Accepted for publication in the Earth and Space Science Journal, 2016.

Towards the Geoscience Paper of the Future: Best Practices for Documenting and Sharing Research from Data to Software to Provenance

Yolanda Gil[1] (orcid.org/0000-0001-8465-8341)
Information Sciences Institute and Department of Computer Science, University of Southern California

Cédric H. David (orcid.org/0000-0002-0924-5907)
Jet Propulsion Laboratory, California Institute of Technology

Ibrahim Demir (orcid.org/0000-0002-0461-1242)
IIHR Hydroscience & Engineering Institute, University of Iowa

Bakinam T. Essawy (orcid.org/0000-0003-2295-7981)
Department of Civil and Environmental Engineering, University of Virginia

Robinson W. Fulweiler (orcid.org/0000-0003-0871-4246)
Department of Earth and Environment, Department of Biology, Boston University

Jonathan L. Goodall (orcid.org/0000-0002-1112-4522)
Department of Civil and Environmental Engineering, University of Virginia

Leif Karlstrom (orcid.org/0000-0002-2197-2349)
Department of Geological Sciences, University of Oregon, Eugene, Oregon, 97403

Huikyo Lee (orcid.org/0000-0003-3754-3204)
Jet Propulsion Laboratory, California Institute of Technology

Heath J. Mills (orcid.org/0000-0002-2882-9238)
Division of Natural Sciences, University of Houston Clear Lake

Ji-Hyun Oh (orcid.org/0000-0001-8989-7691)
Computer Science Department of the University of Southern California & Jet Propulsion Laboratory, California Institute of Technology

Suzanne A Pierce (orcid.org/0000-0002-3050-1987)
Texas Advanced Computing Center and Jackson School of Geosciences, University of Texas at Austin

Allen Pope (orcid.org/0000-0001-9699-7500)
National Snow and Ice Data Center, University of Colorado Boulder & Polar Science Center, Applied Physics Lab, University of Washington

Mimi W. Tzeng (orcid.org/0000-0001-9396-3217)
Data Management Center, Dauphin Island Sea Lab, Alabama

Sandra R. Villamizar (orcid.org/000-0003-4961-0187)
Sierra Nevada Research Institute, University of California, Merced

Xuan Yu (orcid.org/0000-0002-9055-7923)
Department of Geological Sciences, University of Delaware, Newark, Delaware, 19716

[1] **Corresponding author:** Yolanda Gil, Information Sciences Institute, University of Southern California, (gil@isi.edu)

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 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.

August 18, 2016

Key Points

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

Keywords: Geoscience paper of the future; Reproducibility; Data sharing; Software reuse; Provenance; Workflow; Data Science

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, 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 data, software, and provenance management 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 datasets, 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 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 workflow 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 datasets 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 and provenance between the data, software, and results of the research.

We envision that reproducibility (i.e., being able to re-create a study's results) and 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 datasets, 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 a journal with clear data sharing policies only one out of ten authors made their datasets available upon request [Savage and Vickers 2009], which argues for data being published when a paper is published.

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 [AGU 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 [GMD 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 dataset that then becomes the recommended citation for the data themselves, although this still remains problematic for many datasets used in a primary scientific publication (e.g., [Pope et al 2014]). Journals devoted solely to describing software are beginning to emerge [JORS 2015; SoftwareX 2015; SCFBM 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 provenance records [Taylor et al 2007; 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 PDF. For example, ReadCube allows readers can navigate the citations through cross-reference facilities across publishers [ReadCube 2015]. Other experimental efforts include the Executable Papers Challenge (e.g., [Van Gorpa and Mazanekb, 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 datapoints).

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 counter arguments (e.g., [Barnes 2010; Costello et al 2013; Nature Geoscience 2015; Michener 2015]).

Publishing and sharing data and software leads 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 1998; Hatton 1997]. This ambiguity is ingrained in natural language descriptions and consequently is unavoidable, so it is best to publish the software and 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 5 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 co-authors 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 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 US 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 they are publicly accessible [Holdren 2013]. US government funding agencies are responding to this mandate by developing plans to require research products to be openly published. The US National Science Foundation (NSF) released a Public Access Plan in March 2015 [NSF 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 US agencies that fund geosciences research, such as the National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and the US Geological Survey (USGS), have issued similar planning documents [NASA 2015; NOAA 2015; USGS 2015b].

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 datasets 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 datasets. 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 provenance 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 3 degrees of reproducibility: easily reproducible (ER) if it can be easily re-run within 10 mins, conditionally reproducible (CR) if it requires proprietary data, licensed software, or more than 10 mins to run, and non-reproducible (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 2012]). Repositories of shared 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 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 [NIH 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, Studdon et al, 2013]. Open science must be a community effort involving scientists, publishers, and funders [Kattge et al 2014]. 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, we need to realize that researchers seem to altruistically 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, sub-disciplines 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:

- **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.
- **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.

- **Document the digital provenance of results** by explicitly describing the series of computations and their outcome in a workflow sketch, a formal workflow, or a 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 provenance of the results so that they can be re-run 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. Common challenges include the reluctance of co-authors 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 the citation of provenance and the publication of provenance in a public repository are both optional. Ideally, both would be done by GPF authors, but we recognize the lack of shared provenance and workflow repositories in geosciences and therefore are recommended here although considered optional.

4. Suggested Best Practices and Current Challenges

This section describes recommended best practices on how to document data, software, and provenance, and to uniquely identify and cite these digital objects.

Table 1 provides a proposed author checklist consisting of twenty 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., [RDA 2015; CODATA 2013; DataCite 2015; FORCE11 2014; ESIP 2012; Starr et al 2015; Uhlir et al 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.

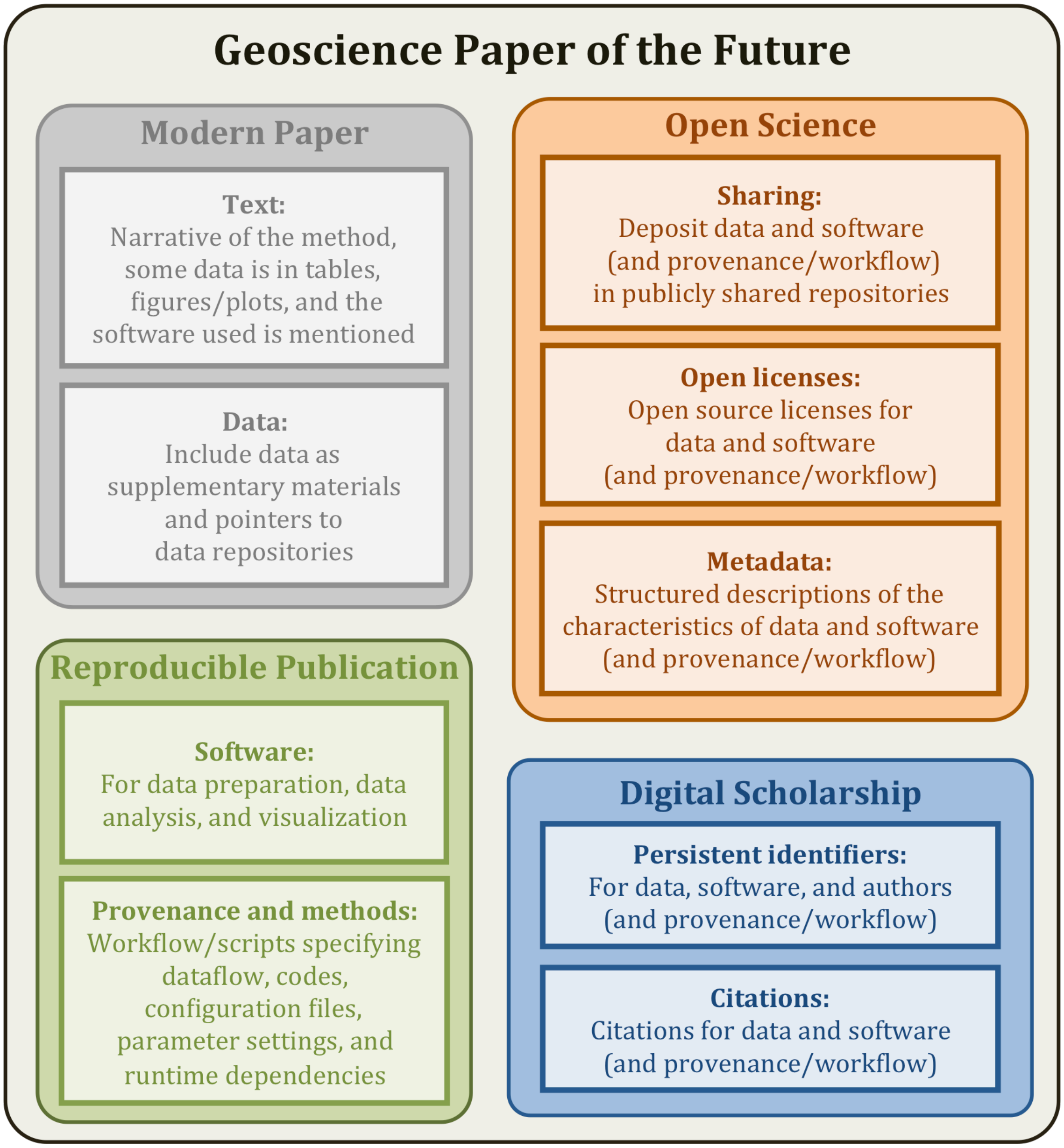

Figure 1. A Geoscience Paper of the Future (GPF) includes data, software and provenance as expected in reproducible publications, but also includes desirable features in open science and digital scholarship: 1) sharing of data, software, and other research products in public repositories, 2) use of open 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 digital 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.

Table 1. A proposed checklist for GPF authors, with twenty 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**	Datasets should be published in a publicly accessible location with a permanent unique identifier
		D2	Datasets should have a license
		D3	Datasets should be cited in the paper
Data Documentation	Initial data, significant intermediate results, and final results	**D4**	Datasets should have general-purpose metadata specified
		D5	Dataset characteristics should be explained in detail
		D6	Dataset origins and availability of related datasets 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	Dataflow 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

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.

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 datasets 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 two 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 [IEDA 2015], NCEI [NCEI 2015], Pangaea [Pangaea 2015]), and finally to the highly specific, managed, and curated (e.g., AGDC [AGDC 2015], NASA's DAACs [DAACs 2015], and the USGS Science Data Catalog [USGS 2015b]). Curation takes time, since it requires adding metadata that is consistent with other entries in the repository. 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. Some repositories are popular with specific disciplines or communities, which increases the chances of data reuse by others. 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 non-durable), 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 [CC 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 reuse by others.

Citation (D3): Data can be cited within text much like an article would be cited, or it can be cited in a special resources section or in the acknowledgements section. Some journals have specific guidelines for data citation. 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 dataset, 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 web site) that can be changed over time. The creator of a PURL must update the link if the URL changes. The PURL can be cited in the paper, and the authors should ensure that the redirection address is updated if anything changes in their local system. A PURL can be obtained through services such as the Online Computer Library Center (OCLC)'s PURL service [PURL 2015]. 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 (/).

4.1.2 Additional Requirements and Issues

Taking data from public repositories: While many researchers collect or create their own datasets, many researchers take data from publicly available repositories. The NASA Global Change Master Directory is a recommended tool to discover data sets in geosciences [GCMD 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 datasets 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 challenges of using data from colleagues are having access to metadata for the case of raw data and having access to provenance for the case of processed data sets.

Publishing intermediate data: Intermediate data should be published when data preparation steps are hard to re-execute 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 dataset. 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 datasets: In many disciplines, the availability of datasets with high spatial/temporal resolutions creates a challenge. Although data storage and transport costs are getting cheaper, sharing and transferring large datasets is still a challenge. Therefore, it is essential to prepare large datasets 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 [Dash 2015]) or that continue to increase the sizes of the datasets allowed (e.g., [Zenodo 2015; Dryad 2015; GitHub 2015a]). One possibility is to publish a sample of the data used, or to document datasets with extensive metadata about the characteristics.

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.

4.2 Documenting Data by Specifying Metadata

Once data are 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 Dataset Characteristics

General-purpose metadata (D4): This is general information about the who/what/when/where/why/how of the dataset. 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 dataset. Most repositories request this kind of metadata.

Dataset characteristics (D5): Scientifically relevant characteristics of the dataset 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 datasets (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 [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 [CD 2011], are increasingly gaining acceptance in their communities.

Data sources and related datasets (D6): Other researchers may be interested not only 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 dataset is extracted from a larger database, or is one of several datasets 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 datasets, 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) [DCMI 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.

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 2015], SourceForge [SourceForge 2015], and Bitbucket [Bitbucket 2015]. Code repositories offer version control systems to facilitate code evolution and collaboration among developers as well as users.

License (S2): Contrary to a copyright automatically applied to software when it is created to grant *the creator* exclusive rights as an intellectual property, an open source license is a mechanism for defining the level of control required by creators over their source code. 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 [OSI 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) [CSDMS 2015; Peckham et al 2013], the Earth System Modeling Framework (ESMF) [ESMF 2015; Hill et al 2004], the Computational Infrastructure for Geodynamics (CIG) [CIG 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], 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 run-time libraries or operating systems, which may not be available.

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 run-time 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].

4.5 Documenting the Provenance of Results

In a computational sense, provenance is an explicit documentation of the data used and the processing performed to reach a scientific result. 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. A provenance record needs to be provided for every figure, table, or new dataset shown in a paper. This record should describe in detail what was actually done, that is, not just what software was used but how it was configured and what specific parameter values were used in the runs that led to the results shown in the paper.

4.5.1 Documenting Provenance: Derivation, Execution, Versions, and Parameters

Derivation traces (P1): A derivation trace can show conceptually how the data were used by the software, and what intermediate and final results were obtained. Traditionally, this is described in the “Methods” section of a paper, usually in the form of text. This format is limited by its inability to fully convey the complexity inherent in computational research, and should therefore be complemented these derivation traces. These can be very effective to convey important details to the reader, and are the first step towards a more complete provenance. Derivation traces can be shown as a graphical sketch, or as a table. They can be sketched using readily available graphics tools.

Execution traces (P2): In addition to a sketch of the derivation trace, fully detailed execution traces can be provided to document the execution details. These execution traces may include statements printed from the code that indicate what is happening. The execution traces must specify unique identifiers for data and software, assigned as described in the sections above. A provenance standard, such as W3C PROV [Gil and Miles 2013; Moreau et al 2013], may be followed to represent execution traces and enable their analysis and reuse. Provenance traces can also be obtained from workflow systems [Taylor et al 2007; Gil et al 2007; Deelman et al 2012], such as Pegasus [Deelman et al 2005], Taverna [Oinn et al 2006], Vistrails [Callahan et al 2006], and Kepler [Ludaescher et al 2006]. The derivation traces and execution traces may be combined in cases where the two would be very similarly specified in a paper.

Versions (P3): The versions should be indicated for all the software used to obtain the results in the paper. Software often evolves and can have many releases over the years, particularly commercial software. Different versions may offer different functionality, some may disappear over time and some may be incorporated in a new release. The versions of the software should be an integral part of a provenance record.

Parameters (P4): A detailed provenance record shows all the data flow across software components, corresponding to the detailed command line invocations and parameter values used. 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 provenance: Ideally, the provenance records for a paper would be published in a public repository and cited in the article. This would put 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 scientific provenance records. It is possible to publish provenance in a data repository, but provenance records have unique structure that should be searchable and comparable. In the future, shared provenance repositories may emerge to enable provenance discovery, comparison, and reuse.

Data preparation steps: Data preparation aspects are often not mentioned in articles, but are crucial to documenting the 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 use of standards.

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 must be prescribed (or proscribed) 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 provenance records: In some cases, the provenance records may be very large and complex to describe. Some research involves running dozens of codes, and in those cases provenance can be documented at varying levels of detail. At the simplest level is a derivation trace that sketches the most important steps and the data flow among them. A more detailed execution trace can document all the steps followed. Other research may involve running experiments with hundreds of datasets. In those cases, the provenance of a few runs can be documented in detail, and the others described at a higher level.

Reproducibility versus inspectability: Reproducibility requires re-running the experiments in the article, but inspectability simply requires examining the 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.

4.6 Documenting the Methods

While provenance documents the specific executions that lead to the results presented in the paper, the methods refer to the underlying general strategies that can be applied to other data. 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): This documents the various steps in the method in terms of how software is composed together. This composition is sometimes a sequential pipeline, but it may consist of many interconnected and interdependent steps. The composition can be indicated as a simple flow diagram, or 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 2007].

Dataflow (M2): The dataflow between the steps indicates how initial data would be processed by software, what intermediate datasets 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.

4.7 Author Identification

Much like datasets 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) [ORCID 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 co-author scientific publications.

Data

1. Find a community data repository, or use a generic one:
   - figshare (http://figshare.com/)
   - Zenodo (http://zenodo.org/)
   - Dryad Digital Repository (http://datadryad.org/)
2. Create a public entry for the data (input, output, intermediate)
3. Specify the metadata including a license
   - Use CC-BY or another license from http://creativecommons.org/licenses/
4. Specify citations for the data, each with a dataset name, persistent identifier, author names, repository, and access date

Software

1. Publish code on lab web site, or use a community code repository:
   - Github (http://github.com/)
   - SourceForge (http://sourceforge.net/)
2. Create a public entry for the software (models, data pre-processing, plotting)
3. Specify the metadata including a license
   - Use a software registry to specify metadata, e.g., http://www.ontosoft.org/portal
   - Use ALv2 or another license from http://opensource.org/licenses
4. Specify citations for the software, with a name, persistent identifier, author names, repository, version, and access date

Provenance and Methods

1. Describe the method in text, referring to data and software with their unique identifiers
2. Develop a workflow sketch, draw a graph to show the dataflow, i.e., how data is processed by each software component, or use a lab notebook or a workflow system
   - http://ipython.org/notebook.html, http://www.wings-workflows.org, https://oodt.apache.org/
3. Show the provenance by summarizing the execution trace
4. Create a public entry for the workflow or provenance, in lab web site or in a data or workflow repository
5. Optionally, specify citations for the workflow and the provenance including a persistent identifier

Figure 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.

4.8 Summary: Preparing a GPF for Publication

Figure 2 provides a roadmap of the best practices and widely-used resources discussed in this section, aligned with the GPF Author Checklist shown in Table 1.

In summary, to prepare a GPF for publication, several key aspects of the research much 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 data should be (D1) published in an accessible location with a permanent unique identifier, (D2) datasets should be published with an accompanying license to delineate acceptable reuse and dissemination options, and (D3) datasets 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) dataset characteristics should be explained in detail, and (D6) dataset origins and availability of related datasets 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 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 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 datasets 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 reporting on the (M1) 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 each author is linked to the products for the research through the use of a permanent unique identifier (A1).

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 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 regards 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.

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 twenty 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.

7. 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 number 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 & 1336839. Mimi Tzeng was supported by a grant from NOAA through the Northern Gulf Institute (NGI). 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 and the anonymous reviewers for valuable comments on this manuscript.

Data accessibility statement: Because this article is a compilation of best practices and recommendations that appear elsewhere, no data was used or produced during the course of the work reported.

8. References

[ACADIS 2015] The ACADIS Gateway: An Arctic Data Repository. Available from http://www.aoncadis.org. Last accessed 3 August 2015.

[AGDC 2015] The Antarctic Glaciological Data Center. Available from https://nsidc.org/agdc. Last accessed 3 August 2015.

[AGU 2013] AGU Publications Data Policy. American Geophysical Union, December 2013. Available from http://publications.agu.org/author-resource-center/publication-policies/data-policy/

[Alias et al 2014] Alias, A., Balaji, V., Bentley, P., Devine, G., Callaghan, S. A., and Guilyardi, E. Geoscientific Model Development, 7, 479-493, doi:10.5194/gmd-7-479-2014, 2014.

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

[Angeli, E., Wagner, J., Lawrick, E., Moore, K., Anderson, M., Soderlund, L., and Brizee, A., 2010] General format. Retrieved from http://owl.english.purdue.edu/owl/resource/560/01/

[Baggerly and Coombes 2009] Baggerly, K. A. and Coombes, K. R. "Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology." Annals of Applied Statistics, 3(4), 2009. Available from http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoas/1267453942

[Baggerly and Coombes 2011] 'What Information Should be Required to Support Clinical Omics Publications?" Keith A. Baggerly and Kevin R. Coombes. Clinical Chemistry, 57:5, 688-690, 2011.

[Baker et al 2010] "Transparency and reproducibility in data analysis: the Prostate Cancer Prevention Trial." Stuart G. Baker, Amy K. Drake, Paul Pinsky, Howard L. Parnes, Barnett S. Kramer. Biostatistics, 11(3), 2010.

[Baker 2012] "Independent Labs to Verify High-Profile Papers." Monya Baker. Nature News, 14 August 2012. doi:10.1038/nature.2012.11176

[Ball and Duke 2012] "How to Cite Datasets and Link to Publications". DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: 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 2010] “Publish Your Computer Code: It’s Good Enough.” Nick Barnes. Nature, 467, pp 753, 2010. doi:10.1038/467753a.

[Bauer et al 2011] Bauer, Brigitte Surer, Matthias Troyer, Dean N Williams, Joel E Tohline, Juliana Freire, Cláudio T Silva. Procedia Computer Science, Volume 4, 2011, Pages 648–657, 2011.

[BCO-DMO 2015] The Data Collection of the Biological and Chemical Oceanography Data Management Office (BCO-DMO). Available from http://www.bco-dmo.org/data. Last accessed 3 August 2015.

[Begley and Ellis 2012] “Drug development: Raise standards for preclinical cancer research.” C. Glenn Begley and Lee M. Ellis. Nature 483, 531–533, 29 March 2012. doi:10.1038/483531a

[Bell et al 2009] “A HUPO test sample study reveals common problems in mass spectrometry–based proteomics.” Bell AW, Deutsch EW, Au CE, Kearney RE, Beavis R, Sechi S, Nilsson T, Bergeron JJ, and the Human Proteome Organization (HUPO) Test Sample Working Group. Nature Methods, 6(6), 2009. Available from http://www.nature.com/nmeth/journal/v6/n6/full/nmeth.1333.html

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

[Bonnet et al 2011] “Repeatability and workability evaluation of SIGMOD 2011.” Philippe Bonnet, Stefan Manegold, Matias Bjørling, Wei Cao, Javier Gonzalez, Joel Granados, Nancy Hall, Stratos Idreos, Milena Ivanova, Ryan Johnson, David Koop, Tim Kraska, René Müller, Dan Olteanu, Paolo Papotti, Christine Reilly, Dimitris Tsirogiannis, Cong Yu, Juliana Freire, Dennis Shasha:. SIGMOD Record 40(2): 45-48, 2011.

[Bourne et al 2010] Bourne, P. “What Do I Want from the Publisher of the Future?” PLoS Computational Biology, 2010. Available from http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000787

[Bourne et al 2011] “Improving Future Research Communication and e-Scholarship.” Phil E. Bourne, Tim Clark, Robert Dale, Anita de Waard, Ivan Herman, Eduard Hovy, and David Shotton (Eds). The FORCE 11 Manifesto, available from http://www.force11.org.

[Callahan et al 2006] “Managing the Evolution of Dataflows with VisTrails.” Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Claudio T. Silva and Huy T. Vo. Proceedings of IEEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow), 2006.

[Cardamone et al 2009] Cardamone, C., Schawinski, K., Sarzi, M., et al. "Galaxy Zoo Green Peas: Discovery of A Class of Compact Extremely Star-Forming Galaxies" 2009, MNRAS, 399, 1191.

[CC 2015] Creative Commons. Available from http://www.creativecommons.org. Last accessed 3 August 2015.

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

[CF 2011] NetCDF Climate and Forecast (CF) Metadata Conventions. 2011. Available from http://cfconventions.org/. Last accessed 3 August 2015.

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

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

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

[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 Science Journal, 2013. DOI: 10.2481/dsj.OSOM13-043

[CODATA 2015] International Council for Science: Committee on Data for Science and Technology (CODATA). Available from http://www.codata.org/. Last accessed 3 August 2015.

[Collins and Tabak 2014] "Policy: NIH Plans to Enhance Reproducibility." Nature 505(7485):612-3, 30 January 2014.

[Costello et al 2013] "Biodiversity data should be published, cited, and peer reviewed." Mark J. Costello, William K. Michener, Mark Gahegan, Zhi-Qiang Zhang, Philip E. Bourne. Trends in Ecology & Evolution 28, 454-461, 2013.

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

[DAACs 2015] The Earth Observing System Data and Information System (EOSDIS) Distributed Archive Centers (DAACs). Available from 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] DataCite. Available from https://www.datacite.org/. Last accessed 3 August 2015.

[DCMI 2012] Dublin Core Metadata Terms. 2012. Available from http://dublincore.org/documents/dcmi-terms/. Last accessed 3 August 2015.

[Deelman et al 2005] "Pegasus: a Framework for Mapping Complex Scientific Workflows onto Distributed Systems." Deelman, E., Singh, G., Su, M. H., Blythe, J., Gil, Y., Kesselman,C., Mehta, G., Vahi, K., Berriman, G. B., Good, J., Laity, A., Jacob, J. C. and Katz, D. S. (2005) Scientific Programming Journal, vol. 13, pp. 219-237.

[Deelman et al 2012] "EarthCube Report on a Workflows Roadmap for the Geosciences." Deelman, E.; Duffy, C.; Gil, Y.; Marru, S.; Pierce, M.; and Wiener, G. 2012. National Science Foundation, Arlington, VA, 2012.

[De Roure et al 2009] De Roure, D; Goble, C.; Stevens, R. "The design and realizations of the myExperiment Virtual Research Environment for social sharing of workflows". Future Generation Computer Systems, 25 (561-567), 2009

[Dellavalle et al 2003] "Going, Going, Gone: Lost Internet References." Robert P. Dellavalle, Eric J. Hester, Lauren F. Heilig, Amanda L. Drake, Jeff W. Kuntzman, Marla Graber, Lisa

[DeRisi et al 2013] "The What and Whys of DOIs." Susanne DeRisi, Rebecca Kennison, Nick Twyman. PLoS Biology 1(2): e57, 2013. DOI

[Diggle and Zeger 2009] "Reproducible research and Biostatistics." Peter J. Diggle and Scott L. Zeger. Biostatistics 10(3), 2009.

[Donoho 2002] "How to be a Highly Cited Author in the Mathematical Sciences" David L. Donoho. In-cites, 2002. Available from http://www.in-cites.com/scientists/DrDavidDonoho.html

[Donoho and Huo 2004] "BEAMLAB and Reproducible Research." David L. Donoho and Xiaoming Huo. International Journal of Wavelets, Multiresolution, and Information Processing. 02, 391, 2004. DOI: 10.1142/S0219691304000615

[Donoho et al 2009] "Reproducible Research in Computational Harmonic Analysis." David Donoho, Arian Maleki, Inam Rahman, Morteza Shahram, Victoria Stodden. Computing in Science and Engineering, January 2009.

[Donoho 2010] "An Invitation to Reproducible Computational Research." David L. Donoho. Biostatistics 11 (3): 385-388, 2010. doi:10.1093/biostatistics/kxq028

[Downs et al 2015] "Data Stewardship in the Earth Sciences." Robert R. Downs, Ruth Duerr, Denise J. Hills, and H. K. Ramapriyan. D-Lib Magazine, 21(7/8). doi:10.1045/july2015-downs

[Dryad 2015] Dryad. Available from http://www.datadryad.org. Last accessed 3 August 2015.

[Easterbrook 2014] "Open code for open science?" Steve M. Easterbrook. Nature Geoscience 7, 779–781, 2014. doi:10.1038/ngeo2283

[Ellison 2010] "Repeatability and transparency in ecological research." Aaron M. Ellison. Ecology 91:2536–2539, 2010. http://dx.doi.org/10.1890/09-0032.1

[eScience 2011] "Transforming Scholarly Communication." Report of 2011 Microsoft Research eScience Workshop. Available from http://msrworkshop.tumblr.com/. Last Accessed 31 July 2015.

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

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

[Falcon 2007] Falcon, S. “Caching code chunks in dynamic documents: The weaver package.” Computational Statistics, (24)2, 2007. Available from http://www.springerlink.com/content/55411257n1473414/

[Fang and Casadevall 2011] Fang, C.F., and Casadevall, A. “Retracted Science and the retracted index”. Infection and Immunity. 2011. doi:10.1128/IAI.05661-11

[Fegraus et al 2005] “Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation.” Eric H. Fegraus, Sandy Andelman, Matthew B. Jones, and Mark Schildhauer. Bulletin of the Ecological Society of America 86:158–168, 2005. 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 et al 2012] “Planet Hunters: the first two planet candidates identified by the public using the Kepler public archive data.” Fischer, Debra A.; Schwamb, Megan E.; Schawinski, Kevin; Lintott, Chris; Brewer, John; Giguere, Matt; Lynn, Stuart; Parrish, Michael; Sartori, Thibault; Simpson, Robert; Smith, Arfon; Spronck, Julien; Batalha, Natalie; Rowe, Jason; Jenkins, Jon; Bryson, Steve; Prsa, Andrej; Tenenbaum, Peter; Crepp, Justin; Morton, Tim; Howard, Andrew; Beleu, Michele; Kaplan, Zachary; Vannispen, Nick; Sharzer, Charlie; Defouw, Justin; Hajduk, Agnieszka; Neal, Joe P.; Nemec, Adam; Schuepbach, Nadine; Zimmermann, Valerij. MNRAS, Volume 419, Issue 4, pp. 2900-2911, 2012.

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

[Freire and Silva 2012] “Making Computations and Publications Reproducible with VisTrails.” Juliana Freire and Claudio Silva. Computing in Science and Engineering 14(4): 18-25, 2012.

[Garijo et al 2013] “Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome” Daniel Garijo, Sarah Kinnings, Li Xie, Lei Xie, Yinliang Zhang, Philip E. Bourne, and Yolanda Gil. PLOS ONE, 27 November 2013.

[Garijo et al 2014] “Towards Workflow Ecosystems Through Semantic and Standard Representations.” Daniel Garijo, Yolanda Gil, and Oscar Corcho. 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 and Donoho 2011] “A Universal Identifier for Computational Results”, Procedia Computer Science, Volume 4, doi:10.1016/j.procs.2011.04.067, 637-647.

[GCMD 2015] The NASA's Global Change Master Directory. Available from http://gcmd.nasa.gov/. Last accessed 3 August 2015.

[Gil 2014] "Intelligent Workflow Systems and Provenance-Aware Software." Yolanda Gil. International Environmental Modelling and Software Society (iEMSs) 7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA, Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.) Available from http://www.iemss.org/society/index.php/iemss-2014-proceedings

[Gil and Miles 2013] "A Primer for the PROV Provenance Model." Yolanda Gil, Simon Miles, Khalid Belhajjame, Helena Deus, Daniel Garijo, Graham Klyne, Paolo Missier, Stian Soiland-Reyes, and Stephan Zednik. Published as a W3C Working Group Note on 30 April 2013. Available from (http://www.w3.org/TR/prov-primer/.

[Gil and Pierce 2015] "Report of the 2015 National Science Foundation Workshop on Intelligent Systems for Geosciences." Yolanda Gil and Suzanne Pierce (Eds). To be published on November 2015. Available from http://www.is-geo.org. Last accessed 3 August 2015.

[Gil et al 2007] "Examining the Challenges of Scientific Workflows," Yolanda Gil, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and Jim Myers. IEEE Computer, vol. 40, no. 12, pp. 24-32, December, 2007. 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 et al 2015] "OntoSoft: Capturing Scientific Software Metadata." Yolanda Gil, Varun Ratnakar, and Daniel Garijo. 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.

[GMD 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. Geoscientific Model Development, 6, 1233–1242, 2013. doi:10.5194/gmd-6-1233-2013

[Goodman et al 2014] Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Stefano, R. D., Gil, Y., Groth, P., Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., and A. Slavkovic (2014), Ten simple rules for the care and feeding of scientific data, PLOS Computational Biology, 10, April 24, 2014, doi: 10.1371/journal.pcbi.1003542.

[Guo 2012] "CDE: A Tool For Creating Portable Experimental Software Packages." Philip J. Guo. Computing in Science and Engineering: Special Issue on Software for Reproducible Computational Science, Jul/Aug 2012.

[Hanson 2014] "AGU 's Data Policy: History and Context." Brooks Hanson. Eos, 95(37), 337. 16 September 2014.

[Harley 2013] “Scholarly Communication: Cultural Contexts, Evolving Models.” Diane Harley. Science 4 October 2013: 80-82. DOI:10.1126/science.1243622

[Hatton 1997] “The T Experiments: Errors in Scientific Software.” Les Hatton. Computational Science & Engineering, Volume 4 Issue 2, 1997.

[Hatton and Roberts 1994] “How Accurate is Scientific Software?” L. Hatton and A. Roberts. IEEE Transactions on Software Engineering, Vol 20, Issue 10, 785-797, 2014.

[Hatton et al 1988] “The Seismic Kernel System – A Large Scale Exercise in Fortran 77 Portability.” Les Hatton, Andy Wright, Stuart Smith, Gregg Parkes, Paddy Bennett. Software—Practice & Experience, Volume 18 Issue 4, April 1988.

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

[Hey and Payne 2015] “Open Science Decoded.” Tony Hey and Mike C. Payne. Nature Physics, Vol 11, May 2015.

[Hill et al 2004] “Architecture of the Earth System Modeling Framework.” Hill, C., C. DeLuca, V. Balaji, M. Suarez, and A. da Silva. Computing in Science and Engineering, Volume 6, Number 1, 2004. doi:10.1109/MCISE.2004.1255817.

[Hothorn and Leisch 2011] “Case Studies in Reproducibility.” Torsten Hothorn and Friedrich Leisch. Briefings in Bioinformatics, 12(3), 2011. Available from http://bib.oxfordjournals.org/content/12/3/288

[Hutson 2010] Hutson, S. “Data Handling Errors Spur Debate Over Clinical Trial,” Nature Medicine, 16(6), 2010. Available from http://www.nature.com/nm/journal/v16/n6/full/nm0610-618a.html

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

[Ince et al 2012] “The Case for Open Computer Programs.” Darrel C. Ince, Leslie Hatton, and John Graham-Cumming. Nature, Vol 482, 23 February 2012.

[Ioannidis 2005] “Why Most Published Research Findings are False.” PLoS Medicine, 2(8): e124, 2005. doi:10.1371/journal.pmed.0020124

[Ioannidis et al 2009] Ioannidis J.P., Allison D.B., Ball C.A., Coulibaly I, Cui X., Culhane A.C., Falchi M, Furlanello C., Game L., Jurman G., Mangion J., Mehta T., Nitzberg M., Page G.P., Petretto E., van Noort V. ”Repeatability of Published Microarray Gene Expression Analyses.” Nature Genetics, 41(2), 2009. Available from http://www.nature.com/ng/journal/v41/n2/full/ng.295.html

[Ioannidis 2014] “How to Make More Published Research True.” John P. A. Ioannidis. PLOS Medicine, Vol 11, No 10, October 2014. DOI:10.1371/journal.pmed.1001747

[ISO 2005] ISO 19110:2005 Geographic information -- Methodology for feature cataloguing. 2005. Available from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=39965.

[ISO 2007] ISO 19139:2007 Geographic Information Metadata XML Schema. 2007. Available from http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557.

[Jasny et al 2011] "Again, and Again, and Again." Barbara Jasny, Gilbert Chin, Lisa Chong, and Sacha Vignieri. Introduction to the Special Issue on Data Replication and Reproducibility. Science,Vol 334, 2 December 2011.

[Jenkins 2015] The Jenkins Extensible Open Source Continuous Integration Server. Available from https://jenkins-ci.org/. Last accessed 3 August 2015.

[Joppa et al 2013] "Troubling Trends in Scientific Software Use." Noppa et al

[JORS 2015] Journal of Open Research Software. http://openresearchsoftware.metajnl.com/.

[Kattge et al 2014] "Of carrots and sticks." Jens Kattge, Sandra Díaz, and Christian Wirth. Nature Geoscience, 7, 778–779, 2014. doi:10.1038/ngeo2280

[Kenett and Shmueli 2015] "Clarifying the terminology that describes scientific reproducibility." Ron S Kenett and Galit Shmueli. Nature Methods 12, 699, 2015. doi:10.1038/nmeth.3489

[Khatib et al 2011] Khatib, F., F. DiMaio, Foldit Contenders Group, Foldit Void Crushers Group, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, M. Jaskolski, and D. Baker. "Crystal structure of a monomeric retroviral protease solved by protein folding game players." Nature Struct Mol Biol. Sep 18;18(10):1175-7, 2011.

[Klein et al 2014] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard 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 et al 2011] "A provenance-based infrastructure to support the life cycle of executable papers." David Koop, Emanuele Santos, Phillip Mates, Huy T Vo, Philippe Bonnet, Bela

[Lehrer et al 2010] Lehrer, J. "The Truth Wears Off: Is There Something Wrong with the Scientific Method?" The New Yorker, December 13, 2010. Available from http://www.newyorker.com/reporting/2010/12/13/101213fa_fact_lehrer

[Leisch 2002] Leisch, F. "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 et al 2009] LeVeque, R J, Mitchell, I M and Stodden, V. (2009). Reproducible research for scientific computing: Tools and strategies for changing the culture. Computing in Science & Engineering 14(4): 13.DOI: http://dx.doi.org/10.1109/MCSE.2012.38

[Lintott et al 2010] Lintott, Chris, Schawinski, Steven Bamford, Anže Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C. Nichol, M. Jordan Raddick, Alex Szalay, Dan Andreescu, Phil Murray, Jan Vandenberg. "Galaxy Zoo 1: data release of morphological classifications for nearly 900,000 galaxies". Monthly Notices of the Royal Astronomical Society, 2010.

[Ludaescher et al 2006] “Scientific workflow management and the Kepler system.” Ludaescher, B.; Altintas, I.; Berkley, C.; Higgins, D.; Jaeger, E.; Jones, M.; Lee, E. A.; Tao, J. & Zhao, Y. Concurrency and Computation: Practice and Experience, 18, 1039-1065, 2006.

[Macleod 2014] “Biomedical research: increasing value, reducing waste.” Malcolm R Macleod, Susan Michie, Ian Roberts, Ulrich Dirnagl, Iain Chalmers, John P A Ioannidis, Rustam Al-Shahi Salman, An-Wen Chan, Paul Glasziou. The Lancet, Volume 383, No. 9912, p101–104, 11 January 2014. DOI: http://dx.doi.org/10.1016/S0140-6736(13)62329-6

[Manegold et al 2010] “Repeatability & workability evaluation of SIGMOD 2009.” Manegold, S, Manolescu I, Afanasiev L, Feng J, Gou G, Hadjieleftheriou M, Harizopoulos S, Kalnis P, Karanasos K, Laurent D, Lupu M, Onose N, Ré C, Sans V, Senellart P, Wu T, Shasha D. SIGMOD Record 38, 2010. Available from http://www.sigmod.org/sigmod/record/issues/0909/p40.open.repeatability2009.pdf

[Manolescu et al 2008] “The repeatability experiment of SIGMOD 2008” Ioana Manolescu, Loredana Afanasiev, Andrei Arion, Jens Dittrich, Stefan Manegold, Neoklis Polyzotis, Karl Schnaitter, Pierre Senellart, Spyros Zoupanos, Dennis Shasha. ACM SIGMOD Record 37(1), 2008. Available from 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 from http://www.mathworks.com/products/matlab/. Last accessed 3 August 2015.

[McCaffrey 2005] McCaffrey, R. E. “Using Citizen Science in Urban Bird Studies.” Urban Habitats, 3, 1, 70-86, 2005.

[Mesirov 2010] Mesirov, J. P. “Accessible Reproducible Research.” Science, 327:415, 2010. Available from http://www.sciencemag.org/cgi/rapidpdf/327/5964/415?ijkey=WzYHd6g6IBNeQ&keytype=ref&siteid=sci

[Michener 2015] William K. Michener. (2015) Ecological data sharing. Ecological Informatics 29, 33-44, 2015. doi:10.1016/j.ecoinf.2015.06.010

[Missier et al 2010] Missier, P., Sahoo, S. S., Zhao, J., 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 et al 2014] “Development and exploitation of a controlled vocabulary in support of climate modeling.” Moine, M.-P., Valcke, S., Lawrence, B. N., Pascoe, C., Ford, R. W.,

[Mooney and Newton 2012] Mooney, H, Newton, MP. (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. Journal of Librarianship and Scholarly Communication 1(1):eP1035. http://dx.doi.org/10.7710/2162-3309.1035

[Moreau and Ludaescher 2007] Moreau, L. and B. Ludaescher, editors. Special Issue on the First Provenance Challenge, volume 20. Wiley, April 2007.

[Moreau et al 2011] Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., and denBussche, J. V. "The Open Provenance Model Core Specification (v1.1)." Future Generation Computer Systems, 27(6), 2011. Preprint available from http://www.bibbase.org/cache/www.isi.edu__7Egil_publications.bib/moreau-etal-fgcs11.html

[Moreau et al 2013] "PROV-DM: The PROV Data Model." Moreau, L.; Missier, P.; Belhajjame, K.; B'Far, R.; Cheney, J.; Coppens, S.; Cresswell, S.; Gil, Y.; Groth, P.; Klyne, G.; Lebo, T.; McCusker, J.; Miles, S.; Myers, J.; Sahoo, S.; and Tilmes, C. World Wide Web Consortium (W3C) Recommendation, April 2013. Available from http://www.w3.org/TR/prov-dm/

[Morozov et al 2006] "A generalized web service model for geophysical data processing and modeling". Morozov, Igor; Reilkoff, Brian; Chubak, Glenn. Computers & Geosciences 32 (9): 1403, 2006. doi:10.1016/j.cageo.2005.12.010.

[Naik 2011] "Scientists' Elusive Goal: Reproducing Study Results." The Wall Street Journal, December 2, 2011.

[NASA 2015] http://science.nasa.gov/media/medialibrary/2014/12/05/NASA_Plan_for_increasing_access_to_results_of_federally_funded_research.pdf

[Nature 2006] Nature Editorial. "Illuminating the Black Box." Nature, 442(7098), 2006. Available from http://www.nature.com/nature/journal/v442/n7098/full/442001a.html

[Nature Metrics 2010] Metrics Survey Results. Nature Metrics, 2010. http://www.nature.com/nature/newspdf/metrics_survey.pdf

[Nature 2012a] "Error Prone: Biologists must realize the pitfalls of work on massive amounts of data." Nature, Vol 487, 26 July 2012.

[Nature 2012b] "Must Try Harder". Nature, Vol 483, March 2012.

[Nature 2013] "Reproducing our Irreproducibility." Nature Vol 496 pp 398, 25 April 2013.

[Nature 2014a] "Journals Unite for Reproducibility." Marcia McNutt. Nature 515, 7, 06 November 2014. doi:10.1038/515007a

[Nature 2014b] "Code Share." Nature 514, pp 536, 30 October 2014. doi:10.1038/514536a

[Nature 2015] "Nature.com Ontologies." Nature, 2015. Available from http://www.nature.com/ontologies/

[Nature Geoscience 2015] "Towards transparency." Nature Geoscience 7, 777, 2014. doi:10.1038/ngeo2294

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

[Nekutrenko and Taylor 2012] “Next-Generation Sequencing Data Interpretation: Enhancing Reproducibility and Accessibility.” Anton Nekutrenko and James Taylor. Nature Reviews Genetics 13, 667-672, September 2012. doi:10.1038/nrg3305

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

[Nielsen 2011] Nielsen M. “Reinventing Discovery.” Princeton University Press, 2011.

[NIH 2015] “Principles and Guidelines for Reporting Preclinical Research.” National Institutes of Health, 2015. Available from http://www.nih.gov/about/reporting-preclinical-research.htm. Last accessed 3 August 2015.

[Nowakowskia et al 2011] “The Collage Authoring Environment.” Piotr Nowakowskia, Eryk Ciepielaa, Daniel Harężlaka, Joanna Kocota, Marek Kasztelnika, Tomasz Bartyńskia, Jan Meiznera, Grzegorz Dyka, Maciej Malawskib. Procedia Computer Science, Volume 4, 2011, Pages 608–617, 2011.

[NOAA 2015] http://docs.lib.noaa.gov/noaa_documents/NOAA_Research_Council/NOAA_PARR_Plan_v5.04.pdf

[Oinn et al 2006] “Taverna: lessons in creating a workflow environment for the life sciences.” Oinn, T., M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe. Concurrency and Computation: Practice and Experience, 18(10), 2006.

[ORCID 2015] The Community Surface Dynamics Modeling System (ORCID). Available from http://www.orcid.org. Last accessed 3 August 2015.

[OS 2015] Open Source Initiative. Available from http://opensource.org/licenses. Last accessed 3 August 2015.

[Pampel et al 2013] “Making Research Data Repositories Visible: The re3data.org Registry.” Heinz Pampel, Paul Vierkant, Frank Scholze, Roland Bertelmann, Maxi Kindling, Jens Klump, Hans-Jürgen Goebelbecker, Jens Gundlach, Peter Schirmbacher, Uwe Dierolf. PLOS ONE, November 4, 2013. DOI: 10.1371/journal.pone.0078080

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

[Pearson and Lipman 1988] “Improved tools for biological sequence comparison.” William R. Pearson and David J. Lipman. Proceedings of the National Academies of Sciences 85(8):2444-8, 1988.

[Pebesma et al 2012] Pebesma E., D.Nüst and R.Bivand, (2012), The R software environment in reproducible geoscientific research, Eos Trans. AGU, 93(16), 163.

[Peckham et al 2013] Scott D. Peckham, Eric W.H. Hutton, and Boyana Norris. “A component-based approach to integrated modeling in the geosciences: The design of CSDMS.” Computers and Geosciences, 53, 2013. DOI: http://dx.doi.org/10.1016/j.cageo.2012.04.002

[Peckham 2014] "The CSDMS Standard Names: Cross-Domain Naming Conventions for Describing Process Models, Data Sets and Their Associated Variables." Scott D. Peckham. Proceedings of the Seventh International Congress on Environmental Modeling and Software, June 2014.

[Peckham 2015] Community Surface Dynamics Modeling System Basic Model Interface, available from: http://csdms.colorado.edu/wiki/BMI_Description, accessed on June 30, 2015.

[Peng 2009] "Reproducible Research and Biostatistics." Roger D. Peng. Biostatistics 10 (3): 405-408, 2009. doi:10.1093/biostatistics/kxp014

[Piwowar et al 2007] "Sharing Detailed Research Data Is Associated with Increased Citation Rate." Heather A. Piwowar, Roger S. Day, Douglas B. Fridsma. PLoS ONE, March 21, 2007. DOI: 10.1371/journal.pone.0000308

[Piwowar and Chapman 2009] Piwowar HA, Chapman WW. Public sharing of research datasets: a pilot study of associations. Journal of Informetrics. 2010;4(2):148-156. doi:10.1016/j.joi.2009.11.010.

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

[Pope et al 2014] Pope, Allen, Rees, W. Gareth, Fox, Adrian J., and Andrew Fleming. "Open Access Data in Polar and Cryospheric Remote Sensing." Remote Sens. 2014, 6, 6183-6220; doi:10.3390/rs6076183. [Prinz et al 2011] "Believe it or not: how much can we rely on published data on potential drug targets?" Florian Prinz, Thomas Schlange, and Khusru Asadullah. Nature Reviews Drug Discovery 10, 712, September 2011. doi:10.1038/nrd3439-c1

[Priem et al 2010] "Altmetrics: A manifesto." Dario Priem, Jason Taraborelli, Paul Groth, and Cameron Neylon 26 October 2010. Available from http://altmetrics.org/manifesto. Last accessed 3 August 2015.

[PURL 2015] Persistent URLs. Available from http://www.purl.org. Last accessed 3 August 2015.

[Re3data 2015] The Registry of Research Data Repositories (re3data). Available from http://www.re3data.org. Last accessed 3 August 2015.

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

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

[Reichman et al 2011] "Challenges and opportunities of open data in ecology." Reichman OJ, Jones MB, Schildhauer MP. Science. 2011 Feb 11;331(6018):703-5. doi: 10.1126/science.1197962.

[Rocca et al 2012] Rocca RA, Magoon G, Reynolds DF, Krahn T, Tilroe VO, et al. "Discovery of Western European R1b1a2 Y Chromosome Variants in 1000 Genomes Project Data: An Online Community Approach." PLoS ONE 7(7), 2012.

[Roston 2015] “Retracted Scientific Studies: A Growing List.” Michael Roston. The New York Times, May 28, 2015.

[Royal Society 2012] “Final Report: Science as an Open Enterprise.” Royal Society, 2012. Available from https://royalsociety.org/policy/projects/science-public-enterprise/Report/

[Russell 2013] “If a job is worth doing, it is worth doing twice.” Jonathan F. Russell. Nature 496, 7, 04 April 2013. doi:10.1038/496007a

[Ryan 2011] “Replication in Field Biology: The Case of the Frog-Eating Bat.” Michael J. Ryan. Science Vol 334, 2 December 2011.

[Savage and Vickers 2009] “Empirical Study of Data Sharing by Authors Publishing in PLoS journals”. Caroline J. Savage and Andrew J. Vickers. PLoS ONE, Volume 4 Issue 9, September 2009.

[Santer et al 2011] "The reproducibility of observational estimates of surface and atmospheric temperature change." Santer BD, Wigley TML, and Taylor KE. Science 334.6060 (2011): 1232-1233. [Savage 2012] Savage, N. “Gaining Wisdom from Crowds.” Communications of the ACM, Vol. 55 No. 3, Pages 13-15, March 2012.

[Schilling 2003] M. Schilling. Science 31 October 2003: Vol. 302 no. 5646 pp. 787-788 DOI: 10.1126/science.1088234

[Science 2014] “Journals Unite for Reproducibility.” Science Vol. 346 no. 6210 p. 679, 7 November 2014. DOI: 10.1126/science.aaa1724

[SCFBM 2015] Source Code for Biology and Medicine, 2015. http://www.scfbm.org/

[Schooler 2014] “Metascience could rescue the ‘replication crisis.’” Jonathan W. Schooler. Nature 515, 9, 06 November 2014. doi:10.1038/515009a

[Schwab et al 2000] “Making Scientific computations reproducible.” Schwab, M.; Karrenbach, N.; Claerbout, J. Computing in Science & Engineering, 2(6), pp.61-67, Nov.-Dec. 2000. Available from http://sep.stanford.edu/lib/exe/fetch.php?id=sep%3Aresearch%3Areproducible&cache=cache&media=sep:research:reproducible:cip.pdf

[Scientific American 2010] Scientific American. “In Science We Trust: Poll Results on How you Feel about Science” Scientific American, October 2010. Available from http://www.scientificamerican.com/article.cfm?id=in-science-we-trust-poll

[Scientist 2010] The Scientist. “Top Retractions of 2010.” The Scientist, December 16, 2010. Available from http://www.the-scientist.com/news/display/57864/

[Shen 2014] “Interactive notebooks: Sharing the code.” Helen Shen. Nature, 515, 151–152 (06 November 2014) doi:10.1038/515151a

[Shen 2014] “Interactive notebooks: Sharing the code.” Helen Shen. Nature Tookbox, 05 November 2014. Available from http://www.nature.com/news/interactive-notebooks-sharing-the-code-1.16261

[SoftwareX 2015] SoftwareX journal. http://www.journals.elsevier.com/softwarex/.

[Soranno et al 2014] “It’s Good to Share: Why Environmental Scientists’ Ethics are Out of Date.” Patricia A. Soranno, Kendra S. Cheruvelil, Kevin C. Elliott and Georgina M. Montgomery. BioScience, 2014. doi: 10.1093/biosci/biu169

[SourceForge 2015] SourceForge. Available from http://sourceforge.net/. Last accessed 3 August 2015.

[Spies et al. 2012] “The reproducibility of psychological science.” Jeffrey Spies et al. Report of the Open Science Collaboration. Available from opensciencframework.org/reproducibility/ [Vandewalle et al 2009] “What, why and how of reproducible research in signal processing.” P. Vandewalle, J. Kovačević and M. Vetterli. IEEE Signal Processing, May 2009.

[Starr et al 2015] “Achieving human and machine accessibility of cited data in scholarly publications.” Starr J, Castro E, Crosas M, Dumontier M, Downs RR, Duerr R, Haak LL, Haendel M, Herman I, Hodson S, Hourclé J, Kratz JE, Lin J, Nielsen LH, Nurnberger A, Proell S, Rauber A, Sacchi S, Smith A, Taylor M, Clark T. PeerJ Computer Science 1:e1, 2015. DOI: 10.7717/peerj-cs.1

[Stodden 2009] "The Legal Framework for Reproducible Research in the Sciences: Licensing and Copyright", Victoria Stodden. IEEE Computing in Science and Engineering, 11(1), January 2009.

[Stodden et al 2013] Stodden V, Guo P, Ma Z (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 et al., 2012] Taylor, K. E. and Stouffer, R. J. and Meehl, G. A. (2012), An Overview of Cmip5 and the Experiment Design, Bulletin of the American Meteorological Society, doi: 10.1175/Bams-D-11-00094.1

[Tenopir et al 2011] “Data Sharing by Scientists: Practices and Perceptions.” Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, Mike Frame. PLOS One, June 2011, DOI: 10.1371/journal.pone.0021101

[Tonella, P. and Potrich, A., 2005] Reverse Engineering of Object Oriented Code, Monographs in Computer Science, Springer Science + Business Media, Inc., Boston, 208 p.

[Travis-CI 2015] The Travis Continuous Integration Service. Available from https://github.com/travis-ci. Last accessed 3 August 2015.

[Uhlir et al 2012] “For Attribution: Developing Data Attribution and Citation Practices and Standards.” Paul F. Uhlir, 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 from http://www.nap.edu/catalog/13564/for-attribution-developing-data-attribution-and-citation-practices-and-standards.

[USGS 2015a] The US Geological Survey Science Data Catalog. Available from http://data.usgs.gov. Last accessed 3 August 2015.

[USGS 2015b] http://www.usgs.gov/datamanagement/policyreferences.php. Last accessed 3 August 2015.

[Van Gorp and Mazanekb 2011] "SHARE: a web portal for creating and sharing executable research papers", Procedia Computer Science, Volume 4, doi:10.1016/j.procs.2011.04.062, 589-597.

[Van Noorden 2013] "Data-sharing: Everything on display." Richard Van Noorden Nature 500, 243-245 (2013) doi:10.1038/nj7461-243a

[Van Noorden 2015] "Sluggish data sharing hampers reproducibility effort." Richard Van Noorden. Nature, 2015. doi:10.1038/nature.2015.17694

[Vasilevsky et al 2013] "On the reproducibility of science: unique identification of research resources in the biomedical literature." Nicole A. Vasilevsky, Matthew H. Brush, Holly Paddock, Laura Ponting, Shreejoy J. Tripathy, Gregory M. LaRocca, Melissa A. Haendel. PeerJ 1:e148, 2013. https://dx.doi.org/10.7717/peerj.148

[VHub 2015] VHub: The Collaborative Volcano and Risk Mitigation Hub. Available from https://vhub.org. Last accessed 3 August 2015.

[Vines et al 2014] "The Availability of Research Data Declines Rapidly with Article Age." Timothy H. Vines, Arianne Y.K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, Diana J. Rennison. Current Biology 24, 94-97, 2014.

[Vision 2010] "Open Data and the Social Contract of Scientific Publishing." Todd J. Vision. BioScience Volume 60, Issue 5, Pp. 330-331, 2010. doi: 10.1525/bio.2010.60.5.2

[WaterML 2015] WaterML 2.0. Open Geospatial Consortium. Available from http://www.opengeospatial.org/standards/waterml/. Last accessed 3 August 2015.

[Wilson et al 2012] "RepliCHI SIG – from a panel to a new submission venue for replication." Max L. Wilson, Wendy Mackay, Ed H. Chi, Michael S Bernstein, Jeffrey Nichols. ACM SIGCHI, 2012.

[Woelfle et al 2011] "Open Science is a Research Accelerator." Michael Woelfle, Piero Olliaro, Matthew H. Todd. Nature Chemistry, Vol 3, October 2011.

[Yong 2012] "Replication studies: Bad copy." Ed Yong. Nature 485, 298–300, 17 May 2012. Available from doi:10.1038/485298a

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

[Zudilova-Seinstra 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, http://www.elsevier.com/connect/designing-the-article-of-the-future