study obtained independently with the same or different methods and/or data, can be seen as "the scientific gold standard" [*Jasny et al*., 2011]. Transparency—a concept that is central to reproducibility—calls for making both data [e.g., *Kattge et al*., 2014] and software [e.g., *Ince et al*., 2012] openly accessible, which implies their publication in shared repositories with appropriate documentation and metadata. Reproducibility can be quite challenging when the methods involve physical samples and laboratory experiments and also in a computational realm where research products are digital. Although this practice increases the burden on research work and research funding [*Buck*, 2015], the ability to cite these digital research products has the potential to allow for acknowledgment and credit, hence motivating the added research workload [*Kattge et al*., 2014; *Kratz and Strasser*, 2015]. Reproducibility also requires a clear description of how the data and the software were used together to produce results (provenance), including any data transformations, parameters used, and software configurations [*Garijo et al*., 2013]. Such information is typically embedded in the text of research papers with the inherent ambiguity of natural languages [*Ince et al*., 2012] and with the restricted length of the "Methods" section. Yet the explicit computational representation of research methods is increasingly becoming practical thanks to tools such as interactive notebooks [*Shen*, 2014]. Finally, open access to digital research products implies careful choice of licenses to clarify permissions and restrictions intended by the authors [*Stodden*, 2009]. The importance of transparency and openness of digital research products is therefore increasingly being noted, such that guidelines now exist on how researchers and journals could promote the evolving research culture [*Nosek et al*., 2015]. However, while much of the published conversation has focused on the "what" (i.e., open research products and methods) and the "why" (i.e., reproducibility), comparatively less emphasis has been put on the "how" in terms of achieving transparency through tools and services that can be embedded in researchers' efforts.

2. Inception and Design of the Special Issue

On 11–13 March 2015, thirteen early-career geoscientists from U.S. universities, government agencies, and national laboratories and a group of computer scientists and geoscientists working on a project on digital stewardship all gathered at the University of Southern California's Information Sciences Institute in Marina Del Rey, CA, to discuss a vision for publishing geoscience papers in our digital era and the associated challenges. Together, the researchers envisaged a near future in which geoscientists would produce research papers that are augmented by all the associated digital objects (data, software, and their interconnections) following best practices of reproducible publications, open science, and digital scholarship. These best practices would make geoscience research products more openly accessible in data repositories with appropriate licenses, facilitate reproducibility through more formal statements of methods, and encourage fair credit for the associated scientific contributions through citations. Out of this meeting came the concept of the "Geoscience Paper of the Future," or GPF.

The GPF rests on a three-legged stool composed of openly accessible data, software, and transparent specification of the steps that linked them to generate results (provenance). This means that each GPF should include the following: (1) well-documented data sets that are available in public repositories and have unique citable identifiers; (2) documentation of software, including preprocessing of data and visualization steps, described with metadata and with unique citable identifiers and with pointers to public code repositories; and (3) documentation and availability of the computational provenance for each figure or result.

The definition of this vision for future papers, however, turned out to be the mere beginning of the GPF journey. Several of the meeting attendees decided to apply the aforementioned GPF guidelines to their own research papers which—as one might expect—turned out to be more challenging in practice than they had anticipated. It gradually became clear that there was not only benefit in demonstrating how the vision could be implemented but also that there was value in documenting and reflecting on the associated difficulties. Together, these papers serve as exemplars of how to implement the GPF best practices. In addition, the challenges exposed have the potential to motivate further work to improve transparency and reproducibility in geoscience.

The efforts that the American Geophysical Union (AGU) has made on transitioning toward more transparent science motivated the group of researchers to ask *Earth and Space Science*, one of AGU's open-access journals, to host a special issue that would unify their papers into a meaningful collection. This special issue was subsequently designed to present the concept of a Geoscience Paper of the Future and to provide GPF exemplars for a range of geoscience disciplines.

3. Content of the Special Issue

This special issue is organized as follows. In a review paper, *Gil et al*. [2016] introduce the recommended GPF best practices based on recent work in the research literature as well as working groups of scientific and professional organizations. The exemplar GPF papers then fall into two major categories of articles that the journal accepts: (1) *research articles* presenting new science results documented as a GPF and (2) *technical notes* documenting previously published papers as a GPF. In a research article, for example, *Essawy et al*. [2016] present a reproducible data-intensive impact study of drought on two U.S. States that is run remotely on a server and that uses a workflow framework that automatically records data products and tracks provenance. The research article of *Yu et al*. [2016] describes their experience with easing the learning curve of a complex surface-subsurface hydrological model through having new users reproduce an example test case and subsequently develop a new application. In a research article, *David et al*. [2016] reflect on their 10 year experience with open development (software and data) of a numerical river model. In a technical report, *Pope* [2016] extends one of his previous papers on supraglacial lakes into a GPF by fully exposing all digital products. The technical report of *Fulweiler et al*. [2016] offers a reproducible computational description of how they prepare data about nitrogen and oxygen fluxes from a bay in the Northeastern U.S., so that their research methods on benthic processes can be applied elsewhere. Other disciplines that also challenge the existing technologies and best practices for reproducibility, open science, and digital scholarship, are equally suited for a GPF. Each of the aforementioned studies follows the GPF guidelines and discusses the benefits and challenges associated with the publication of a GPF. All papers consistently emphasize the added burden associated with openly sharing data, software, and methods [*David et al*., 2016; *Essawy et al*.,

2016; *Fulweiler et al*., 2016; *Pope*, 2016; *Yu et al*., 2016]. Documenting and sharing data is a particular challenge when files are large and diverse [*Essawy et al*., 2016] and/or unearthed from old studies [*Fulweiler et al*., 2016; *Pope*, 2016]. Challenges associated with documenting and sharing software and methods are especially acute for data-intensive analyses [*Essawy et al*., 2016], further stressed when performed a posteriori of paper publication [*Pope*, 2016], and sometimes rise from the self-perceived inferiority of geoscientists' computer skills [*David et al*., 2016]. Additionally, openly sharing digital products highlights that the issue of licensing is complex and has to be addressed early and openly with coauthors [*David et al*., 2016; *Pope*, 2016]. Yet all papers unanimously highlight the benefits of transparency for reproducibility. The open availability of digital products and the automation of the methods enable the verification and/or update of researchers' own results [*David et al*., 2016; *Fulweiler et al*., 2016] and also allow for reproducibility by others during peer review or for research purposes [*David et al*., 2016; *Yu et al*., 2016]. Additionally, open data sharing motivates detailed data curation and guarantees safer data storage [*David et al*., 2016; *Fulweiler et al*., 2016; *Yu et al*., 2016]. Note that digital scholarship requires computer science skills that are best learnt as part of a teaching curriculum [*David et al*., 2016; *Fulweiler et al*., 2016] and are best applied in concert with research instead of years later [*Pope*, 2016]. Nevertheless, digital scholarship can enable better tracking of the impact of research results through citations [*Essawy et al*., 2016; *Yu et al*., 2016]. The added burden of transparent research can also be alleviated by community contributions [*David et al*., 2016; *Yu et al*., 2016] and rewarded through inclusion of open scientists in papers or funding proposals, or during their evaluation and promotion [*David et al*., 2016]. Overall, reproducibility not only strengthens the scientific method but also allows for eased communication, clarity, and transparency with peers [*David et al*., 2016; *Fulweiler et al*., 2016] as well as the greater public [*Fulweiler et al*., 2016; *Yu et al*., 2016].

4. Conclusion

This special issue on Geoscience Papers of the Future therefore provides examples that others can follow to achieve transparency of their research and highlights the associated challenges and benefits. It is our hope that readers will find valuable information and inspiration in this special issue and that an increasing number of geoscience publications will follow this approach to bring a desirable vision of future publications into a reality of the present, hence taking a step further toward the gold standard of scientific reproducibility.

Acknowledgments

The meeting that led to the concept of the Geoscience Paper of the Future was supported by the U.S. National Science Foundation under the EarthCube projects ICER-1343800 and ICER-1440323 that are designed to promote the development and adoption of transformative infrastructure for multidisciplinary geosciences research. Cédric H. David is supported by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the U.S. National Aeronautics and Space Administration (NASA) and by a grant from the NASA SERVIR Applied Sciences Team. Yolanda Gil, Christopher J. Duffy, and Scott D. Peckham are supported by the U.S. National Science Foundation under the EarthCube projects ICER-1343800 and ICER-1440323. S. Karan Venayagamoorthy is supported by the U.S. National Science Foundation under a CAREER grant OCE-1151838.

References

Barnes, N. (2010), Publish your computer code: It is good enough, *Nature*, *467*, 753, doi:10.1038/467753a.

Buck, S. (2015), Solving reproducibility, *Science*, *348*(6242), 1403.

Collins, F. S., and L. A. Tabak (2014), NIH plans to enhance reproducibility, *Nature*, *505*, 612–613, doi:10.1038/505612a.

David, C. H., J. S. Famiglietti, Z.-L. Yang, F. Habets, and D. R. Maidment (2016), A decade of RAPID—Reflections on the development of an open source geoscience code, *Earth Space Sci.*, *3*, 226–244, doi:10.1002/2015EA000142.

Easterbrook, S. M. (2014), Open code for open science?, *Nat. Geosci.*, *7*(11), 779–781.

Essawy, B. T., J. L. Goodall, H. Xu, A. Rajasekar, J. D. Myers, T. A. Kugler, M. M. Billah, M. C. Whitton, and R. W. Moore (2016), Server-side workflow execution using data grid technology for reproducible analyses of data-intensive hydrologic systems, *Earth Space Sci.*, *3*(4), 163–175, doi:10.1002/2015EA000139.

Fulweiler, R. W., H. E. Emery, and T. J. Maguire (2016), A workflow for reproducing mean benthic gas fluxes, *Earth Space Sci.*, *3*, 318–325, doi:10.1002/2015EA000158.

Garijo, D., S. Kinnings, L. Xie, L. Xie, Y. Zhang, P. E. Bourne, and Y. Gil (2013), Quantifying reproducibility in computational biology: The case of the tuberculosis drugome, *PLoS One*, *8*(11), 1–11.

Gil, Y., et al. (2016), Towards the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance, *Earth Space Sci.*, *2*, doi:10.1002/2015EA000136.

Hey, T., and M. C. Payne (2015), Open science decoded, *Nat. Phys.*, *11*(5), 367–369.

Holdren, J. P. (2013), Memorandum for the Heads of Executive Departments and Agencies. Increasing Access to the Results of Federally Funded Scientific Research, Executive Office of the President, Office of Science and Technology Policy, Washington, D. C.

Hutson, S. (2010), Data handling errors spur debate over clinical trial, *Nat. Med.*, *16*(6), 618–618, doi:10.1038/nm0610-618a.

Ince, D. C., L. Hatton, and J. Graham-Cumming (2012), The case for open computer programs, *Nature*, *482*(7386), 485–488, doi:10.1038/nature10836.

Ioannidis, J. P. A., et al. (2009), Repeatability of published microarray gene expression analyses, *Nat. Genet.*, *41*(2), 149–155, doi:10.1038/ng.295.

Jasny, B. R., G. Chin, L. Chong, and S. Vignieri (2011), Again, and Again, and Again …, *Science*, *334*(6060), 1225, doi:10.1126/science.334.6060.1225.

Kattge, J., S. Diaz, and C. Wirth (2014), Of carrots and sticks, *Nat. Geosci.*, *7*(11), 778–779.

Kratz, J. E., and C. Strasser (2015), Researcher perspectives on publication and peer review of data, *PLoS One*, *10*(2), e0117619, doi:10.1371/journal.pone.0117619.

Mesirov, J. P. (2010), Accessible reproducible research, *Science*, *327*(5964), 415–416, doi:10.1126/science.1179653.

Nature (2014), Code share, *Nature*, *514*, 536.

Nature Geoscience (2014), Towards transparency, *Nat. Geosci.*, *7*(11), 777–777.
Nosek, B. A., et al. (2015), Promoting an open research culture, *Science*, *348*(6242), 1422–1425, doi:10.1126/science.aab2374.
Pope, A. (2016), Reproducibly estimating and evaluating supraglacial lake depth with Landsat 8 and other multispectral sensors, *Earth Space Sci.*, *3*(4), 176–188, doi:10.1002/2015EA000125.
Russell, J. F. (2013), If a job is worth doing, it is worth doing twice, *Nature*, *496*, 7, doi:10.1038/496007a.
Scientific Data (2014), More bang for your byte, *Sci. Data*, *1*, 140010, doi:10.1038/sdata.2014.10.
Shen, H. (2014), Interactive notebooks: Sharing the code, *Nature*, *515*, 151–152.
Stodden, V. (2009), The legal framework for reproducible scientific research licensing and copyright, *Comput. Sci. Eng.*, *11*(1), 35–40.
Yu, X., C. J. Duffy, A. N. Rousseau, G. Bhatt, Á. Pardo Álvarez, and D. Charron (2016), Open science in practice: Learning integrated modeling of coupled surface-subsurface flow processes from scratch, *Earth Space Sci.*, *3*(5), 190–206, doi:10.1002/2015EA000155.