AGU PUBLICATIONS

Earth and Space Science

REVIEW

10.1002/2015EA000136

Special Section:
Geoscience Papers of the Future

Key Points:

- Describes best practices for documenting research to support open science
- Publishing computational provenance with software and data improves science transparency
- Promotes approaches to achieve equitable credit for all digital research products

Correspondence to:
Y. Gil,
gil@isi.edu

Citation:
Gil, Y., et al. (2016), Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance, *Earth and Space Science*, *3*, 388–415, doi:10.1002/2015EA000136.

Received 19 AUG 2015
Accepted 10 JUL 2016
Accepted article online 26 JUL 2016
Published online 14 OCT 2016

Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance

Yolanda Gil[1], Cédric H. David[2], Ibrahim Demir[3], Bakinam T. Essawy[4], Robinson W. Fulweiler[5], Jonathan L. Goodall[4], Leif Karlstrom[6], Huikyo Lee[2], Heath J. Mills[7], Ji-Hyun Oh[2,8], Suzanne A. Pierce[9], Allen Pope[10,11], Mimi W. Tzeng[12], Sandra R. Villamizar[13], and Xuan Yu[14]

[1]Information Sciences Institute and Department of Computer Science, University of Southern California, Los Angeles, California, USA, [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, USA, [3]IIHR Hydroscience and Engineering Institute, University of Iowa, Iowa City, Iowa, USA, [4]Department of Civil and Environmental Engineering, University of Virginia, Charlottesville, Virginia, USA, [5]Department of Earth and Environment, Department of Biology, Boston University, Boston, Massachusetts, USA, [6]Department of Earth Sciences, University of Oregon, Eugene, Oregon, USA, [7]Division of Natural Sciences, University of Houston–Clear Lake, Houston, Texas, USA, [8]Computer Science Department, University of Southern California, Los Angeles, California, USA, [9]Texas Advanced Computing Center and Jackson School of Geosciences, University of Texas at Austin, Austin, Texas, USA, [10]National Snow and Ice Data Center, University of Colorado Boulder, Boulder, Colorado, USA, [11]Polar Science Center, Applied Physics Laboratory, University of Washington, Seattle, Washington, USA, [12]Data Management Center, Dauphin Island Sea Lab, Dauphin Island, Alabama, USA, [13]Universidad Pontificia Bolivariana, Colombia, [14]Department of Geological Sciences, University of Delaware, Newark, Delaware, USA

Abstract Geoscientists now live in a world rich with digital data and methods, and their computational research cannot be fully captured in traditional publications. The Geoscience Paper of the Future (GPF) presents an approach to fully document, share, and cite all their research products including data, software, and computational provenance. This article proposes best practices for GPF authors to make data, software, and methods openly accessible, citable, and well documented. The publication of digital objects empowers scientists to manage their research products as valuable scientific assets in an open and transparent way that enables broader access by other scientists, students, decision makers, and the public. Improving documentation and dissemination of research will accelerate the pace of scientific discovery by improving the ability of others to build upon published work.

1. Introduction

Increasingly, scientists are asked to share their data, software, and other results of their research. These requests, which used to come only from fellow scientists, are now coming from a variety of sources including funders, publishers, policy makers, and journalists. The ultimate goal of this emerging movement is not only to make research products openly accessible to interested parties but also to enhance reproducibility, collaboration, and the directions and capability of future research. However, in order to be effective, making research products accessible requires careful planning as well as novel approaches that enable credit for these new forms of scientific contributions. It also requires awareness and social change in the scientific community, including clear communication of the benefits and best practices that may be new to geoscientists.

This paper presents a core set of best practices, reports on practical challenges in their implementation, and suggests practical workarounds when they cannot be followed. This work resulted from the efforts of the authors of this article in creating a Geoscience Paper of the Future (GPF) in their respective geoscience disciplines, in collaboration with computer scientists that guided them through the state of the art in digital scholarship. The authors span diverse geoscience disciplines, each with different kinds of data, community standards, and stages of adoption of cyberinfrastructure.

The implementation of best practices for digital publication of scientific research products (such as data sets, software, methods, etc.) requires effort. While there is a learning curve to understanding best practices for publishing research products, the curve is not as steep as it may initially seem. New infrastructure supporting the effective and successful management of these digital objects is being developed to make these tasks increasingly reasonable and manageable, and to ensure future availability and sustainability.

The goal of this paper is to instigate a nucleus of early adopters who will be the first to reap the benefits from open science, digital publication, and new forms of credit for their digital contributions to science. We believe that the best practices described in this paper will streamline data analysis and reporting in ways that will propel the geosciences forward in new and unanticipated directions.

2. Background and Motivation

The impact that digital publications can have on traditional science scholarship has been discussed in many forums. This section introduces key ideas and recent findings that serve as a motivation for our work. Although the views presented here may not be new to digital scholarship researchers, they have had limited dissemination and early adoption in geosciences. Our goal is to disseminate these ideas and more importantly to articulate best practices and make them easy to embed into the daily routines of geoscientists.

2.1. A Vision for Future Geoscience Papers

The publication of research papers is slowly changing to adapt to the digital age. We envision that in the near future (5–10 years), scientists will use radically new tools to author papers and disseminate information about the process and products of their research. These tools will document and publish the computational workflow as well as all the associated digital objects (data, software, etc.) that form the basis of a paper. This evolution in research publication will substantially improve scientific communication, promote a fair basis for crediting science contributions, and offer a transparent way for other scientists to evaluate and even reproduce the research. Today, several research tools exist to perform these tasks, but they are not routinely used and have not been integrated into the typical publication processes in geosciences.

It is our view that in the future publishers will accept submissions that include not only text and figures but also the data (both final and intermediate results), software, and other digital objects resulting from the study. These objects will be interlinked and contain metadata that allow readers to understand how the data and software were used to generate the study's results. Today, many journals accept supplemental data sets with an article, and some journals accept software or other digital objects, but geoscience journals do not require the complete details necessary to understand the connections between the data, software, and results of the research as computational provenance.

We envision that reproducibility (i.e., being able to recreate a study's results) and computational provenance (i.e., the digital documentation of what data and methods were used to obtain a new result) will be key review criteria for future geoscience publications. Readers of future geoscience papers will be able to actively interact with a published article, for example, by reorganizing the data or altering the computations that produced a figure. It should be straightforward to reproduce the results of a study because the connections between data, software, and resulting figures and findings will be more clearly expressed and documented in metadata. It will also be easier to build from past work by taking published methods/models associated with a paper and modifying them or running them with new data. Today, readers simply get a static paper, and in the rare cases where data are downloadable, reproduction of the analysis requires significant additional work or may not even be possible.

Another aspect of this vision of future geoscience papers is that these publications will include citations to the data and software resources used to complete the study. Although such resources are an important contribution to science, data producers and software developers often do not get credit for their work in the same way that authors of scientific papers get credit through citations. In the future, data and software should be citable resources with unique identifiers that allow all publications that build on their work to properly acknowledge them. This would reward those who create the data and software that form the basis of much of geoscience research and would encourage the production of high-quality products that can be reused by others to amplify the research potential of shared data and software.

2.2. A Changing Environment for Scientific Research

There are several major forces that push scientists to make their research open and accessible. We discuss here changes in publishing, the public's interest in science, funding, and scientists themselves.

2.2.1. Scientific Publishing Is Changing

Many journals accommodate the publication of data sets and in some cases other associated materials including code and other research products. Studies show that journals requiring a repository submission number as a condition of publication increase the likelihood of sharing data [*Piwowar and Chapman*, 2009]. Unfortunately, even in journals with clear data-sharing policies when the authors are allowed to announce that they will make data sets available upon request, studies have found that only one out of 10 authors comply [*Savage and Vickers*, 2009], which argues for requiring that data is published at paper publication time.

Publishers often have specific data and software requirements. The American Geophysical Union (AGU) does include research code in its definition of "data" that must be shared for publication in its journals [*American Geophysical Union*, 2013; *Hanson*, 2014]. More and more journals recognize the importance of documenting software in publications [e.g., *Nature*, 2014a]. Author guidelines for the Geoscientific Model Development journal require publication of code with documentation and a license, and reviewers are required to run the code with test cases supplied by the authors and comment on the ease of access [*Geoscientific Model Development*, 2013]. In addition to reproducibility and transparency, a major driver is the need for traceability across new versions of a model.

Although *data papers* and *software papers* are beginning to emerge as citable articles, most data and software in geosciences go unpublished and uncited [e.g., *Reichman et al*., 2011]. In data papers, authors write a scientific paper on the production and analysis of a data set that then becomes the recommended citation for the data themselves, although this still remains problematic for many data sets used in a primary scientific publication [e.g., *Pope et al*., 2014]. Journals devoted solely to describing software are beginning to emerge [*Journal of Open Research Software*, 2015; *SoftwareX*, 2015; *Source Code for Biology and Medicine*, 2015].

The record of origin or provenance of the results of published articles is often only provided at a high level in current articles, missing important details due to lack of space, or to the ambiguity inherent in natural language text. Interactive notebooks are gaining popularity, including iPhython Notebook for Python [*Shen*, 2014], Sweave for R in Latex [*Leisch*, 2002; *Falcon*, 2007], and the Computable Document Format for Mathematica [*Wolfram*, 2015]. Executable papers are designed to enable readers to run experiments [*Koop et al*., 2011]. Many scientific workflow systems now include the ability to publish computational provenance records [*Taylor et al*., 2006; *Koop et al*., 2011; *Mesirov*, 2010]. The Open Provenance Model was developed by the scientific workflow community and has been used extensively [*Moreau et al*., 2011], paving the way for the more recent W3C PROV standard for open publication of provenance [*Moreau et al*., 2013].

Publishers have been interested in improving digital scholarship practices. New approaches have been developed to document scientific articles so that they are more interactive than just a static probability density function. For example, ReadCube allows readers to navigate the citations through cross-reference facilities across publishers [*ReadCube*, 2015]. Other experimental efforts include the Executable Papers Challenge [e.g., *Van Gorp and Mazanek*, 2011; *Nowakowskia et al*., 2011; *Gavish and Donoho*, 2011], although this effort is focused on Computer Science, and the Article of the Future [*Zudilova-Seinstra*, 2013], which focuses on enhanced interaction between the reader and the publication (e.g., inclusion of published maps in Google maps, ability to zoom in on figures, and select data points).

Making all digital research products citable is also a major concern of publishers. Studies have found that more than half of the resources (reagents, organisms, etc.) mentioned in biomedical articles are not uniquely identifiable [*Vasilevsky et al*., 2013]. However, digital objects can be assigned persistent unique identifiers, such as Permanent URLs (PURLs) or Digital Object Identifiers (DOIs) [*DeRisi et al*., 2013], for unique identification. In addition, authors are also often assigned a unique identifier to distinguish among authors with identical names.

Scientific publications are increasingly linked to other digital information on the Web. Some publishers are linking digital assets to structured web data about people, locations, and all kinds of scientific objects [e.g., *Nature*, 2015].

2.2.2. Scientists Are Changing

Scientific organizations encourage open science [*Royal Society*, 2012; *Nature*, 2014b; *Science*, 2014]. Many research communities, editorials, and individual researchers have eloquently advocated for open science [e.g., *Costello et al*., 2013; *Nature Geoscience*, 2015; *Michener*, 2015; *Bourne*, 2010]. For every reason given for not sharing data or code, there are strong counterarguments [e.g., *Barnes*, 2010; *Costello et al*., 2013; *Nature Geoscience*, 2015; *Michener*, 2015].

Publishing and sharing data and software lead to better science [*Easterbrook*, 2014; *Joppa et al*., 2013]. Natural language descriptions of methods in papers have tremendous ambiguity that can lead to different interpretations and therefore different outcomes [*Ince et al*., 2012]. Focusing on geosciences as a case study, errors were reported in different implementations of the same algorithms [*Hatton and Roberts*, 1994; *Hatton et al*., 1988; *Hatton*, 1997]. This ambiguity is ingrained in natural language descriptions and consequently is unavoidable, so it is best to publish the software and computational provenance in addition to the data [*Nekutrenko and Taylor*, 2012].

A recent survey found that researchers want to be recognized more for their development of research resources for the community than for their invited presentations or the prestigious positions of their students [*Nature Metrics*, 2010]. Scientists are also recognizing the increased visibility and credit for their open science practices. A computational harmonic analysis research lab, WaveLab, reported on more than a decade of publications that included not only data and code for the paper but also examples, documentation, and credits [*Donoho et al*., 2009]. Their lab papers were on the top few cited mathematical sciences papers in the year they appeared, and such practices were described by the head of the lab as key reasons for becoming one of the top five most-cited authors in mathematics in the 1990s [*Donoho*, 2002; *Donoho and Huo*, 2004]. Important innovations in scientific credit and impact measures are beginning to emerge [*Priem et al*., 2010].

2.2.3. The Public's Interest in Science Is Changing

Science is a costly enterprise, and opening science creates new opportunities to leverage resources. Open sharing of research products enables the democratization of science and satisfies the public's interest in scientific data sharing [*Soranno et al*., 2014].

Opening science to the public enables scientists to harness massive amounts of volunteer effort from people who are able to make meaningful contributions [*Savage*, 2012]. Many citizen science projects have been wildly successful, including eBirds [*McCaffrey*, 2005], Zooniverse [*Lintott et al*., 2010], and FoldIt [*Khatib et al*., 2011]. In these projects, volunteers with no particular background in science create useful data for scientists. In some projects, citizen scientists have created their own science questions based on personal motivations and in some cases have made scientific contributions and are coauthors of publications in first-rate journals [*Cardamone et al*., 2009; *Fischer et al*., 2012]. The Polymath [*Nielsen*, 2011] project provides a massively collaborative online site wherein mathematicians collaborate with high school teachers, engineers, and other volunteers to solve mathematics conjectures and open problems by decomposing, reformulating, and contributing to all aspects of a problem. Several citizens collaborated to discover a gene mutation that was of interest to their families, learning to use science-grade data and tools and collecting additional data from volunteers [*Rocca et al*., 2012].

Open science practices also allow academic and industry to collaborate, creating beneficial and cost-effective synergies and broadening the societal impact of scientific research [*Woelfle et al*., 2011].

Finally, there is significant effort wasted when research results are not shared [*Macleod et al*., 2014], which is a practical and ethical concern for research supported by public funds.

2.2.4. Funding Agencies Are Changing

In response to a massive petition to make the results of federally funded research publicly accessible, the U.S. Office of Science and Technology issued a mandate for all government agencies that fund research to put a plan in place to release all research products so that they are publicly accessible [*Holdren*, 2013]. U.S. government funding agencies are responding to this mandate by developing plans to require research products to be openly published. The U.S. National Science Foundation (NSF) released a Public Access Plan in March 2015 [*National Science Foundation*, 2015] requiring that all research products be published for grants awarded after January 2016. The NSF already has a mandatory Data Management Plan in place, although it is not formally enforced. Plans are underway to determine how other research products are to be released. Other U.S. agencies that fund geosciences research, such as the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and the U.S. Geological Survey (USGS), have issued similar planning documents [*National Aeronautics and Space Administration*, 2015; *National Oceanic and Atmospheric Administration*, 2015; *U.S. Geological Survey* (*USGS*), 2015].

Some agencies are aggressively pursuing changes to project reviewing and evaluation processes. For example, the National Institutes of Health (NIH) are experimenting with a variety of approaches to make science more open [*Collins and Tabak*, 2014], including a pilot on having a special reviewer in each panel that checks the validity of the published articles that are the premise for a proposal. The NIH are also enhancing transparency through a new Data Discovery Index for unpublished primary data, online forums for discussion of published articles, and author checklists to facilitate verification by reviewers.

2.3. The Reproducibility Crisis

Scientific articles describe computational methods informally, often requiring a significant effort from others to understand and to reuse. Attempts to replication of published work naturally reveal uncertainties, which enable further scientific progress [*Jasny et al.*, 2011]. It is useful to distinguish between replication under identical conditions but different testers (*repeatability*), and replication with different testers and testing conditions (*reproducibility*), although the terminology used in different fields is not always consistent [*Kenett and Shmueli*, 2015]. Reproducibility can be challenging in some disciplines such as in ecology, but it can be attained and has significant benefits [*Ellison*, 2010; *Ryan*, 2011]. Reproducibility is a cornerstone of the scientific method, so it is important that reproducibility be possible not just in principle but in practice in terms of time and effort to the original team and to the reproducers. The reproducibility process can be so difficult and time consuming that it has been referred to as "forensic" research [*Baggerly and Coombes*, 2009]. Studies have also shown that reproducibility is in many cases not achievable from the article itself, even when data sets are published [*Bell et al.*, 2009; *Ioannidis*, 2005; *Ioannidis et al.*, 2009]. In a recent effort in cancer biology to reproduce 50 important papers, the slow response from authors to requests to release data made the effort difficult [*Van Noorden*, 2015], which argues for requiring the publication of data when the paper is published. Without access to the source codes for the papers, reproducibility has been shown elusive [*Hothorn and Leisch*, 2011; *Hey and Payne*, 2015]. In a recent study, only 11% of selected landmark papers in cancer research were found reproducible [*Begley and Ellis*, 2012]. An internal survey at Bayer pharmaceuticals found that about two thirds of their projects are canceled because of inconsistencies during attempts to reproduce published research [*Prinz et al.*, 2011]. In this era of big data, computational processes are becoming increasingly more complex and more challenging to reproduce [*Nature*, 2012a].

The justification of reproducible research has received increasing attention, particularly in climate science [*Santer et al.*, 2011]. The latest Coupled Model Intercomparison Project Phase 5 (CMIP5) provides vast amounts of model simulations useful for scrutinizing the past and future climate change [*Taylor*, 2012]. The computational expense and size of outputs for CMIP5 are much larger than its previous phase, CMIP3, due to the high-resolution and complicated processes included in CMIP5 models. As more models are publicly available for intercomparison projects, it is expected that major climate science journals require sharing the data analysis procedure in publications and making analysis results reproducible and applicable to similar data sets. Retractions of publications do occur more often than is desirable [*Roston*, 2015]. Indeed, *Fang and Casadevall* [2011] proposed tracking the "retraction index" of scientific journals to indicate the proportion of published articles that are later found to be problematic. In psychology, where several important studies have been called into question [*Yong*, 2012], labs have volunteered to do replication projects in collaboration with the original researchers [*Schooler*, 2014]. The Reproducibility Initiative offers to do validation studies to replicate papers of interest [*Baker*, 2012]. Ultimately, open sharing of data, code, and the computational provenance of the results will allow colleagues and reviewers to examine papers more closely and will increase validation of scientific research.

Computational reproducibility is a relatively modern concept. The Stanford Exploration Project led by Jon Claerbout published an electronic book containing a dissertation and other articles from their geosciences lab [*Claerbout and Karrenbach*, 1992; *Claerbout*, 2006]. The lab adopted "Reproducible Electronic Documents" (ReDocs), with sets of make rules that help build and run the application from scratch and take care of temporary files [*Schwab et al.*, 2000]. They described three degrees of reproducibility: easily reproducible (ER) if it can be easily rerun within 10 min, conditionally reproducible (CR) if it requires proprietary data, licensed software, or more than 10 min to run, and nonreproducible (NR) if it is material that is manually created (e.g., a figure). Advocates of reproducibility have grown over the years in many disciplines, from signal processing [*Vandewalle et al.*, 2009] to computational harmonic analysis [*Donoho et al.*, 2009] to psychology [*Spies et al.*, 2012]. Organized community efforts include reproducibility tracks at conferences [*Manolescu et al.*, 2008; *Bonnet et al.*, 2011; *Wilson et al.*, 2012], reproducibility editors in journals [*Diggle and Zeger*, 2009; *Peng*, 2009], and numerous community workshops and forums [e.g., *Bourne et al.*, 2011]. Repositories of shared computational workflows enable scientists to reuse workflows published by others and facilitate reproducibility, although these repositories do not yet have significant uptake in geosciences [*De Roure et al.*, 2009;

Missier et al., 2010; *Garijo et al*., 2014]. Other active research in this area is addressing a range of topics including copyright [*Stodden*, 2009], privacy [*Baker et al*., 2010], social [*Yong*, 2012], and validation issues [*Guo*, 2012].

The recommendations for making scientific research reproducible generally agree on requiring the publication and documentation of data, software, and methods [*Baggerly and Coombes*, 2011; *Claerbout*, 2006; *Donoho et al*., 2009; *Garijo et al*., 2013]. Advocates also propose broader changes such as adopting collaborative research practices, creating a replication culture, and training the scientific workforce [*Ioannidis*, 2014]. Reproducibility requirements would help principal investigators to be more accountable for the work in their labs [*Nature*, 2012b]. *Russell* [2013] proposes to tie grant funding to replication, since that work will be more likely to have increased returns. There is a need for better infrastructure beyond current tools and services [*LeVeque et al*., 2009; *Pebesma et al*., 2012].

[*Donoho*, 2010] mentions several important advantages of reproducibility, including improved work habits since others can examine the work, improved teamwork due to more efficient communication, greater impact since others can easily reuse the work, improved continuity since others can build on the work, and responsibility to taxpayers that the work is preserved.

Some publishers are agreeing to new guidelines for journals to develop author checklists that promote reproducibility [*Nature*, 2014a; *Science*, 2014; *Nature*, 2013], sometimes in coordination with funding agencies [*National Institutes of Health*, 2015].

2.4. Digital Scholarship in the Geosciences

Despite the notable efforts mentioned above, the geosciences are still behind in the practice of digital scholarship. Why the sluggish uptake?

Open science requires work that is often challenging for individual scientists to undertake. Credit for data, software, and other digital research products that benefit the scientific community must be recognized, particularly in academic promotion cases [*Harley*, 2013]. Policy issues and the role of journals and funding agencies are discussed in recent studies [*LeVeque et al*., 2009; *Stodden et al*., 2013]. Funders and publishers are significant driving forces to promote open science in the scientific community [*Kattge et al*., 2014]. Data facilities and academic institutions play a significant role as well. These are important issues that are being seriously considered by the community, and will bring about significant changes in scientific practice and publications in the coming years.

In geosciences, in particular, a significant challenge is the effort involved in evolving a traditionally descriptive and field centric discipline. There is always a cost in documenting any research product. The “why” and “how” of an artifact are ideally captured but this takes effort. Recognizing that there is a cost to documenting anything, researchers seem to perform these tasks despite the associated commitments and it may just be a matter of education and culture. No one would have imagined the open source movement and how it motivates programmers to release well-documented code to enable others to build on their work. Such endeavors are trending in data-intensive fields such as bioinformatics and computer science. In geosciences, subdisciplines such as climate modeling and geologic mapping have only recently begun transitioning to digital methods. Geoscience researchers also need the mechanisms, infrastructure, and benefits to transition into modern digital scholarship practices.

Another major challenge is that there is a diversity of sources for best practices, and none are very familiar or easily accessible to geoscientists. Although there are organizations that promote recommendations for data and software sharing, citation, and documentation, such as the Federation of Earth Science Information Partners (ESIP) and the Research Data Alliance (RDA) among others, they tend to reach people who focus on data management and informatics. Publishers announce new guidelines and requirements for their journals in response to those recommendations, but they tend to be very minimal in order to reduce the burden on authors. These guidelines are changing rapidly, in concert with changes on their business models given the open access trends in science mentioned above. In the end, many planned recommendations are still under development, particularly those concerning the description and citation of software, physical samples, and digital mapping and other visualizations.

Finally, another major factor is the lack of awareness of best practices and of opportunities to learn about them. Geosciences researchers by and large have minor familiarity with software sharing practices and little knowledge or enthusiasm about data sharing [*Reichman et al.*, 2011]. There exist only rare opportunities to learn about digital scholarship in practice. Therefore, the dissemination of best practices and new approaches to publishing research results in the digital age and of the benefits associated with open publication and sharing of data and other research products are both greatly needed.

This article aims to overcome these barriers by articulating and disseminating best practices and by suggesting how to implement them in ways that are realistic to accomplish in writing a scientific article today reaching to become a geoscience paper of the future.

3. The Geoscience Paper of the Future (GPF)

We propose a characterization of the Geoscience Paper of the Future (GPF) that aims to capture the core concepts behind open science, reproducibility, and modern digital scholarship. A GPF intends to satisfy the following requirements:

1. *Make data reusable* through publication in a public repository, with documentation (metadata), a clear license specifying conditions of use, and citable using a unique and persistent identifier.
2. *Make software reusable* through publication in a public repository, with documentation, a license for reuse, and citable with a unique and persistent identifier. This includes modeling software as well as all software for data (re)formatting, conversions, filtering, analysis, and visualization.
3. *Document the computational provenance of results* by explicitly describing the series of computations and their outcome in a high-level workflow diagram, a formal workflow, or a computational provenance record, possibly stored in a shared repository and citable with a unique and persistent identifier.

Figure 1 characterizes a GPF and highlights the differences with a reproducible paper. A reproducible paper focuses on the publication of data, software, and computational provenance of the results so that they can be rerun and reproduced. Those are all desirable characteristics of a GPF. In addition, a GPF focuses on the sharing of all research products and emphasizes their publication in public repositories with open licenses, unique and permanent identifiers that make them citable, and appropriate metadata to document their characteristics.

Given the current technical and cultural limitations to performing our envisioned leap in geoscience publications, we expect that it will take some time for papers in geosciences to satisfy all these criteria, and we acknowledge that papers that are not data- or software-focused (e.g., collection of physical samples or laboratory experiments) may not benefit from adopting them. Even when authors commit to publishing all data and software in an open shared repository, they often face difficulties such as the reluctance of coauthors to share specific data or software, the difficulty of fully describing experiments, the inability to share due to technological limitations (size, dependencies, existing repositories, infrastructure, etc.), and the necessity to simplify the approach for broad use (i.e., generating figures with easier formatting than generally used in published form). When faced with such challenges, GPF authors should reflect on the difficulties they face, pursue workarounds, and propose areas for future improvements.

We note that we use the terms “publication” and “publishing” to refer to the action of making objects (e.g., data sets, software) publicly accessible. This is a common way to use the term in a Web context, as in the Web just making a resource available through a URI is considered a publication. We note that in other contexts, “publication” and “publishing” involve archiving, curation, and quality assurance [*Parsons and Fox*, 2013].

We note that the citation and the publication of the computational provenance in a public repository are both optional. Ideally, both would be done by GPF authors, but we recognize the lack of shared computational provenance and workflow repositories in geosciences and therefore are recommended here although considered optional.

One of the goals of a GPF is to facilitate reproducibility. Reproducibility requires rerunning the experiments in the article, but inspectability simply requires examining the computational provenance records provided. While reproducibility takes significant effort, enabling inspectability can be relatively straightforward. Authors should make inspectability very easy for any reader of a paper and make reproducibility practical by making the effort required small enough that it is not out of the question for other researchers.

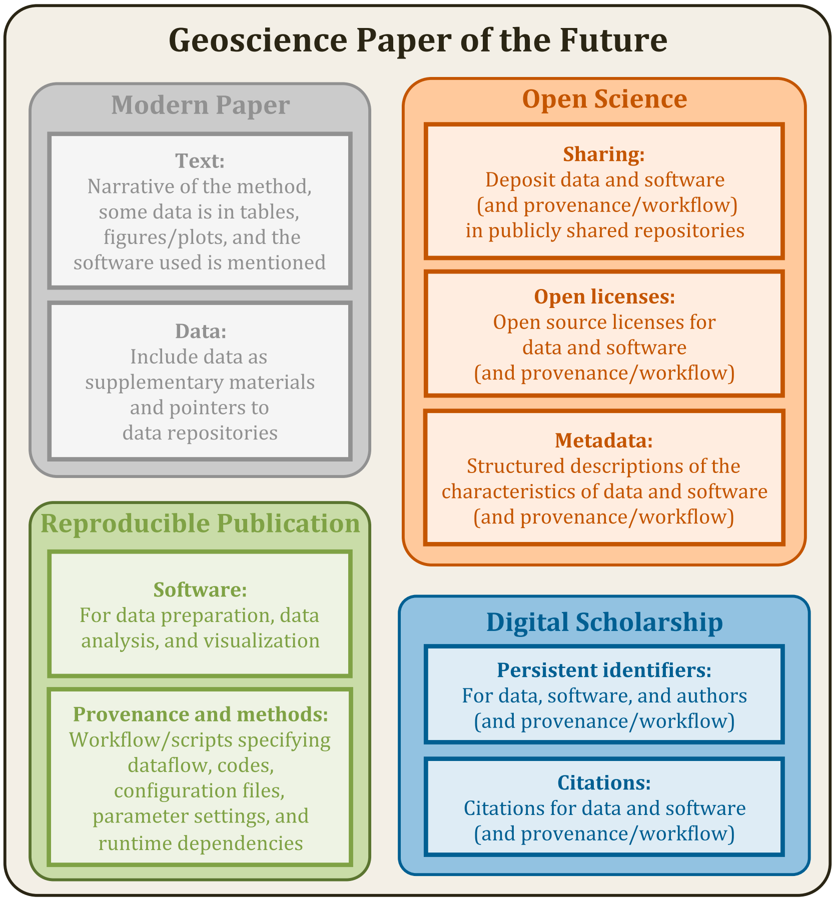

Geoscience Paper of the Future
Modern Paper
Text:
Narrative of the method, some data is in tables, figures/plots, and the software used is mentioned
Data:
Include data as supplementary materials and pointers to data repositories
Open Science
Sharing:
Deposit data and software (and provenance/workflow) in publicly shared repositories
Open licenses:
Open source licenses for data and software (and provenance/workflow)
Metadata:
Structured descriptions of the characteristics of data and software (and provenance/workflow)
Reproducible Publication
Software:
For data preparation, data analysis, and visualization
Provenance and methods:
Workflow/scripts specifying dataflow, codes, configuration files, parameter settings, and runtime dependencies
Digital Scholarship
Persistent identifiers:
For data, software, and authors (and provenance/workflow)
Citations:
Citations for data and software (and provenance/workflow)

Figure 1. A Geoscience Paper of the Future (GPF) includes data, software, and computational provenance as expected in reproducible publications but also includes desirable features in open science and digital scholarship: (1) sharing of data, software, and other research products in public repositories, (2) use of open licenses, (3) metadata that describes the characteristics of data, software, and other research products, (4) persistent unique identifiers for data, software, and other research products, and (5) citations for all data and other resources mentioned in the paper. GPF authors may find practical impediments to follow some of these recommendations, and in that case they should state their desire to do so and document the reasons for not following them.

4. Suggested Best Practices and Current Challenges

This section describes recommended best practices on how to document data, software, and computational provenance, and to uniquely identify and cite these digital objects. Table 1 provides a proposed author checklist consisting of 20 recommendations for creating a GPF and serves as a roadmap for this section. These best practices were compiled from recommendations by both scholars and organizations concerning digital publications [e.g., *Research Data Alliance*, 2015; *Committee on Data for Science and Technology*, 2013; *DataCite*, 2015; *FORCE11*, 2014; *Federation of Earth Science Information Partners*, 2012; *Open Archival Information System*, 2012; *Starr et al.*, 2015; *Uhlir*, 2012; *Downs et al.*, 2015; *Ball and Duke.*, 2012; *Mooney and Newton*, 2012; *Goodman et al.*, 2014; *Garijo et al.*, 2013; *Altman and King*, 2007]. They were developed as some of the authors of this article endeavored to write a GPF about their own work.

Our recommendations and best practices are independent of the particular area of research, computing platforms and languages, or approach to publishing. For those seeking more specific advice, *eScience* [2011] provides an excellent trove of pointers to resources for improving scholarly communications, including not just community repositories but also modern science communication such as blogging, screencasting, and collaborative idea generation.

Table 1. A Proposed Checklist for GPF Authors, With 20 Recommendations That Can Guide Them to Assemble the Information That Should Be Included in a GPF

Category	Applicability	Recommendations
Data accessibility	Initial data, significant intermediate results, and final results	D1: Data sets should be published in a publicly accessible location with a permanent unique identifier D2: Data sets should have a license D3: Data sets should be cited in the paper
Data documentation	Initial data, significant intermediate results, and final results	D4: Data sets should have general-purpose metadata specified D5: Data set characteristics should be explained in detail D6: Data set origins and availability of related data sets should be documented
Software accessibility	Software used to process initial data and to generate any intermediate or final results	S1: Software should be published in a publicly accessible location with a permanent unique identifier S2: Software should have a license S3: Software should be cited in the paper
Software documentation	Software used to process initial data and to generate any intermediate or final results	S4: Software function and purpose should be described S5: Software download and execution requirements should be documented S6: Software testing and reuse with new data should be documented S7: Software support for extensions and updates should be mentioned
Provenance documentation	Provenance of all computational results reported in the article, including figures, tables, and other findings	P1: Derivations of newly generated data from initial data should be provided P2: Software execution traces for newly generated results should be provided P3: Versions and configurations of the software should be specified P4: Parameter values used to run the software should be specified
Methods documentation	Computational methods that are generally applicable to data other than the data in the paper	M1: Compositions of software that form a general reusable method should be specified M2: Data flow across software components should be described
Authors Identification	Authors of the paper and of any new data and software cited in the paper	A1: Authors have a permanent unique identifier

4.1. Making Data Accessible

All the input data and results should be made accessible, as well as any key intermediate data that may help others understand or reproduce the work being described in the paper.

4.1.1. Data Accessibility: Location, Citation, and License

Location (*D1*). Data should be in a publicly accessible location. Many researchers include in their papers links to data sets published in their lab or personal web sites, which is easy and convenient. However, studies have shown that the majority of the articles that use such links have at least one broken link within 2 years [*Klein et al.*, 2014; *Dellavalle et al.*, 2003], and the availability of data declines quickly over time [*Vines et al.*, 2014]. An alternative and more desirable approach is to use a data repository. Many scientists view the sharing of data as onerous, but there are now many general repositories that make it very easy to publish data [*Tenopir et al.*, 2011; *Van Noorden*, 2013]. There are many data repositories available to scientists that ensure longevity and accessibility. Several meta registries contain pointers to data repositories, such as re3data [*Re3data*, 2015; *Pampel et al.*, 2013]. Data repositories can be institutional, discipline-specific, or generic to accommodate "orphan" data [*Vision*, 2010]. They differ in their community of use, search-ability/discoverability, ease of use, degree of curation (e.g., organization and preservation), and reputation. Repositories range from general domain and not curated [*Figshare*, 2015; *Zenodo*, 2015; *Dryad*, 2015], to more focused and curated (e.g., ACADIS [*ACADIS*, 2015], IEDA [*The Interdisciplinary Data Alliance*, 2015], NCEI [*The National Centers for Environmental Information*, 2015], Pangaea [*Pangaea*, 2015]), and finally to the highly specific, managed, and curated (e.g., AGDC [*The Antarctic Glaciological Data Center*, 2015], NASA's DAACs [*Distributed Active Archive Centers*, 2015], and the USGS Science Data Catalog [*USGS*, 2015]). Curation takes time, since data must be described well enough that another user, possibly even other communities with different knowledge bases and expectations, can use it. Cost is also an important issue to many researchers, especially those early in their careers. Many of these repositories are free, but have a limit in the size of the data they accept. Disciplinary or community-based repositories tend to highly curate their holdings, which increases the possibility that it is possible for others to reuse the data. The choice of a repository should also take into account other aspects of data management planning. Considerations include data formats (which may be proprietary or nondurable), data integrity (file naming/versioning, backups, "permanent" availability, etc.), data context (through documentation and metadata), discoverability of data, ease of access, ease of use, ease of citation, licensing, and, where appropriate, privacy concerns. All of these factors should be weighed when deciding on a data repository.

License (*D2*). The data should have a license that specifies any constraints for its reuse, including how the authors should be acknowledged, whether it can be modified before redistributing, or whether it can be used for commercial purposes. A widely used set of licenses is offered by *Creative Commons* [2015]. The most permissive licenses are CC-BY, which allows any modifications and uses provided attribution is stated, and CC-0, which waives all the rights of the creators to be reused by others.

Citation (*D3*). Data should be cited within text much like an article would be cited. Some journals have specific guidelines for data citation, in some cases requiring that data be cited in a special resources section or in the acknowledgments section. While there is no universal standard for data citation, agreement is emerging among various style guides, institutions, and publishers in that a data citation should include author names, the name of the data set, retrieval and/or publication date, publisher (or repository) name, version, access date, and access information in the form of a persistent unique identifier. Persistent unique identifiers to cite data include persistent URLs (PURLs) and Digital Object Identifiers (DOIs) [*DeRisi et al*., 2013]. A PURL is a URL that is permanent and will not change, but when accessed it redirects to another URL in a local system (e.g., a lab website) that can be changed over time. The creator of a PURL must update the link if the underlying URL changes. The PURL can be cited in the paper, and the authors must be responsible and ensure that the redirection address is updated if anything changes in their local system. A PURL can be obtained through services such as the W3C's Permanent Identifiers for the Web [*W3C*, 2016]. Although PURLs are better than URLs from the point of view of persistence, DOIs have the best archival guarantees."A DOI is a character string used to uniquely identify a digital object, such as an electronic document. DOIs are only issued by authorized sites, and most data repositories issue DOIs. A DOI consists of a publisher ID (prefix) and an item ID (suffix), separated by a forward slash (/). Any DOI can be easily turned into a URL format.

4.1.2. Additional Requirements and Issues

Taking data from public repositories. While many researchers collect or create their own data sets, many researchers take data from publicly available repositories. The NASA Global Change Master Directory is a recommended tool to discover data sets in geosciences [*Global Change Master Directory*, 2015]. Data repositories often indicate the license agreements to be followed and specify how the data extracted should be cited.

Using data from colleagues. Frequently, researchers will incorporate data sets from a combination of sources including data obtained formally or informally from colleagues. In this case, the author must make sure to have their permission to publish it, taking special care in clearly defining the authorship and the licensing conditions. The main challenge of using data from colleagues is having access to metadata.

Publishing intermediate data. Intermediate data should be published when data preparation steps are hard to reexecute or understand, or when there are manual processes involved. These steps take raw data (from local or external repositories) and produce data sets in the desired formats for further use within the analysis process. Data preparation includes quality control (removal of outliers, gap filling, etc.), unit conversions, corrections for time zone differences or daylight savings time, and extraction of subsets from a larger data set. These processes can change many of the data's characteristics including format, structure, quality, accuracy, and precision. It is important to document data preparation steps and to publish any key intermediate data generated.

Large data sets. In many disciplines, the availability of data sets with high spatial/temporal resolutions creates a challenge. Although data storage and transport costs are getting cheaper, sharing and transferring large data sets is still a challenge. Therefore, it is essential to prepare large data sets in efficient formats supported by repositories, software, and visualization tools. For example, NetCDF and HDF are widely used for meteorological data [*NetCDF*, 2015; *HDF*, 2015]. If a trusted discipline-specific repository is not available (or is too costly), general-purpose repositories can be used that accept unlimited data sizes [e.g., *Dash*, 2015] or that continue to increase the sizes of the data sets allowed [e.g., *Zenodo*, 2015; *Dryad*, 2015; *GitHub*, 2015a]. One possibility is to publish a sample of the data used and provide extensive metadata about the characteristics of the data. This way, even if the work cannot be replicated (i.e., another lab would not be able to run the method on the same data), this helps with reproducibility, since another lab would be able to see what kind of data were used and try to find similar data to reproduce the results (i.e., another lab using the same method and different data).

Timing data and paper publication. Some researchers and some publications impose moratoriums on data, which argues for coordination of the release of data and papers. While some journals require data to be archived and available through a trusted repository, some repositories will require data to be documented and published in a peer-reviewed journal (e.g., Dryad Digital Repository), often creating a chicken-and-egg situation. This situation must be streamlined for scientists in the future.

4.1.3. Data Accessibility: Recommendations

The best approach to ensure data accessibility is to find curated repositories that are widely used in the community and upload the data there. These repositories will provide a unique persistent identifier and will offer default licenses that will facilitate reuse. They will also offer a data citation that can be used to cite the data in the paper. They will also collect general metadata such as authors and license, as well as domain-specific metadata as we will discuss in the next section.

Ideally, every researcher should be familiar with the repositories used in their community and understand the requirements, capabilities, and long-term commitments of these repositories. By highlighting the value of these repositories, and by pointing out shortcomings that should be addressed, researchers will help the groups and institutions that manage these repositories to continue to serve the community. New repositories may be designed as a result of these kinds of community discussions.

In the cases that such a repository does not exist or may not be easy to find, a simple short-term solution is to use a general repository such as figshare, Zenodo, Dryad, Dataverse, or Pangaea. Although these repositories lack deep domain metadata that makes the data easier for other researchers to discover and reuse, at the very least they offer important capabilities such as archiving, persistent identifiers, citation, and version control. They also capture general metadata such as authors, basic keywords, and license information.

4.2. Documenting Data by Specifying Metadata

When making data available for public access, it is important to describe them in a structured form using metadata so that other researchers can understand what the data represent as well as enable them to find the data through queries and reuse them for their purposes. Metadata can take many forms, from unstructured text to standardized, structured, machine-readable, extensible content. Many repositories provide a specific format for metadata using a formal standard.

4.2.1. Documenting Data: General-Purpose Metadata and Data Set Characteristics

General-purpose metadata (*D4*). This is general information about the who/what/when/where/why/how of the data set. It should include the creator, date, funding agencies, purpose of the study, what was collected, timeframe and area covered, contact information, and other basic information about a data set. Most repositories request this kind of metadata.

Data set characteristics (*D5*). Scientifically relevant characteristics of the data set should also be documented. For example, the sensor used to collect data, descriptions of column headers in tabular data, units of measurement, and other characteristics that affect usability of the data. Different disciplines and areas of research care about different kinds of data characteristics, but there are many efforts to standardize how this kind of metadata is organized and collected. Most of the common formats for storing large data sets (e.g., NetCDF, HDF, XML) allow for inclusion of detailed descriptions concerning the specifics of each variable (average or instantaneous, time zone, long variable names, time step, spatial range, etc.). Several ISO metadata standards (e.g., ISO-19139 [*International Organization for Standardization* (*ISO*), 2007], ISO-19110 [*ISO*, 2005]) are popular in geosciences. Some discipline-specific standards, such as the Climate and Forecast (CF) metadata conventions for NetCDF files [*CF*, 2011], are increasingly gaining acceptance in their communities.

Data sources and related data sets (*D6*). Other researchers may not only be interested in reusing the data in a particular article but may also want to find similar data that suit their purposes. For example, a paper may use climate data for a particular region but other researchers may want to apply the same analysis for a different region of interest. For this reason, it is useful to document what data sources could be accessed in order to retrieve data similar to what is used in a paper. If the data set is extracted from a larger database or is one of several data sets collected for the same consortium project with multiple PIs covering different aspects of a large study, it is worth mentioning the existence of the other related data sets, the project that collected them, and the program that funded the work.

4.2.2. Additional Requirements and Issues

Metadata standards. In some disciplines there are coordination efforts to develop extensive metadata standards (e.g., *Moine et al*. [2014] for climate). Some specific examples of metadata standards, both general and domain specific, include the Dublin Core Metadata Terms (a domain-independent metadata standard for attribution) [*Dublin Core Metadata Initiative*, 2012], the Ecological Metadata Language (EML) [*Fegraus et al*., 2005], the Water Markup Language (WaterML) [*WaterML*, 2015], and FASTA for genetic sequence information [*Pearson and Lipman*, 1988]. More disciplines in geosciences are organizing community efforts to develop standards for metadata.

Data about physical samples. When digital information is generated from physical samples, the sample itself should be referenced and cited. The International Geo Sample Number (IGSN) was created for this purpose.

References to instruments and sensors. It is possible to refer to a specific instrument or sensor when describing how data were collected or analyzed. Sensors and instruments may have unique persistent identifiers.

4.2.3. Data Documentation: Recommendations

When using a well-curated data repository, extensive domain-specific metadata will be collected to facilitate discovery and reuse by other researchers. Providing this metadata takes time, but the goal is to make the data most useful to the community. There are always fields to document the data further, and any information to help others reuse the data should be specified.

When using general repositories, particularly noncurated ones, general metadata will be collected. It is also useful to specify keywords that will help others discover a data set, and to include documentation about the characteristics and origins of data sets. Curated repositories are another option, which involves additional investment to provide appropriate metadata. Repository curators ensure that enough metadata is provided that others will be able to find and understand the data. Typically, domain-specific standards are designed to facilitate this. For example, geospatial information is ubiquitous in geosciences and many curated repositories adopt geospatial standards to facilitate discovery and reuse.

There are several directories of data repositories [e.g., *Nature*, 2016; *Re3Data*, 2015], and some are emerging in geosciences [*Coalition for Publishing Data in the Earth and Space Sciences*, 2016].

Provenance is a major component of metadata no matter what data repository is used. It includes details about how the data were collected or generated, such as the instruments, processes, and entities involved. In this paper we focus on computational provenance, specifically on the computations performed to generate a new data set.

4.3. Making Software Accessible

When considering software availability, we often think about the big packages or models. But in addition, any other software written to transform data or generate a plot, or any ancillary data manipulation should be made publicly available. This kind of pervasive software sharing could greatly benefit scientific communities by reducing development cost, saving time to develop software, and improving the quality of the software through collaboration between software developers and end users.

4.3.1. Software Accessibility: Location, License, and Citation

Location (*S1*). Software can be made public by hosting it in code repositories, such as GitHub [*GitHub*, 2015a], SourceForge [*SourceForge*, 2015], and Bitbucket [*Bitbucket*, 2015]. These repositories do not necessarily guarantee persistence. For example, Google Code was a popular code repository that was discontinued. On the other side of the spectrum, GitHub teamed up with Zenodo to offer DOIs for specific versions of codes that are worth archiving permanently. Code repositories offer version control systems to facilitate code evolution and collaboration among developers as well as users. These code repositories may be challenging to learn, and a reasonable compromise is to make software accessible from a shared repository such as figshare and Zenodo.

License (*S2*). Copyright automatically applies to software when it is created, which grants the creator exclusive rights over the software as intellectual property. An open source license is a mechanism for releasing that copyright while retaining some control by creators over how they want their source code to be reused and acknowledged. Open source licenses specify whether the creator allows modifications of source code and/or the distribution of the modified source code under the same terms as the license of the original source code. These licenses also specify whether the creator wants to be acknowledged when the software is reused. The Open Source Initiative (OSI) offers widely used open source license options [*Open Source Initiative*, 2015], such as the GNU General Public License (GPL), the MIT license, the Berkeley Software Distribution License (BSD), and the Apache Public License (APL). Without a license, the creator is not protected by reuse of their software in ways they did not intend it to be.

Citation (*S3*). Citation of the software can assure that the developers get credit. Like with data, software can be cited through a DOI or a PURL. Some software repositories assign DOIs for particular software versions. For example, GitHub offers DOIs through the Zenodo data repository [*GitHub*, 2015b]. Some researchers choose to use data repositories to publish their code and get a DOI for software citation, keeping the code and the data in the same site.

4.3.2. Additional Requirements and Issues

Domain-specific software repositories. Software repositories for model software in geosciences include the Community Surface Dynamics Modeling System (CSDMS) [*The Community Surface Dynamics Modeling System*, 2015; *Peckham et al*., 2013], the Earth System Modeling Framework (ESMF) [*Earth System Modeling Framework*, 2015; *Hill et al*., 2004], the Computational Infrastructure for Geodynamics (CIG) [*The Computational Infrastructure for Geodynamics*, 2015; *Morozov et al*., 2006], and the VHub collaborative volcano and risk mitigation hub [*VHub*, 2015]. However, they do not include other ancillary software, such as code for data preparation or data reformatting. A great advantage of using these repositories is that they often enable scientists to integrate and run models at scale. Another significant advantage is that they enable model coupling (i.e., executing several models in consonance) through the specification of common interfaces and automatic regridding to standardize the granularity and scale of the models [*Peckham*, 2015].

Making software executable by others. Although it is recommended that code is shared as it is written [*Barnes*, 2010], there are a few best practices for preparing code for publication. File paths or variable settings may be better done as parameters in configuration files, so that when those need to be changed the source code itself does not have to be changed. When possible, code should not have dependencies on the particular operating system or directory structure, so if there are any it is worth investigating general ways to accomplish the same using more portable commands. The dependencies of the software on other libraries or programs must be explicitly documented. Another valuable step in enabling others to run software is to provide test cases that include data files that are known to work with the software, to explain the steps involved in the execution, and the expected results. However, despite best efforts put in preparing users' guides and sharing test cases, the ultimate voucher for the enablement of others to execute software is user feedback. A major advantage of using software sharing sites is that they enable this kind of user feedback. This feedback is perhaps the highest benefit of the time-consuming process of teaching others how to execute software that will eventually lead to better software.

Software updates. Another aspect of third-party software execution, albeit often overlooked, is the capacity to enable automatic installation and execution of software when updates are made. This is called Continuous Integration (CI), and there are a variety of tools to support it (e.g., Travis-CI [*Travis-CI*, 2015] and Jenkins [*Jenkins*, 2015]). As software grows in size or in number of contributors, the automation of repetitive steps such as installation and testing can lead to faster debugging and significant time savings. Enabling such capabilities is relatively straightforward once the installation steps are fully described and test cases are made available. The specification of these instructions can all be included in a series of small simple instruction files (e.g., shell scripts). Therefore, the relatively small additional burden of translating software dependency and short tests into a machine readable format can have great benefits.

Legacy software. Like data and other artifacts of research, it is common for software code to undergo a period of use by an individual or small group of researchers only to be abandoned, lost, or become outdated. Yet these "legacy" codes may be important assets to some studies and can aid researchers if unearthed and updated for long-term access and reuse. In many cases, with modest effort, a legacy code can be documented and potentially translated or moved into a maintainable code repository. Recreating or refactoring a program may introduce errors, bugs, or other issues not present in the original code. Yet after committing the time and effort to develop a useful code, it is worth investing the additional effort needed to facilitate its reuse in the future. Documentation of recovered software can aid future development and maintenance or reuse of that code. Common approaches to document software systems include writing natural language documents, creating formal specifications, producing standard design documents and providing interpretable test cases [*Tonella and Potrich*, 2005]. Any of these documentation formats will always be useful to a researcher attempting to revive legacy code. A difficulty in reusing legacy code arises when the existing documentation does not match the actual code. Another major difficulty is that to run some legacy code again may require recreating the older versions of runtime libraries or operating systems, which may not be available.

4.3.3. Software Accessibility: Recommendations

For scientists that develop software routinely, it makes good sense to learn to use a proper software repository (such as GitHub, SourceForge, and Bitbucket) to develop the code there. The repository will provide version control and other issue-tracking facilities that facilitate the management of the code as it evolves over time. These software repositories collect basic software metadata, such as authors, contributors, and versions. Licenses should be always specified, preferably open source permissive licenses such as the Apache License.

For many scientists, who do not routinely develop code or who find code repositories to be complex and challenging to learn, a very simple approach is to use a data repository to archive software. When a paper is written, whatever version of the software is used to generate the results should be the version archived. General metadata will be provided as for any data set, although care should be exercised to provide a software license (an open source permissive license like the Apache License is best) and to specify a version.

When using software developed by third parties that is not in an archival repository, it is important to open a dialogue to get the software to be properly accessible in an archival site, with a unique identifier for the version, and a citation.

Provenance is a major component of metadata no matter what data repository is used. It includes details about how the data were collected or generated, such as the instruments, processes, and entities involved. In this paper we focus on computational provenance, specifically the computations performed to generate a new data set.

4.4. Documenting Software by Specifying Metadata

Metadata documentation describes software so that others can find the software, understand what it does, run it, do research with it, get support, and contribute to future software development. It is useful to distinguish between *code repositories* and *software registries*. The code itself can be deposited in a code repository (such as those mentioned in section 4.3), and the metadata can be stored in one or more software registries that are linked to the code repository entry for the software. Documenting software within a software registry helps the software author describe their product while also making it more discoverable and open to use by a larger community. In geosciences, model repositories often serve as software registries and collect extensive metadata (e.g., CSDMS). General software registries, such as OntoSoft [*Gil et al*., 2015], can be linked to code repositories and automatically extract metadata from them.

4.4.1. Documenting Software: Function, Execution, Testing, and Updates

Function (*S4*). This describes the intended use of the software, its purpose, and function. The exact inner workings and modeling details may be very complex and be best described in a scientific article, but this kind of metadata highlights the main usage characteristics of the software so others understand what to use it for. Representative information is needed from the simplest perspective of clearly labeling units of measure, all the way to documenting data models and core algorithmic structures to communicate the underlying assumptions, key values, relational definitions among attributes, and fundamental descriptions used for a specific research endeavor.

Execution (*S5*). This metadata points to documents that describe what is needed to install and run the software, as well as any runtime dependencies and requirements (e.g., libraries).

Testing (*S6*). This metadata refers to test data provided to enable others to run the software and check whether it works. Testing information should include input data and parameter configurations, as well as the output data that should be expected if the software is running correctly.

Updates (*S7*). Any commitments of support for software are useful to those considering using it. This can include a specification of a point of contact to submit bug reports and requests for extensions, a mailing list to send questions and get help with any potential problems, and a description of how any future releases are planned and disseminated.

4.4.2. Additional Requirements and Issues

Domain-specific software metadata. To describe the function and purpose of scientific software, it is best done using standard vocabularies in the domain. The variables and parameters used within a piece of code would then be in alignment with standard naming rules, such as the CSDMS Standard Names [*Peckham*, 2014].

4.4.3. Software Documentation: Recommendations

When a code repository, such as GitHub, SourceForge, and Bitbucket, is adopted to make the software accessible, the repository offers mechanisms that are traditionally associated with software documentation, such as README files, test data sets, wikis, and known bug reports.

A software registry is an additional mechanism that is useful to provide additional documentation that is not present in code repositories. CSDMS, ESMF, CIG, and VHub, mentioned in section 4.3.2, collect software metadata that is useful for others to understand when reusing the software. A domain-independent and extensive software registry is OntoSoft, mentioned in section 4.4. It collects extensive software metadata that is then mapped to an ontology to enable search and discovery. If desired, OntoSoft can export the software metadata as HTML, XML, and RDF so it can be included in the code repository where the software resides.

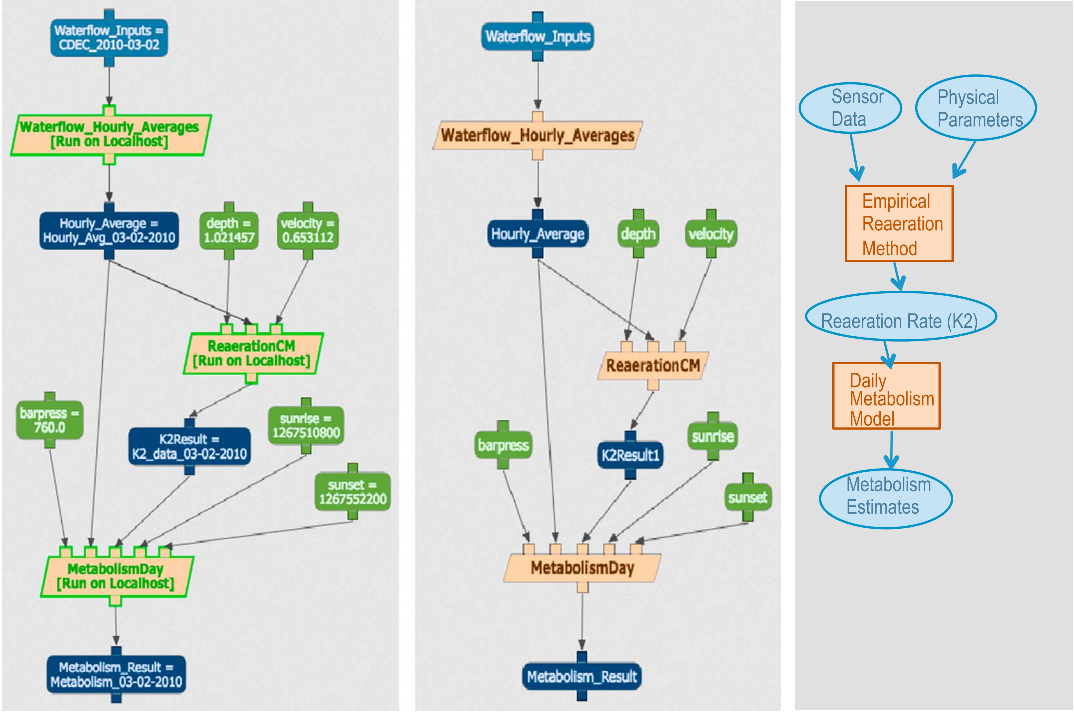
4.5. Documenting the Computational Provenance of Results

In this article, we focus on computational provenance as an explicit documentation of the data used and the processing performed to reach a scientific result. Computational provenance fully links together all of the (digital) objects used in the GPF, going from data, through any other software or code, and finally to completed results and figures. We define computational provenance as the documentation of the inputs, processing steps, and outputs of a particular scientific result. This may be implemented through a provenance record, an automatically generated data structure that fully links together all of the (digital) objects used in the GPF, going from data through any software, and finally to completed results and associated visualizations. These provenance records should also include any additional details on software configuration or specific parameter values that are necessary to reproduce the results of the GPF. Although tools such as workflow systems automatically capture computational provenance, programming environments do not typically record provenance. In such cases, trace logs (i.e., text lines describing an execution process) can be generated through print statements in the code that capture key information about the computations executed to generate the results reported in the paper.

While provenance is concerned with actual events that have occurred in the past, the general method (or computational workflow) is concerned with reusable processes that can be executed in the future either with existing or future data. The distinction between these two concepts is shown in Figure 2. Figure 2a shows the computational provenance to obtain whole-stream metabolism estimates from sensor data collected in March 2010, using specific parameter values for that period. In contrast, Figure 2b shows the general workflow or method for calculating the metabolism estimates using sensor data and parameters for any time period. Figure 2c shows a high-level workflow diagram created by hand, showing only some of the steps and some of the parameters. Traditionally, this kind of information is described in the “Methods” section of a paper, usually at a high-level and in a text format. Text is a limited medium to fully convey the complexity inherent in computational research and should therefore be complemented with trace logs (execution summary) or complete provenance records.

4.5.1. Documenting Computational Provenance

Trace logs and provenance records (*P1*). A trace log captures execution events and information typically for debugging purposes. Trace logs are often generated through print statements of any important information


Waterflow_Inputs = CDEC_2010-03-02
Waterflow_Hourly_Averages [Run on Localhost]
Hourly_Average = Hourly_Avg_03-02-2010
depth = 1.021457
velocity = 0.653112
ReaerationCM [Run on Localhost]
barpress = 760.0
K2Result = K2_data_03-02-2010
sunrise = 1267510800
sunset = 1267552200
MetabolismDay [Run on Localhost]
Metabolism_Result = Metabolism_03-02-2010
Waterflow_Inputs
Waterflow_Hourly_Averages
Hourly_Average
depth
velocity
ReaerationCM
barpress
K2Result1
sunrise
sunset
MetabolismDay
Metabolism_Result
Sensor Data
Physical Parameters
Empirical Reaeration Method
Reaeration Rate (K2)
Daily Metabolism Model
Metabolism Estimates

(a) Computational provenance of an execution run

(b) Reusable workflow corresponding to the general method

(c) High-level workflow diagram

Figure 2. Different approaches to document the methods in an article: (a) computational provenance includes the inputs, processes, and outputs for a particular execution. (b) computational workflow showing the inputs, processes, and outputs for any execution thus providing the general, reusable method; (c) a high-level workflow diagram showing major steps in the method. Both Figures 2a and 2b were generated automatically by a workflow system, while Figure 2c was drawn by hand.

about the execution and may be generated by the programming environment. They can show conceptually how the data were used by the software, what the parameter settings were, and what intermediate and final results were obtained. Computational provenance refers to the explicit data structures that record the computational steps used and the dataflow among them. Computational provenance records can be obtained from workflow systems [*Taylor et al.*, 2006; *Gil et al.*, 2007; *Deelman et al.*, 2012; *Moreau et al.*, 2014], such as Pegasus [*Deelman et al.*, 2005], Taverna [*Oinn et al.*, 2006], Vistrails [*Callahan et al.*, 2006], and Kepler [*Ludaescher et al.*, 2006]. A provenance standard, such as W3C PROV [*Gil et al.*, 2013; *Moreau et al.*, 2013], may be adopted to enable analysis and reuse of provenance records.

Provenance summaries (*P2*). Trace logs and provenance records can be long and not suitable for human consumption. When that is the case, a provenance summary can be created to highlight the main computational steps in the form of a subset of key statements, a graphical sketch, or a table. These can be very effective to convey important details to the reader of the article, although only the detailed provenance records and trace logs contain enough detail to enable full reproducibility.

Versions (*P3*). Software often evolves and can have many releases over the years. Different versions of a piece of software may offer different functionalities, some may disappear over time, and some may be incorporated in a new release. Because of this, it is very important that the versions of all the software used to obtain results in the paper be clearly indicated within the computational provenance record.

Parameters (*P4*). For each software used, the detailed command line invocations and parameter values used should be specified. The parameters may be in configuration files and should be provided alongside with the data used in the paper.

4.5.2. Additional Requirements and Issues

Publishing and citing computational provenance. Ideally, the computational provenance records for a paper would be published in a public repository and cited in the article. This would put computational provenance at the same level of importance as the data and software used in the paper, which is appropriate. However, unlike data and software, there are no public shared repositories for computational provenance records. Since provenance records have a unique structure that should be searchable and comparable, it is possible that in the future shared provenance repositories may emerge to enable provenance discovery and reuse. A temporary solution to this issue is to publish computational provenance in a general data repository.

Data preparation steps. Data preparation aspects are often not mentioned in articles but are crucial to documenting the computational provenance of results in a proper manner. Data preparation can take a significant amount of effort and may include important choices regarding quality control, imputation for missing values, and conversion to standards. In documenting the computational provenance of results in a paper, all these steps are important for reproducing the work.

Manual steps to create figures. A special case is the creation of data visualizations that go beyond the computational generation of results. Indeed, figure production is often as much an artistic endeavor as it is a computational process. Thus, it is incumbent upon the author to identify if and when certain visualization steps need to be specified in order for readers to fully reproduce a paper's results. Otherwise, providing the source data is sufficient. While sometimes only a manual process is possible (e.g., using a GIS to create a map), many tools (e.g., MATLAB [*MATLAB*, 2015]) allow manual creation of a figure and will then allow subsequent generation of the code needed to automatically generate it. Taking this concept one step further, new tools (e.g., Plotly [*Plotly*, 2015]) can generate fully shareable, interactive data plots.

Size of the computational provenance records. In some cases, the computational provenance and trace logs records may be very large and complex to describe. Some research involves running dozens of codes, and in those cases computational provenance can be documented at varying levels of detail. At the simplest level is a provenance summary that highlights the most important steps and the data flow among them. The detailed trace logs and provenance records can be provided in a repository. Other research may involve running experiments with hundreds of data sets. In those cases, the computational provenance of a few runs can be documented in detail, and the others described at a higher level.

4.5.3. Computational Provenance: Recommendations

Most software environments used by scientists today do not capture computational provenance. Commonly used frameworks such as spreadsheet software, R and MATLAB, and other programming environments have no mechanisms for capturing what functions were executed and with what parameters and data. This is slowly changing (e.g., with tools like IPython Notebook and Jupyter Notebook), but capturing computational provenance is still not an easy task for a scientist today. When using a system that does not track computational provenance, the only solution is to track it by hand by adding trace or print functions to the code to generate trace logs.

Increased adoption of electronic notebooks (e.g., IPython and Jupyter Notebook) and workflow systems (e.g., Kepler, Pegasus, WINGS) that produce computational provenance traces is important to increasing transparency and reproducibility. In some cases, the computational provenance captured by these systems can be exported and stored in public repositories. The best option would be a system that exports provenance using a standard such as W3C PROV or ISO 19115 to provide interoperability across notebook and workflow systems. Ideally, these computational provenance traces would have detailed information about software versions as well as DOIs to each data set and software used. A short summary of these computational provenance records of trace logs should be included in the main paper, pointing to supporting information that provide more details and key excerpts.

4.6. Documenting the Methods

While the computational provenance describes specific data and method setup that led to a specific result (e.g., a figure in the paper), the method describes the general way to get similar results with other data and/or parameters. Computational provenance was described in section 4.5, this section describes how to document the method. Scientific articles typically include a "Methods" section that describes them. However, computational methodology should be explicitly documented.

4.6.1. Documenting Methods: Composition, and Dataflow

Composition (*M1*). A composition documents the various steps in the method in terms of how different pieces of software are used together in a pipeline. This composition is sometimes a set of sequential steps, but it may consist of many interconnected and interdependent steps. The composition can be indicated as a simple workflow diagram drawn by hand as a graph of computations (nodes) are linked by the dataflow among them. Alternatively, the composition can be formally specified as a computational workflow using a workflow system such as those mentioned in section 4.5.1. Each workflow system offers different capabilities that suit different requirements and communities [*Deelman et al*., 2012; *Taylor et al*., 2006].

Dataflow (*M2*). The dataflow between the steps indicates how initial data would be processed by software, what intermediate data sets would be generated, and how the results would be obtained. The composition may already include this information, but if it does not, it should be provided.

4.6.2. Additional Requirements and Issues

Steps involving samples. In geosciences, some of the steps may involve collecting and handling samples or processing materials in the laboratory. These steps are not computational and may involve manual intervention. It is important to document these steps in as much detail as possible, particularly where they result in digital data that is to be processed computationally.

4.6.3. Methods: Recommendations

The general method followed to generate the results in the paper can be documented as a high-level workflow diagram, which is a diagram that captures the dataflow across components. The diagram should mention specific software by name and version, and the types of data that can be used. The workflow diagram itself can be published in a general data repository (e.g., figshare), and assigned a DOI and citation.

When using a workflow system, the general method will be captured as a formal dataflow structure, which can be published in a general workflow repository (e.g., myExperiment.org), a general data repository (e.g., figshare, Dryad, Zenodo), or a domain-specific data repository that accepts materials other than data. A graphical workflow diagram is also generated by workflow systems, which can be included in the entry and/or in the paper itself. A persistent identifier should be obtained so the workflow can be cited.

The main text of the paper should describe the high-level workflow diagram and how the general method should be adapted to other cases. If the descriptions are lengthy, they can be included as supporting information. Finally, the workflow should be cited in the paper, with a pointer to the archival repository in the citation.

4.7. Author Identification

Much like data sets and software must have a permanent unique identifier, researchers must have one as well. The name alone is not sufficient to identify a person uniquely, and the institution or other affiliation information helps but it is often transient information.

4.7.1. Author Identification: Unique Identifiers

Unique identifiers (*A1*). Authors should get a persistent unique identifier that is associated with all their digital research products. A common identifier for researchers is the Open Researcher and Contributor ID (ORCID) [*Open Researcher and Contributor ID*, 2015], which can be easily obtained from orcid.org.

4.7.2. Additional Requirements and Issues

Authorship of data and software citations. Author identifiers should also be used in the data and software citations of the article. Ideally, all digital research products of a researcher should be linked to their identifier.

Authorship of data and software contributions. Key contributors of data and software who are not authors of the GPF should also be assigned a persistent unique identifier to be used in the attribution of data and software cited in the article. This helps create a healthy ecosystem of credit and recognition through citation to those who do not coauthor scientific publications.

References to funding sources and institutions. It is useful to refer to funding sources in the acknowledgments using a persistent unique identifier. When authors list affiliations with institutions, these can also have a persistent unique identifier.

4.8. Summary: Preparing a GPF for Publication

Table 2 shows a high-level description for the simplest and advanced approaches that we recommend for GPF authors, aligned with the GPF Author Checklist shown in Table 1. The simplest approach typically takes a few hours to implement and is feasible to adopt for any manuscript that an author is preparing. The advanced approaches require more investment of time and effort, and go beyond the preparation of a manuscript to impact their routine practices for data, software, and computational provenance management.

Table 2. A High-Level Roadmap for Authors of a Geoscience Paper of the Future (GPF), With Pointers to Popular Resources for Publication and Licensing of Scientific Research Products

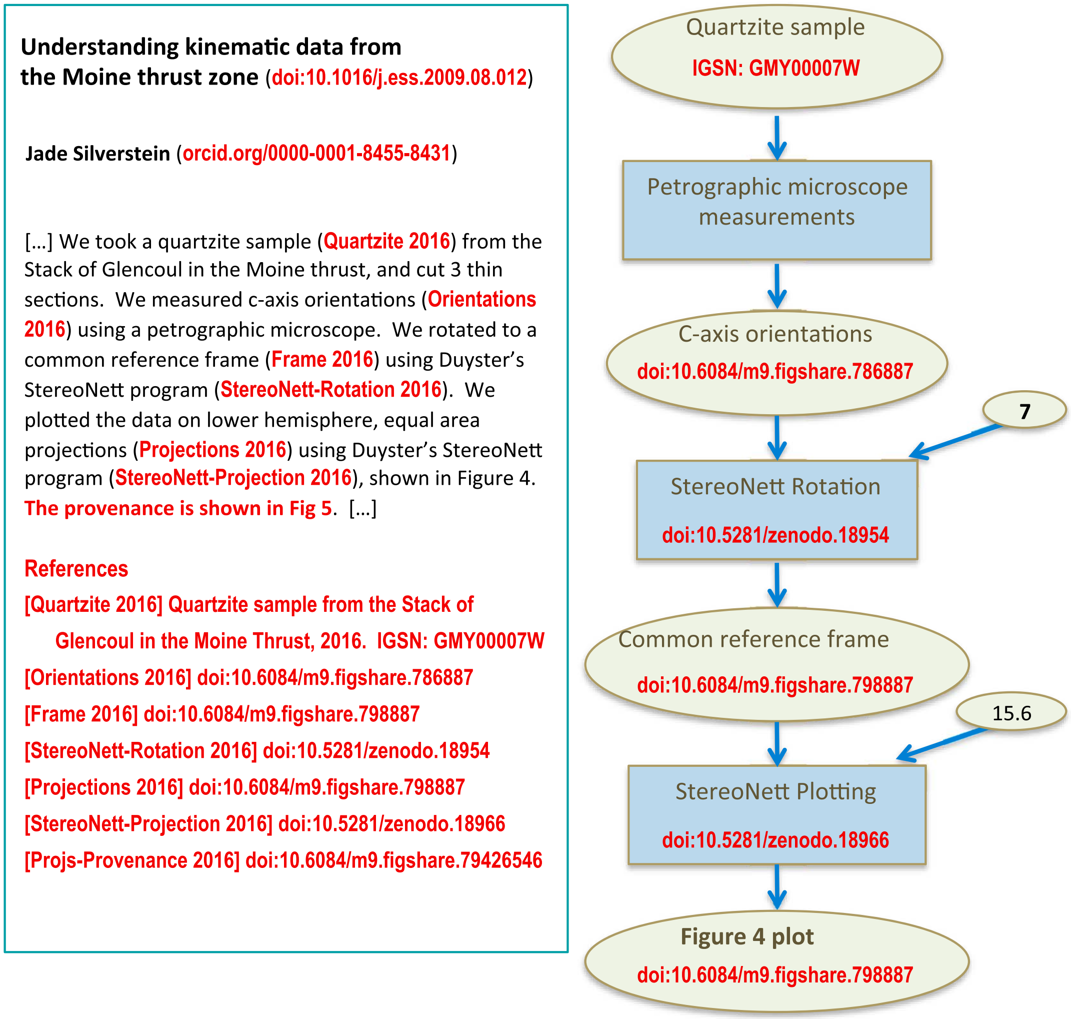
Data		
Available in a public repository, with metadata, a license specifying conditions of use, and citable using a unique and persistent identifier.		
Data Accessibility		What to Show in the Paper
Simplest Approach	Advanced Approach	
1. Choose a general repository (e.g., figshare, Zenodo, Dryad, Pangaea, etc.) 2. Create a public entry for your data set with a persistent unique identifier 3. Specify basic metadata (name, authors) • Including license—choose from creativecommons.org 4. Upload/point to the data 5. The repository will give you a data citation	1. Find a repository that your community uses, if there is not one then organize one! 2. Create a public entry for your data set with a persistent unique identifier 3. Specify the basic metadata required by that repository • Including license—choose from creativecommons.org 4. Upload/point to the data 5. Get a data citation from the repository	• Cite data set in the paper references • Citation includes data set name, creators, publication date, repository name, persistent identifier, and time of retrieval • If there is a separate paper about a data set, cite it as well • Mention in the text that the persistent identifier site has the metadata and includes a detailed description of the data • Mention availability of related data sets
Data Documentation		
Simplest Approach	Advanced Approach	
• Data sets should have at least domain-specific keywords • Include metadata and documentation that will help reuse	• Domain-specific metadata should be well documented using metadata standards for that community	
Software		
Available in a public repository, with documentation, a license for reuse, and a unique and citable persistent identifier. This includes any ancillary software for data reformatting, data conversions, data filtering, and data visualization.		
Software Accessibility		What to Show in the Paper
Simplest Approach	Advanced Approach	
1. Create a public entry for your software with a persistent unique identifier 2. Post on your web site and use a PURL, upload to a data repository and get a DOI 3. Specify basic metadata 4. Include license—choose from opensource.org/licenses, e.g., Apache 5. Specify desired citation	1. Learn to use a code repository that allows version tracking and collaborative software development (e.g., GitHub, BitBucket, etc.) 2. Create a public entry for your software with a persistent unique identifier 3. Specify basic metadata • Include license—choose from opensource.org/licenses, e.g., Apache 4. Get a software citation from the repository.	• Cite software in the paper references ○ Citation similar to data but includes software version • If there is a software paper, cite it • Mention that the persistent identifier location for your software points to its metadata • Optionally, include the software metadata as supporting information
Software Documentation		
Simplest Approach	Advanced Approach	
• Describe as much metadata as will help reuse ○ Document basic metadata including authors, contributors, version, license, and release date	• Use a software registry ○ www.ontosoft.org/portal, csdms.colorado.edu, etc. ○ Guides through questions to provide metadata • Save the metadata as HTML, XML, etc. • Post the metadata on your code site	

Table 2. (continued)

Computational Provenance and Methods

Documented for all results by explicitly describing the series of computations and their outcome with a trace log (or computational provenance record) and a high-level workflow diagram (or a formal workflow), possibly in a shared repository and with a unique and persistent identifier.

Documentation		What to Show in the Paper
Simplest Approach	Advanced Approach	• Describe workflow in text and provide a workflow diagram ○ Optionally, provide the formal workflow or lab notebook, use a persistent identifier, and cite it • Include a provenance summary as supplementary material, or use a persistent identifier and cite it ○ Optionally, include the trace log pr the computational provenance records using a standard (e.g., W3C PROV), also with a persistent identifier and cited
1. Provide a trace log or a formal computational provenance record 2. Provide a provenance summary in text • Data + software • Specify unique identifiers for data and software, versions, credit all sources 3. Develop a high-level workflow diagram • Capture dataflow across major computational steps as a diagram	1. Provide a computational provenance summary in text • Data + software • Specify unique identifiers for software versions 2. Develop a high-level workflow diagram • Capture high-level dataflow across major computational steps 3. Specify the formal workflow using a workflow system, electronic notebook, etc. • Command lines + parameter values • Dataflow across components 4. Include the computational provenance record • If generated automatically, preferably using a standard (e.g., PROV) 5. Publish the workflow and computational provenance record in a repository (e.g., myExperiment.org) or a data repository 6. Get a unique persistent identifier for the workflow, the computational provenance, or both	



Understanding kinematic data from the Moine thrust zone (doi:10.1016/j.ess.2009.08.012)

Jade Silverstein (orcid.org/0000-0001-8455-8431)

[…] We took a quartzite sample (Quartzite 2016) from the Stack of Glencoul in the Moine thrust, and cut 3 thin sections. We measured c-axis orientations (Orientations 2016) using a petrographic microscope. We rotated to a common reference frame (Frame 2016) using Duyster's StereoNett program (StereoNett-Rotation 2016). We plotted the data on lower hemisphere, equal area projections (Projections 2016) using Duyster's StereoNett program (StereoNett-Projection 2016), shown in Figure 4. **The provenance is shown in Fig 5**. […]

References

[Quartzite 2016] Quartzite sample from the Stack of Glencoul in the Moine Thrust, 2016. IGSN: GMY00007W

[Orientations 2016] doi:10.6084/m9.figshare.786887

[Frame 2016] doi:10.6084/m9.figshare.798887

[StereoNett-Rotation 2016] doi:10.5281/zenodo.18954

[Projections 2016] doi:10.6084/m9.figshare.798887

[StereoNett-Projection 2016] doi:10.5281/zenodo.18966

[Projs-Provenance 2016] doi:10.6084/m9.figshare.79426546

Figure 3. An example of how a Geoscience Paper of the Future (GPF) would refer to data, software, and computational provenance.

However, they have significant benefits that we describe in the next sections including reproducibility, reusability, and due credit.

Figure 3 includes an illustration of how a GPF would cite data sets, software, and computational provenance. This is a simple example where a rock sample is analyzed in the microscope and then the images are transformed to get particular projections. On the right-hand side, a simplified dataflow diagram of the computational provenance record is included, showing the unique persistent identifiers of the initial sample, the data sets, and the software. In the text, the physical sample, the data sets, and the software are all cited. The computational provenance itself is also cited, as the detailed record is in a shared repository. The list of citations appears in the References of the paper, as shown in the bottom left.

In summary, to prepare a GPF for publication, several key aspects of the research must be documented:

1. *Data accessibility and documentation*. Generally, data that are used from the initial point of an analysis or evaluation, through any significant intermediate results, and data generated for the final results of research should be made accessible and documented. To achieve data accessibility, (D1) data should be published in an accessible location with a permanent unique identifier, (D2) data sets should be published with an accompanying license to delineate acceptable reuse and dissemination options, and (D3) data sets should be cited in the accompanying GPF. In addition, data documentation is needed to assure that the representative values and parameters can be understood by others. Basic documentation for data

should include the following (D4) specification of general purpose metadata, (D5) data set characteristics should be explained in detail, and (D6) data set origins and availability of related data sets should be documented.

2. *Software accessibility and documentation*. Like data, the software used to process initial data and to generate any intermediate or final results for research needs to be documented and shared with attention to a similar set of recommendations. (S1) The code and an executable version of software should be published in an accessible location or repository with a permanent unique identifier, (S2) assigning a license that defines acceptable use and distribution, and (S3) citing the software in the article of reference or GPF. Documenting software requires (S4) a clear description of the function and purpose of the software, (S5) descriptions of download and execution requirements or dependencies, (S6) documentation describing how to test and reuse with new data, and (S7) a description of the expected levels of software support, if any, for extensions and updates.
3. *Provenance documentation*. The computational provenance of an information source reports the origination and chain of transformations used to generate all computational results reported in a GPF article, including figures, tables, and other findings. To assure complete documentation of computational provenance GPF authors should include descriptions of the (P1) derivation traces of newly generated data from initial data, (P2) traces of software executions used for newly generated results, (P3) versions and configurations of the software, and (P4) parameter values used to run the software.
4. *Methods documentation*. Methods that are applied to data sets or used to analyze information related to the scientific research, particularly computational methods applied to data other than the data in the paper, should be documented. Methodological documentation should present information that enables the replication of computational approaches and requires (M1) reporting on the compositions of software that form a general reusable method and (M2) a description of dataflow across software components.
5. *Author identification*. Assuring that research efforts are transparent, reproducible, and accessible while also connecting credit for the work and impact of a particular investigator can only be achieved if (A1) each author is linked to the products for the research through the use of a permanent unique identifier.

When all the research products are open and shared, it makes good sense that the paper itself is open access so it is accessible to everyone. There is a general movement by funding agencies to require that papers be accessible, at least after a few months of publication. Publishers offer open access journals where the contents of the articles are freely available to readers.

5. Discussion

As we mentioned earlier, major challenges to improve open science, reproducibility, and digital scholarship in geosciences include the lack of clarity on best practices, the lack of awareness of those best practices, and the level of effort involved. The vision of Geoscience Papers of the Future helps address these barriers through a concise and practical articulation of requirements and associated best practices. In our own experiences in writing our own GPFs, the availability of these guidelines turned impossible into manageable.

Having a set of guidelines and appropriate training expedites the process of producing a complete GPF. Given the broad extent of the geosciences, researchers in their particular areas of study need to communicate among themselves to finesse their own definition of a complete GPF. There are many choices for sharing and documenting data and code, and each field of study may define the aspects of our proposed GPF vision that best suit their needs. Defining minimum requirements and preferred repositories for a particular research area would make digital objects more usable.

From our perspective, the greatest roadblock in implementing the proposed vision for geoscience papers of the future is the lack of knowledge in the community about best practices and available tools to implement them, and lack of critical mass usage. Increased communication and education on existing technology, potential limitations, and best practices will be key to making this vision a reality.

The level of effort involved in following these best practices is not negligible, but it is also not unreasonable. There is no question that there is a learning curve, both in grasping the basic concepts behind the best practices and in implementing them with an approach and tools that suit an individual researcher. The more scientists that adopt these best practices and have experience writing a GPF, the easier it will be for others to find a colleague within arm's reach who can help with shortcuts and commonly used tools. These best practices are technically very simple, so they are within reach for everyone to learn in a few hours.

The effort required is greatly reduced by available tools and platforms. There are already many tools for publishing data and software, for documenting metadata, for obtaining identifiers, for capturing computational provenance and workflow (although this remains even less integrated in geoscience workflows), and many other aspects mentioned in this article. Although they are not yet seamlessly integrated with geosciences practice, they improve constantly and for many scientists they become a staple once they are discovered.

The investment related to the implementation of these best practices can have many benefits to the authors. Data sharing makes authors double-check their work, improving science at the first stage as well as future reuse. Software sharing can improve the practices of scientists who are informally-taught coders. A payoff for sharing digital objects is that it improves science by making available better quality products resulting from the spontaneous feedback and sometimes curation provided by users of the shared digital objects. Some of the best practices can be implemented with tools that save time and can generate some of the content for the article (e.g., writing the Methods section by showing a workflow and describing it).

With very minimal effort, it is still possible to implement an important subset of the best practices recommended here. Each scientist should find the right balance with regard to the effort needed and the best practices that are suitable for their own needs, their field of research, and the broader community.

It is harder to invest the effort as an afterthought of the research and to document the paper once the work has been completed. It takes more time to write a GPF in retrospect than it would to document the work from the beginning. It is often said that the quality of data description and documentation is inversely proportional to the time since data collection and analysis, so it is important for scientists to continuously describe and document data whenever possible. A continuous process of provenance documentation may, in fact, be a helpful practice for authors to ensure that all data and methods are fully understood, documented, and shared before any results are interpreted and considered for publication.

Federal agencies have formed working groups to improve open access to publications, data, software, and generally any research products. Another important area of improvement is reproducibility and transparency in science. As these discussions proceed, we believe the outcomes will be consistent with the recommendations put forward in this article. Our aim here is to lay the foundation for more open and reproducible science, acknowledging that this requires extra effort on the side of the authors and conveying that there are clear rewards in doing so. Authors of GPFs are in an excellent position to write Data Management Plans in their proposals that will cover all the bases that federal agencies are currently considering.

Scientists may soon be forced to document their papers in a manner similar to the GPF best practices described here. Publishers and funders are increasingly requiring the kinds of documentation that a GPF would include, in order to improve open access to research products, reproducibility, accountability, and credit. The best practices discussed in this paper can be easily taught to junior researchers who can adopt them in their daily practice and make them routine in their work.

6. Conclusions and Future Work

This paper motivates the vision for geoscience papers of the future and describes best practices and their recommended implementations for GPF authors based on open science practices, reproducibility, and digital scholarship. It also articulates 20 specific recommendations for GPF authors to facilitate their uptake in the geoscience community.

While we have endeavored in this paper to disseminate best practices and available tools, a major roadblock is that they are not fully integrated into the processes and systems currently used by geoscientists. Many of these tools and platforms are insular and the overall process for writing a GPF requires using several of them. There are lots of moving parts that need to be coordinated, which can be a challenge. Publication embargo dates complicate matters and are not handled by many of these tools. As a result, they introduce a burden on the geoscientist and, although many will agree with the need for reproducibility and transparency, the barriers remain high. Close collaborations between computer scientists and geoscientists are needed to develop tools that reduce these barriers by becoming an essential part of geoscience research workflows.

Beyond the GPF vision, additional enhancements to geosciences papers include making the methods composable with one another, making the main claims of a paper explicit in formal logic, and comparing alternative hypotheses or contradictory results across papers. The more explicit and documented papers are, the more likely it is that we will have automated means to answer common questions such as "What is known in the literature about *X*?", which scientists face all the time but take a lot of effort to research. Such explicit and formal representations of papers would also support intelligent systems in geosciences [*Gil and Pierce*, 2015]. These explicit representations of the content of papers would significantly improve the productivity of geoscientists and greatly facilitate cross-disciplinary collaborations. Ultimately, these explicit representations of scientific knowledge will significantly amplify the capabilities and impacts of geosciences research.

Acknowledgments
This work was supported in part by the U.S. National Science Foundation under awards ICER-1343800 and ICER-1440323. Cédric H. David was supported by the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. Ibrahim Demir was supported by the Iowa Flood Center at the University of Iowa. Jonathan Goodall and Bakinam Essawy were supported by a grant from the National Science Foundation with award 0940841. Huikyo Lee's work was done as a private venture and not in the author's capacity as an employee of the Jet Propulsion Laboratory, California Institute of Technology. Allen Pope was supported by USGS contract G12PC00066. Sandra Villamizar was supported by the National Science Foundation with awards CBET-1204841 and 1336839. Mimi Tzeng was supported by a grant from NOAA through the Northern Gulf Institute (NGI). Xuan Yu was supported by a grant from the National Science Foundation with award IIA-1301765. The authors would like to thank other members of the OntoSoft project for their contributions, including Chris Duffy, Chris Mattmann, Scott Peckham, Erin Robinson, and other members of the EarthCube initiative and the geosciences community for their valuable comments, in particular, Ruth Duerr, James Howison, Matt Jones, Lisa Kempler, Matt Mayernik, and Greg Wilson. The authors would also like to thank Kate Musen for valuable comments on this manuscript and are very grateful for the feedback provided by the reviewers and editors. Because this article is a compilation of best practices and recommendations that appear elsewhere, no data were used or produced during the course of the work reported.

References

ACADIS (2015), The ACADIS gateway: An Arctic data repository. [Available at http://www.aoncadis.org, Last accessed 3 August 2015.]

Altman, M., and G. King (2007), A proposed standard for the scholarly citation of quantitative data, *D-Lib Mag.*, *13*(3/4), doi:10.1045/march2007-altman.

American Geophysical Union (2013), AGU publications data policy, December 2013. [Available at http://publications.agu.org/author-resource-center/publication-policies/data-policy/.]

Baggerly, K. A., and K. R. Coombes (2009), Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology, *Ann. Appl. Stat.*, *3*(4). [Available from http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoas/1267453942.]

Baggerly, K. A., and K. R. Coombes (2011), What information should be required to support clinical Omics publications?, *Clin. Chem.*, *57*(5), 688–690.

Baker, M. (2012), Independent labs to verify high-profile papers, *Nat. News*, doi:10.1038/nature.2012.11176.

Baker, S. G., A. K. Drake, P. Pinsky, H. L. Parnes, and B. S. Kramer (2010), Transparency and reproducibility in data analysis: The prostate cancer prevention trial, *Biostatistics*, *11*(3), 413–418.

Ball, A., and M. Duke (2012), How to cite dataset s and link to publications, DCC How-to Guides. Edinburgh: Digital Curation Centre. [Available at http://www.dcc.ac.uk/resources/how-guides - See more at: http://www.dcc.ac.uk/resources/how-guides/cite-datasets#sthash.MJQjNn3i.dpuf.]

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

Begley, C. G., and L. M. Ellis (2012), Drug development: Raise standards for preclinical cancer research, *Nature*, *483*, 531–533, doi:10.1038/483531a.

Bell, A. W., E. W. Deutsch, C. E. Au, R. E. Kearney, R. Beavis, S. Sechi, T. Nilsson, J. J. Bergeron, and the Human Proteome Organization (HUPO) Test Sample Working Group (2009), A HUPO test sample study reveals common problems in mass spectrometry-based proteomics, *Nat. Methods*, *6*(6). [Available at http://www.nature.com/nmeth/journal/v6/n6/full/nmeth.1333.html.]

Bitbucket (2015). [Available from http://bitbucket.org/, Last accessed 3 August 2015.]

Bonnet, P., et al. (2011), Repeatability and workability evaluation of SIGMOD 2011, *SIGMOD Rec.*, *40*(2), 45–48.

Bourne, P. (2010), What do I want from the publisher of the future?, *PLoS Comput. Biol.*. [Available at http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000787.]

Bourne, P. E., T. Clark, R. Dale, A. de Waard, I. Herman, E. Hovy, and D. Shotton editors (2011), Improving future research communication and e-scholarship, The FORCE 11 Manifesto. Available from http://www.force11.org.]

Callahan, S. P., J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo (2006), Managing the evolution of dataflows with VisTrails, Proceedings of IEEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow).

Cardamone, N., et cal (2009), Galaxy Zoo Green Peas: Discovery of a class of compact extremely star-forming galaxies 2009, MNRAS, 399, 1191.

CF (2011), NetCDF climate and forecast (CF) metadata conventions. [Available at http://cfconventions.org/, Last accessed 3 August 2015.]

Claerbout, J. (2006), Preface to SEP report 124, Technical Project Report, Stanford Exploration Project, 22 February 2006. [Available at http://sepwww.stanford.edu/data/media/public/sep//jon/repropreface.html.]

Claerbout, J., and M. Karrenbach (1992), Electronic documents give reproducible research a new meaning, 62nd Annual International Meeting of the Society of Exploration Geophysics., Expanded Abstracts, 92: Society of Exploration Geophysics, 601-604, 1992. [Available at http://sepwww.stanford.edu/doku.php?id=sep:research:reproducible:seg92.]

Coalition for Publishing Data in the Earth and Space Sciences (2016), Directory of repositories. [Available from https://copdessdirectory.osf.io/search/, Last Accessed 24 May 2016.]

Collins, F. S., and L. A. Tabak (2014), Policy: NIH plans to enhance reproducibility, *Nature*, *505*(7485), 612–613.

Committee on Data for Science and Technology (CODATA) (2013), Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data, CODATA-ICSTI Task Group on Data Citation Standards and PractOut of Cite, Out of Mind: The Current Sices, *Data Sci. J.*, doi:10.2481/dsj.OSOM13-043.

Costello, M. J., W. K. Michener, M. Gahegan, Z.-Q. Zhang, and P. E. Bourne (2013), Biodiversity data should be published, cited, and peer reviewed, *Trends Ecol. Evol.*, *28*, 454–461.

Creative Commons (2015), Available from http://www.creativecommons.org. Last accessed 3 August 2015.

Distributed Active Archive Centers (2015), The Earth Observing System Data and Information System (EOSDIS) Distributed Active Archive Centers (DAACs). Available at https://earthdata.nasa.gov/about/daacs, Last accessed 3 August 2015.]

Dash (2015), The dash tool: Data sharing made easy, Available from https://dash.cdlib.org/. Last accessed 3 August 2015.

DataCite (2015). [Available from https://www.datacite.org/, Last accessed 3 August 2015.]

Dublin Core Metadata Initiative (2012), Dublin core metadata terms. [Available from http://dublincore.org/documents/dcmi-terms/, Last accessed 3 August 2015.]

Deelman, E., et al. (2005), Pegasus: A framework for mapping complex scientific workflows onto distributed systems, *Sci. Program. J.*, *13*, 219–237.

Deelman, E., C. Duffy, Y. Gil, S. Marru, M. Pierce, and G. Wiener (2012), EarthCube Report on a workflows roadmap for the geosciences, National Science Foundation, Arlington, VA.

Dellavalle, R. P., E. J. Hester, L. F. Heilig, A. L. Drake, J. W. Kuntzman, M. Graber, and L. M. Schilling (2003), Going, going, gone: Lost internet references, *Science*, *302*(5646), 787–788, doi:10.1126/science.1088234, http://www.sciencemag.org/cgi/reprint/sci;302/5646/787.pdf.

DeRisi, S., R. Kennison, and N. Twyman (2013), The what and whys of DOIs, *PLoS Biol.*, *1*(2), e57.

De Roure, D., C. Goble, and R. Stevens (2009), The design and realizations of the myexperiment virtual research environment for social sharing of workflows, *Future Generation Comput. Syst.*, *25*, 561–567.

Diggle, P. J., and S. L. Zeger (2009), Reproducible research and biostatistics, *Biostatistics*, *10*(3), 405–408.

Donoho, D. L. (2002), How to be a highly cited author in the mathematical sciences. In-cites. [Available at http://www.in-cites.com/scientists/DrDavidDonoho.html.]

Donoho, D. L. (2010), An invitation to reproducible computational research, *Biostatistics*, *11*(3), 385–388, doi:10.1093/biostatistics/kxq028.

Donoho, D. L., and X. Huo (2004), BEAMLAB and Reproducible Research, *Int. J. Wavelets, Multi., Info. Process.*, *02*, 391, doi:10.1142/S0219691304000615.

Donoho, D., A. Maleki, I. Rahman, M. Shahram, and V. Stodden (2009), Reproducible Research in Computational Harmonic Analysis, *Comput. Sci. Eng.*, *11*, 8–18.

Downs, R. R., R. Duerr, D. J. Hills, and H. K. Ramapriyan (2015), Data stewardship in the earth sciences, *D-Lib Mag.*, *21*(7/8), doi:10.1045/july2015-downs.

Dryad (2015), Available from http://www.datadryad.org. Last accessed 3 August 2015.

Easterbrook, S. M. (2014), Open code for open science?, *Nat. Geosci.*, *7*, 779–781, doi:10.1038/ngeo2283.

Ellison, A. M. (2010), Repeatability and transparency in ecological research, *Ecology*, *91*, 2536–2539, doi:10.1890/09-0032.1.

eScience (2011), Transforming scholarly communication, Report of 2011 Microsoft Research eScience Workshop. [Available at http://msrworkshop.tumblr.com/, Last Accessed 31 July 2015.]

Federation of Earth Science Information Partners (2012), Interagency Data Stewardship/Citations/provider guidelines, Federation of Earth Science Information Partners (ESIP), 2 January 2012. [Available at http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines.]

Earth System Modeling Framework (2015), The Earth System Modeling Framework (ESMF). [Available from https://www.earthsystemcog.org/projects/esmf/, Last accessed 3 August 2015.]

Falcon, S. (2007), Caching code chunks in dynamic documents: The weaver package, *Comput. Stat.*, *24*(2), 255–261. [Available from http://www.springerlink.com/content/55411257n1473414/.]

Fang, C. F., and A. Casadevall (2011), Retracted science and the retracted index, *Infect. Immun.*, doi:10.1128/IAI.05661-11.

Fegraus, E. H., S. Andelman, M. B. Jones, and M. Schildhauer (2005), Maximizing the value of ecological data with structured metadata: An introduction to Ecological Metadata Language (EML) and principles for metadata creation, *Bull. Ecol. Soc. Am.*, *86*, 158–168, doi:10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.

Figshare (2015), figshare. Available from http://www.figshare.org. Last accessed 3 August 2015.

Fischer, D. A., et al. (2012), Planet hunters: The first two planet candidates identified by the public using the Kepler public archive data, *MNRAS*, *419*(4), 2900–2911.

FORCE11 (2014), *Joint Declaration of Data Citation Principles*, edited by M. Martone, and the Data Citation Synthesis Group, FORCE11, San Diego CA. [Available from https://www.force11.org/datacitation.]

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), e80278.

Garijo, D., Y. Gil, and O. Corcho (2014), Towards workflow ecosystems through semantic and standard representations, Proceedings of the Ninth Workshop on Workflows in Support of Large-Scale Science (WORKS), held in conjunction with the IEEE ACM International Conference on High-Performance Computing (SC), New Orleans, LA, 2014.

Gavish, M., and D. Donoho (2011), A universal identifier for computational results, *Procedia Comput. Sci.*, *4*, 637–647, doi:10.1016/j.procs.2011.04.067.

Global Change Master Directory (2015), The NASA's Global Change Master Directory. [Available at http://gcmd.nasa.gov/, Last accessed 3 August 2015.]

Gil, Y., and S. Pierce editor (2015), Report of the 2015 national science foundation workshop on intelligent systems for geosciences, To be published on November 2015. [Available at http://www.is-geo.org, Last accessed 3 August 2015.]

Gil, Y., E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers (2007), Examining the challenges of scientific workflows, *IEEE Comput.*, *40*(12), 24–32, http://www.computer.org/portal/web/csdl/doi/10.1109/MC.2007.421 (preprint available at http://www.isi.edu/~gil/papers/computer-NSFworkflows07.pdf)

Gil, Y., S. Miles, K. Belhajjame, H. Deus, D. Garijo, G. Klyne, P. Missier, S. Soiland-Reyes, and S. Zednik (2013), A primer for the PROV provenance model, Published as a W3C Working Group Note on 30 April 2013. [Available from http://www.w3.org/TR/prov-primer/.]

Gil, Y., V. Ratnakar, and D. Garijo (2015), OntoSoft: Capturing scientific software metadata, Proceedings of the ACM International Conference on Knowledge Capture, October 2015

GitHub (2015a), GitHub. Available from http://www.github.org. Last accessed 3 August 2015.

GitHub (2015b), GitHub: Citable Code. [Available from https://guides.github.com/activities/citable-code/, Last accessed 3 August 2015.]

Geoscientific Model Development (2013), Editorial: The publication of geoscientific model developments v1.0." GMD Executive Editors: J. Annan, J. Hargreaves, D. Lunt (Chief Editor), A. Ridgwell, I. Rutt and R. Sander, *Geosci. Model Dev.*, *6*, 1233–1242, doi:10.5194/gmd-6-1233-2013.

Goodman, A., et al. (2014), Ten simple rules for the care and feeding of scientific data, *PLoS Comput. Biol.*, *10*(4), e1003542, doi:10.1371/journal.pcbi.1003542.

Guo, P. J. (2012), CDE: A tool for creating portable experimental software packages, Computing in Science and Engineering: Special Issue on Software for Reproducible Computational Science, Jul/Aug 2012.

Hanson, B. (2014), AGU 's data policy: History and context, *Eos Trans. AGU*, *95*(37), 337, doi:10.1002/2014EO370008.

Harley, D. (2013), Scholarly communication: Cultural contexts, evolving models, *Science*, *342*, 80–82, doi:10.1126/science.1243622.

Hatton, L. (1997), The T experiments: Errors in scientific software, *Comput. Sci. Eng.*, *4*(2), 27–38.

Hatton, L., and A. Roberts (1994), How accurate is scientific software?, *IEEE Trans. Softw. Eng.*, *20*(10), 785–797.

Hatton, L., A. Wright, S. Smith, G. Parkes, and P. Bennett (1988), The seismic kernel system—A large scale exercise in Fortran 77 portability, *Softw.Pract. Exp.*, *18*(4), 301–329.

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

Hill, C., C. DeLuca, V. Balaji, M. Suarez, and A. da Silva (2004), Architecture of the earth system modeling framework, *Comput. Sci. Eng.*, *6*(1), 18–28, doi:10.1109/MCISE.2004.1255817.

Holdren, J. P. (2013), Increasing access to the results of federally funded scientific research memorandum for the heads of executive departments and agencies, Office of Science and Technology Policy, The White House. [Available from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.]

Hothorn, T., and F. Leisch (2011), Case studies in reproducibility, *Brief. Bioinform.*, *12*(3), 288–300. [Available from http://bib.oxfordjournals.org/content/12/3/288.]

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

Ioannidis, J. P. A. (2005), Why most published research findings are false, *PLoS Med.*, *2*(8), e124, doi:10.1371/journal.pmed.0020124.

Ioannidis, J. P. A. (2014), How to make more published research true, *PLOS Med.*, *11*(10), e1001747, doi:10.1371/journal.pmed.1001747.

Ioannidis, J. P. et al. (2009), Repeatability of published microarray gene expression analyses, *Nat. Genet.*, *41*(2), 149–155. [Available at http://www.nature.com/ng/journal/v41/n2/full/ng.295.html.]

International Organization for Standardization (ISO) (2005), ISO 19110:2005 Geographic information—Methodology for feature cataloguing. Available at http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39965.]

International Organization for Standardization (ISO) (2007), ISO 19139:2007 geographic information metadata XML schema. [Available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557.]

Jasny, B., G. Chin, L. Chong, and S. Vignieri (2011), Again, and again, and again, introduction to the special issue on data replication and reproducibility, *Science*, *334*, 1225–1225.

Jenkins (2015), The Jenkins extensible open source continuous integration server. [Available at https://jenkins-ci.org/, Last accessed 3 August 2015.]

Joppa, L. N., G. McInerny, R. Harper, L. Salido, K. Takeda, K. O'Hara, D. Gavaghan, and S. Emmott (2013), Troubling trends in scientific software use, *Comput. Sci.*, *340*(6134), 814–815.

Journal of Open Research Software (2015). [Available athttp://openresearchsoftware.metajnl.com/.]

Kattge, J., S. Díaz, and C. Wirth (2014), Of carrots and sticks, *Nat. Geosci.*, *7*, 778–779, doi:10.1038/ngeo2280.

Kenett, R. S., and G. Shmueli (2015), Clarifying the terminology that describes scientific reproducibility, *Nat. Methods*, *12*, 699, doi:10.1038/nmeth.3489.

Khatib, F., et al. (2011), Crystal structure of a monomeric retroviral protease solved by protein folding game players, *Nat. Struct. Mol. Biol.*, *18*(10), 1175–1177.

Klein, M., H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, and R. Tobin (2014), Scholarly context not found: One in five articles suffers from reference rot, *PLoS One*, *9*(12), e115253. doi:10.1371/journal.pone.0115253.

Koop, D., et al. (2011), A provenance-based infrastructure to support the life cycle of executable papers, *Proceedings of the International Conference on Computational Science, ICCS 2011*, doi:10.1016/j.procs.2011.04.068.

Leisch, F. (2002), Sweave: Dynamic generation of statistical reports using literate data analysis, Proceedings of Computational Statistics, 2002. Preprint available from http://www.statistik.lmu.de/~leisch/Sweave/Sweave-compstat2002.pdf.

LeVeque, R. J., I. M. Mitchell, and V. Stodden (2009), Reproducible research for scientific computing: Tools and strategies for changing the culture, *Comput. Sci. Eng*. *14*(4), 13, doi:10.1109/MCSE.2012.38

Lintott, C., et al. (2010), Galaxy Zoo 1: Data release of morphological classifications for nearly 900,000 galaxies, *Mon. Not. R. Astron. Soc.*, *410*(1), 166–178.

Ludaescher, B., I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao (2006), Scientific workflow management and the Kepler system, *Concurr. Comput.: Pract. Exp.*, *18*, 1039–1065.

Macleod, M. R., S. Michie, I. Roberts, U. Dirnagl, I. Chalmers, J. P. A. Ioannidis, R. Al-Shahi Salman, A.-W. Chan, and P. Glasziou (2014), Biomedical research: Increasing value, reducing waste, *Lancet*, *383*(9912), 101–104, doi:10.1016/S0140-6736(13)62329-6.

Manolescu, I., L. Afanasiev, A. Arion, J. Dittrich, S. Manegold, N. Polyzotis, K. Schnaitter, P. Senellart, S. Zoupanos, and D. Shasha (2008), The repeatability experiment of SIGMOD 2008, *ACM SIGMOD Rec.*, *37*(1). [Available at http://portal.acm.org/citation.cfm?id=1374780.1374791&coll=&dl=&idx=J689∂=newsletter&WantType=Newsletters&title=ACM%20SIGMOD%20Record.]

MATLAB (2015), MATLAB: The language of technical computing. MathWorks. [Available at http://www.mathworks.com/products/matlab/, Last accessed 3 August 2015.]

McCaffrey, R. E. (2005), Using citizen science in urban bird studies, *Urban Habitats*, *3*(1), 70–86.

Mesirov, J. P. (2010), Accessible reproducible research, *Science*, *327*, 415. [Available from http://www.sciencemag.org/cgi/rapidpdf/327/5964/415?ijkey=WzYHd6g6IBNeQ&keytype=ref&siteid=sci.]

Michener, W. K. (2015), Ecological data sharing, *Ecol. Inform.*, *29*(33-44), doi:10.1016/j.ecoinf.2015.06.010.

Missier, P., S. S. Sahoo, J. Zhao, Goble, C., and Sheth, A. (2010), Janus: From workflows to semantic provenance and linked open data. Provenance and Annotation of Data and Processes Third International Provenance and Annotation Workshop IPAW 2010 Troy NY USA June 1516 2010 Revised Selected Papers 6378, 129-141. [Available at http://www.mygrid.org.uk/files/presentations/SP-IPAW10.pdf.]

Moine, M.-P., et al. (2014), Development and exploitation of a controlled vocabulary in support of climate modelling, *Geosci. Model Dev.*, *7*, 479–493, doi:10.5194/gmd-7-479-2014.

Mooney, H., and M. P. Newton (2012), The anatomy of a data citation: Discovery, reuse, and credit, *J. Librar. Scholarly Commun.*, *1*(1), eP1035, doi:10.7710/2162-3309.1035.

Moreau, L., et al. (2008), Special issue: The first Provenance Challenge, *Concurrency Computat.: Pract. Exper.*, *20*, 409–418, doi:10.1002/cpe.1233.

Moreau, L., et al. (2011), The open provenance model core specification (v1.1), *Future Gener. Comput. Syst.*, *27*(6), 743–756, Preprint available from http://www.bibbase.org/cache/www.isi.edu__7Egil_publications.bib/moreau-etal-fgcs11.html

Moreau, L., et al. (2013), PROV-DM: The PROV data model, World Wide Web Consortium (W3C) Recommendation, April 2013. [Available at http://www.w3.org/TR/prov-dm/.]

Morozov, I., B. Reilkoff, and G. Chubak (2006), A generalized web service model for geophysical data processing and modeling, *Comput. Geosci.*, *32*(9), 1403, doi:10.1016/j.cageo.2005.12.010.

National Aeronautics and Space Administration (2015). [Available at http://science.nasa.gov/media/medialibrary/2014/12/05/NASA_Plan_for_increasing_access_to_results_of_federally_funded_research.pdf.]

National Institutes of Health (2015), Principles and guidelines for reporting preclinical research. [Available at http://www.nih.gov/about/reporting-preclinical-research.htm, Last accessed 3 August 2015.]

National Science Foundation (2015), NSF's Public Access Plan: Today's Data, Tomorrow's Discoveries — Increasing Access to the Results of Research Funded by the National Science Foundation, Arlington, Va., 18 March. [Available from https://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf.]

Nature (2012a), Error prone: Biologists must realize the pitfalls of work on massive amounts of data, *Nature*, *487*(7408), 406.

Nature (2012b), Must try harder, *Nature*, *483*, 509.

Nature (2013), Reproducing our irreproducibility, *Nature*, *496*, 398.

Nature (2014a), Journals unite for reproducibility, *Nature*, *515*, 7, doi:10.1038/515007a.

Nature (2014b), Code share, *Nature*, *514*, 536, doi:10.1038/514536a.

Nature (2015), Nature.com ontologies, *Nature*. [Available at http://www.nature.com/ontologies/.]

Nature (2016), Recommended data repositories. [Available at http://www.nature.com/sdata/data-policies/repositories, Last Accessed 24 May 2016.]

Nature Geoscience (2015), Towards transparency, *Nat. Geosci.*, *7*, 777, doi:10.1038/ngeo2294.

Nature Metrics (2010), Metrics survey results. [Available at http://www.nature.com/nature/newspdf/metrics_survey.pdf.]

Nekutrenko, A., and J. Taylor (2012), Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility, *Nat. Rev. Genet.*, *13*, 667–672, September 2012. doi:10.1038/nrg3305.

Nielsen, M. (2011), *Reinventing Discovery*, Princeton Univ. Press, Oxfordshire, U. K.

National Oceanic and Atmospheric Administration (2015). [Available at http://docs.lib.noaa.gov/noaa_documents/NOAA_Research_Council/NOAA_PARR_Plan_v5.04.pdf.]

Nowakowskia, P., E. Ciepielaa, D. Harężlaka, J. Kocota, M. Kasztelnika, T. Bartyńskia, J. Meiznera, G. Dyka, and M. Malawskib (2011), The collage authoring environment, *Procedia Comput. Sci.*, *4*, 608–617.

Open Archival Information System (2012), Reference model for an Open Archival Information System (OAIS), management council of the Consultative Committee For Space Data Systems (CCSDS), June 2012. [Available at http://public.ccsds.org/publications/archive/650x0m2.pdf, Last accessed 24 May 2016.]

Oinn, T., et al. (2006), Taverna: Lessons in creating a workflow environment for the life sciences, *Concurr. Comput.: Pract. Exp.*, *18*(10), 1067–1100.

Open Source Initiative (OSI) (2015), Available from http://opensource.org/licenses. Last accessed 3 August 2015.

Open Researcher and Contributor ID (2015), The community surface dynamics modeling system (ORCID). [Available at http://www.orcid.org. Last accessed 3 August 2015.]

Pampel, H., P. Vierkant, F. Scholze, R. Bertelmann, M. Kindling, J. Klump, H.-J. Goebelbecker, J. Gundlach, P. Schirmbacher, and U. Dierolf (2013), Making research data repositories visible: The re3data.org registry, *PLoS One*, doi:10.1371/journal.pone.0078080.

Pangaea (2015), The Pangaea Data Publisher for Earth and Environmental Science. [Available at www.pangaea.de, Last accessed 3 August 2015.]

Parsons, M., and P. Fox (2013), Is data publication the right metaphor? *Data Sci. J.*, *12*, WDS32–WDS46, doi:10.2481/dsj.WDS-042.

Pearson, W. R., and D. J. Lipman (1988), Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci.*, *85*(8), 2444–2448.

Pebesma, E., D. Nüst, and R. Bivand (2012), The R software environment in reproducible geoscientific research, *Eos Trans. AGU*, *93*(16), 163, doi:10.1029/2012EO160003.

Peckham, S. D. (2014), The CSDMS standard names: Cross-domain naming conventions for describing process models, data sets and their associated variables, Proceedings of the Seventh International Congress on Environmental Modeling and Software, June 2014.

Peckham, S. D. (2015), Community surface dynamics modeling system basic model interface. [Available at http://csdms.colorado.edu/wiki/BMI_Description, accessed on June 30, 2015.]

Peckham, S. D., E. W. H. Hutton, and B. Norris (2013), A component-based approach to integrated modeling in the geosciences: The design of CSDMS, *Comput. Geosci.*, *53*, doi:10.1016/j.cageo.2012.04.002.

Peng, R. D. (2009), Reproducible research and biostatistics, *Biostatistics*, *10*(3), 405–408, doi:10.1093/biostatistics/kxp014.

Piwowar, H. A., and W. W. Chapman (2009), Public sharing of research datasets: A pilot study of associations, *J. Informetrics*, *4*(2), 148–156, doi:10.1016/j.joi.2009.11.010.

Plotly (2015). [Available from https://plot.ly/, Last accessed 3 August 2015.]

Pope, A., W. G. Rees, A. J. Fox, and A. Fleming (2014), Open access data in polar and cryospheric remote sensing, *Remote Sens.*, *6*, 6183–6220, doi:10.3390/rs6076183.

Priem, D., J. Taraborelli, P. Groth, and C. Neylon (2010), Altmetrics: A manifesto, 26 October 2010. [Available at http://altmetrics.org/manifesto, Last accessed 3 August 2015.]

Prinz, F., T. Schlange, and K. Asadullah (2011), Believe it or not: How much can we rely on published data on potential drug targets?, *Nat. Rev. Drug Discovery*, *10*, 712, doi:10.1038/nrd3439-c1.

Research Data Alliance (2015), Outcomes of the Research Data Alliance (RDA). [Available at https://rd-alliance.org/outcomes, Last accessed July 30, 2015.]

Re3data (2015), The registry of research data repositories (re3data), Available from http://www.re3data.org. Last accessed 3 August 2015.

ReadCube (2015). [Available from https://www.readcube.com/, Last accessed 3 August 2015.]

Reichman, O. J., M. B. Jones, and M. P. Schildhauer (2011), Challenges and opportunities of open data in ecology, *Science*, *331*(6018), 703–705, doi:10.1126/science.1197962.

Rocca, R. A., G. Magoon, D. F. Reynolds, T. Krahn, and V. O. Tilroe (2012), Discovery of Western European R1b1a2 Y chromosome variants in 1000 genomes project data: An online community approach, *PLoS One*, *7*(7), e41634.

Roston, M. (2015), Retracted scientific studies: A growing list, The New York Times.

Royal Society (2012), Final report: Science as an open enterprise. [Available at https://royalsociety.org/policy/projects/science-public-enterprise/Report/.]

Russell, J. F. (2013), If a job is worth doing, it is worth doing twice, *Nature*, *496*, 7, doi:10.1038/496007a.

Ryan, M. J. (2011), Replication in field biology: The case of the frog-eating bat, *Science*, *334*(6060), 1229–1230.

Santer, B. D., T. M. L. Wigley, and K. E. Taylor (2011), The reproducibility of observational estimates of surface and atmospheric temperature change, *Science*, *334*(6060), 1232–1233.

Savage, C. J., and A. J. Vickers (2009), Empirical study of data sharing by authors publishing in PLoS journals, *PLoS One*, *4*(9), e7078.

Savage, N. (2012), Gaining wisdom from crowds, *Commun. ACM*, *55*(3), 13–15.

Schooler, J. W. (2014), Metascience could rescue the 'replication crisis', *Nature*, *515*, 9, doi:10.1038/515009a.

Schwab, M., N. Karrenbach, and J. Claerbout (2000), Making scientific computations reproducible, *Comput. Sci. Eng.*, *2*(6), 61–67. [Available at http://sep.stanford.edu/lib/exe/fetch.php?id=sep%3Aresearch%3Areproducible&cache=cache&media=sep:research:reproducible:cip.pdf.]

Science (2014), Journals unite for reproducibility, *Science*, *346*(6210), 679, doi:10.1126/science.aaa1724.

Shen, H. (2014), Interactive notebooks: Sharing the code, *Nature*, *515*, 151–152, doi:10.1038/515151a.

SoftwareX (2015), SoftwareX journal. [Available at http://www.journals.elsevier.com/softwarex/.]

Soranno, P. A., K. S. Cheruvelil, K. C. Elliott, and G. M. Montgomery (2014), It's good to share: Why environmental scientists' ethics are out of date, *BioScience*, doi:10.1093/biosci/biu169.

Source Code for Biology and Medicine (2015), http://www.scfbm.org/

SourceForge (2015), Available from http://sourceforge.net/. Last accessed 3 August 2015.

Spies, J., et al. (2012), The reproducibility of psychological science, Report of the Open Science Collaboration. [Available at openscienceframework.org/reproducibility/.]

Starr, J., et al. (2015), Achieving human and machine accessibility of cited data in scholarly publications, *PeerJ Comput. Sci.*, *1*, e1, doi:10.7717/peerj-cs.1.

Stodden, V. (2009), The legal framework for reproducible research in the sciences: Licensing and copyright, *IEEE Comput. Sci. Eng.*, *11*(1), 35–40.

Stodden, V., P. Guo, and Z. Ma (2013), Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PLoS One*, *8*, e67111. doi:10.1371/journal.pone.0067111.

Taylor, I., E. Deelman, D. Gannon, and M. Shield (2006), *Workflows for e-Science*, Springer.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of Cmip5 and the experiment design, *Bull. Am. Meteorol. Soc.*, doi:10.1175/Bams-D-11-00094.1.

Tenopir, C., S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame (2011), Data sharing by scientists: Practices and perceptions, *PLoS One*, doi:10.1371/journal.pone.0021101.

The Antarctic Glaciological Data Center (2015). [Available at https://nsidc.org/agdc, Last accessed 3 August 2015.]

The Community Surface Dynamics Modeling System (2015). [Available at http://csdms.colorado.edu, Last accessed 3 August 2015.]

The Computational Infrastructure for Geodynamics (CIG) (2015). [Available at https://geodynamics.org/cig/, Last accessed 3 August 2015.]

The Hierarchical Data Format (HDF) (2015). [Available at https://www.hdfgroup.org/, Last accessed 3 August 2015.]

The Interdisciplinary Data Alliance (2015). [Available at http://www.iedadata.org, Last accessed 3 August 2015.]

The National Centers for Environmental Information (2015). [Available at http://www.ncei.noaa.gov, Last accessed 3 August 2015.]

The Network Common Data Format (NetCDF) (2015). [Available at http://www.unidata.ucar.edu/software/netcdf/, Last accessed 3 August 2015.]

Tonella, P., and A. Potrich (2005), *Reverse Engineering of Object Oriented Code, Monographs in Computer Science*, pp. 208, Springer Science + Business Media, Inc., Boston.

Travis-CI (2015), The Travis continuous integration service. [Available at https://github.com/travis-ci, Last accessed 3 August 2015.]

Uhlir, P. F. (2012), For Attribution: Developing Data Attribution and Citation Practices and Standards, Rapporteur; Board on Research Data and Information; Policy and Global Affairs; National Research Council. Report of CODATA Data Citation Workshop. National Academies Press, 2012. [Available at http://www.nap.edu/catalog/13564/for-attribution-developing-data-attribution-and-citation-practices-and-standards.]

US Geological Survey (USGS) (2015). Available at http://www.usgs.gov/datamanagement/policyreferences.php, Last accessed 3 August 2015.]

Van Gorp, P., and S. Mazanek (2011), SHARE: A web portal for creating and sharing executable research papers, *Procedia Comput. Sci.*, *4*, 589–597, doi:10.1016/j.procs.2011.04.062.

Van Noorden, R. (2013), Data-sharing: Everything on display, *Nature*, *500*, 243–245, doi:10.1038/nj7461-243a.

Van Noorden, R. (2015), Sluggish data sharing hampers reproducibility effort, *Nature*, doi:10.1038/nature.2015.17694.

Vandewalle, P., et al. (2009), What, why and how of reproducible research in signal processing, IEEE Signal Processing, May 2009.

Vasilevsky, N. A., M. H. Brush, H. Paddock, L. Ponting, S. J. Tripathy, G. M. LaRocca, and M. A. Haendel (2013), On the reproducibility of science: Unique identification of research resources in the biomedical literature, *PeerJ*, *1*, e148, doi:10.7717/peerj.148.

VHub (2015), VHub: The collaborative volcano and risk mitigation hub. [Available at https://vhub.org, Last accessed 3 August 2015.]

Vines, T. H., A. Y. K. Albert, R. L. Andrew, F. Débarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, and D. J. Rennison (2014), The availability of research data declines rapidly with article age, *Curr. Biol.*, *24*, 94–97.

Vision, T. J. (2010), Open data and the social contract of scientific publishing, *BioScience*, *60*(5), 330–331, doi:10.1525/bio.2010.60.5.2.

W3C (2016), Permanent identifiers for the Web, World Wide Web Consortium, 2016. [Available at https://w3id.org/, Last accessed 24 May 2016.

WaterML (2015), WaterML 2.0. Open geospatial consortium. [Available at http://www.opengeospatial.org/standards/waterml/, Last accessed 3 August 2015.]

Wilson, M. L., W. Mackay, E. H. Chi, M. S. Bernstein, and J. Nichols (2012), RepliCHI SIG—From a panel to a new submission venue for replication, ACM SIGCHI.

Woelfle, M., P. Olliaro, and M. H. Todd (2011), Open science is a research accelerator, *Nat. Chem.*, *3*, 745–748.

Wolfram (2015), Computable Document Format (CDF). [Available at http://www.wolfram.com/cdf, Last accessed 3 August 2015.]

Yong, E. (2012), Replication studies: Bad copy, *Nature*, *485*, 298–300. [Available at doi:10.1038/485298a.]

Zenodo (2015). [Available from http://www.zenodo.org, Last accessed 3 August 2015.]

Zudilova-Seinstra, E. (2013), Designing the Article of the Future: Elsevier's user centered design specialists show how they worked with users to transform the format of online articles, Elsevier, accessed on June 29, 2015. [Available athttp://www.elsevier.com/connect/designing-the-article-of-the-future.]