

Do You Speak Open Science? Resources and Tips to Learn the Language.

Paola Masuzzo^{1, 2} - ORCID: 0000-0003-3699-1195, Lennart Martens^{1,2} - ORCID: 0000-0003-4277-658X

Author Affiliation

¹ Medical Biotechnology Center, VIB, Ghent, Belgium

² Department of Biochemistry, Ghent University, Ghent, Belgium

Abstract

The internet era, large-scale computing and storage resources, mobile devices, social media, and their high uptake among different groups of people, have all deeply changed the way knowledge is created, communicated, and further deployed. These advances have enabled a radical transformation of the practice of science, which is now more open, more global and collaborative, and closer to society than ever. Open science has therefore become an increasingly important topic. Moreover, as open science is actively pursued by several high-profile funders and institutions, it has fast become a crucial matter to all researchers. However, because this widespread interest in open science has emerged relatively recently, its definition and implementation are constantly shifting and evolving, sometimes leaving researchers in doubt about how to adopt open science, and which are the best practices to follow.

This article therefore aims to be a field guide for scientists who want to perform science in the open, offering resources and tips to make open science happen in the four key areas of data, code, publications and peer-review.

The Rationale for Open Science: Standing on the Shoulders of Giants

One of the most widely used definitions of open science originates from Michael Nielsen [1]: “Open science is the idea that scientific knowledge of all kinds should be openly shared as early as is practical in the discovery process”. With this in mind, the overall goal of open science is to accelerate scientific progress and discoveries and to turn these discoveries into benefits for all. An essential part of this process is therefore to guarantee that all sorts of scientific outputs are publicly available, easily accessible, and discoverable for others to use, re-use, and build upon.

As Mick Watson has recently wondered, “[...] isn’t that just science?” [2]. One of the basic premises of science is that it should be based on a global, collaborative effort, building on open communication of published methods, data, and results. In fact, the concept of discovering truth by building on previous findings can be traced back to at least the 12th century in the metaphor of dwarfs standing on the shoulders of giants: “*Nanos gigantum humeris insidentes*”¹.

While creativity and intuition are contributed to science by individuals, validation and confirmation of scientific findings can only be reached through collaborative efforts, notably peer-driven quality control and cross-validation. Through open inspection and critical, collective analysis, models can be refined, improved, or rejected. As such, conclusions formulated and validated by the efforts of many take prominence over personal opinions and statements, and this

¹ Metaphor attributed to Bernard of Chartres, and better known in its English form as found in a 1676 letter of Isaac Newton: “If I have seen further, it is by standing on the shoulders of giants”

is, in the end, what science is about. While science has been based for centuries on an open process of creating and sharing knowledge, the quantity, quality, and speed of scientific output have dramatically changed over time. The beginning of scholarly publication as we intend it today can be traced back to the 17th century with the foundation of the ‘Philosophical Transactions’. Before that, it was not at all unusual for a new discovery to be announced in an encrypted message (*e.g.*, as an anagram) that was usually indecipherable for anyone but the discoverer: both Isaac Newton and Leibniz used this approach. However, since the 17th century, the increasing complexity of research efforts led to more (indirect) collaborations between scientists. This in turn led to the creation of scientific societies, and to the emergence of scientific journals dedicated to the diffusion of scientific research. Paradoxically however, knowledge diffusion has dramatically slowed down over the same time. In his review of Michael Neilsen’s book “Reinventing Discovery” [3], Timo Hannay describes science as “self-serving” and “uncooperative”, “replete with examples of secrecy and resistance to change”, and furthermore defines the natural state of researchers as “one of extreme possessiveness” [4]. Hannay might have a point: the majority of research papers are behind a paywall [5], researchers still fail at making data and metadata available [6], reproducibility is hampered by the lack of appropriate reporting of methodologies [7], software is often not released [8], and peer-review is anonymous and slow [9].

As a reaction, the open science movement was born, almost as a counterculture to the too-closed system that re-emerged over the past few decades. More and more academic and research institutions are currently opening up the science they produce, making the scientific research, produced data and associated papers accessible to all levels of an ever more inquiring society, amateur or professional. And increasingly, major funding agencies are mandating the same. For example, the European Commission requires participants of the H2020 funding framework to adhere to the Open Access mandate and the Open Research Data Pilot. Furthermore, both the National Institutes of Health (NIH) and the Wellcome Trust have developed specific mandates to enforce more open and reproducible research. As a result, practicing open science is no longer only a moral matter, but has become a crucial requirement for the funding, publication, and evaluation of research.

Because the many benefits of open science have already been extensively studied and reported [10–16], this article instead intends to be a user guide for open science. The next sections of this article therefore provide an overview of the key pillars of open science, along with resources and tips to make open science happen in everyday research practices. This collection of resources can then serve as an open science guidebook for early-career researchers, research laboratories, and the scientific community at large.

Four Pillars of Open Science

Almost all scientists today will have bumped into the expression “open science”. As an umbrella term used to cover any kind of change towards availability and accessibility of scientific knowledge, “open science” evokes many different concepts and covers many different fronts, from the right to have free access to scholarly publications (dubbed “open access”), over the demand for a wider public engagement (typically referred to as citizen science), to the development of free tools for collaboration and open peer-review (as implemented in science-oriented social media platforms).

This diversity and perhaps even ambiguity of open science can be explained by the many stakeholders that are directly affected by a changing scientific environment: researchers,

administrators, funders, policy makers, libraries, publishing companies, and even the general public. Five different schools of thought on open science have been identified², each with their stakeholder groups, their aims, and their tools and methods to achieve and promote these aims [12]. While these schools depict the whole scope of open science, their fundamental aim is to enhance openness in the four widely recognized thematic pillars: open research data, open software code, open access to papers, and open peer-review (*Figure 1*). The following sections will briefly introduce the rationale for each of these pillars, and will then provide resources for their adoption in daily research practice.

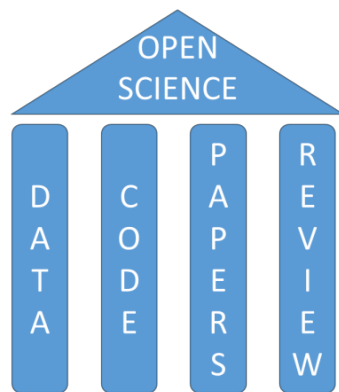


Figure 1: The four pillars of open science discussed in this article.

Image adapted from [17], distributed under a CC BY 4.0 International license (<http://creativecommons.org/licenses/by/4.0/>).

Open Data: Sharing the Main Actor of a Scientific Story

By open data in science we mean data that are freely available on the public internet permitting any user to download, copy, analyze, re-process, or use these for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself³.

In the digital era, data are more and more considered to be the main part of a scientific publication, while the paper serves the secondary role of describing and disseminating scientific results. This because open data tend to outlive the associated paper. In fact, others (professional researchers as well as interested members from the general public) can conduct re-analyses on these data, and can do so within the context of new questions, leading to new scientific discoveries. In 2015 Borgman identified four rationales for sharing research data: to reproduce research, to make those data that can be considered public assets, available to the public⁴, to leverage investments in research, and to advance research and innovation [18]. Several studies have furthermore reported that scientific papers accompanied by publicly available data are on average cited more often [19,20], and are moreover characterized by fewer statistical errors and a greater degree of robustness [21].

² Democratic, Pragmatic, Infrastructure, Public and Measurement

³ see the full Open Definition at: <http://opendefinition.org/od/2.0/en/> and the Pantan Principles for Open Data in Science at <http://pantonprinciples.org>

⁴ Privacy sensitive data for instance, do not belong to this category.

Releasing data, however, is not sufficient by itself. For re-use to happen efficiently, which is ultimately the goal of open data, data sharing needs to become a custom routine, should encompass the full research cycle, and needs to assure long-term preservation. Furthermore, data sharing requires some amount of manual work, and a specific shift in research habits, for which the current credit system in research should accommodate. A nice example of this shift is provided by the journal *Psychological Science*, which adopted such an incentive for open research data in January 2014, by offering “badges” to acknowledge and signal open practices in publications. To receive an ‘open data’ badge, authors must make all digitally shareable data relevant to the publication available on an open access repository. Similarly, to earn an ‘open materials’ badge, authors must make all digitally shareable materials available on an open access repository. Those who apply for a badge and meet open data or open materials specifications receive the corresponding badge symbol at the top of their paper and provide an explicit statement in the paper including a URL to the data or materials at an open repository. A recent study has shown that these badges are effective incentives to improve the openness, accessibility, and persistence of data and materials that underlie scientific research [22].

Finally, for data sharing to encourage re-use, data curation and metadata annotations are key factors, together with reliable basic infrastructure for data sharing: the availability of data infrastructures that are well curated and well maintained in the long-term, and a rich catalogue of standards and formats that are moreover continuously updated to keep up with shifts in technology and knowledge.

Where to Submit Research Data? General-Purpose and Domain-Specific Repositories

As a general rule, data should be submitted to a repository prior to submission of a relevant manuscript that describes these data. Thus, the authors can point the readers to the location of the data in the manuscript itself, increasing transparency, reproducibility and validation of the results, and aiding efficient peer-review. Two types of such data repositories exist: general-purpose and domain-specific repositories. The former are inter-disciplinary repositories meant to host data for which domain-specific repositories do not exist, as well as general research output (such as posters, presentations, code). The latter on the other hand, are well-established subject or data-type specific repositories that typically serve specific fields. **Table 1** lists the most widely used repositories across both types. Although not exhaustive, this list provides a good cross-section of repositories that should be considered both for publication of data, and for the location and retrieval of relevant data for (re)use in research.

A global registry of research data repositories for different scientific disciplines can be found at the Registry of Research Data Repositories (<http://www.re3data.org>). Furthermore, NCBI and EBI online databases can be found at <http://goo.gl/0KwIq8> and <http://goo.gl/j3stqD>, respectively. Biomed Central suggests a list of possible repositories at <https://goo.gl/dBHeZf>, while another interesting list, maintained by Nature Scientific Data, can be found at <https://goo.gl/G7cLFp>. Finally, the Biosharing catalogue includes bioscience databases described according to domain guidelines and standards (<https://biosharing.org/databases/>, 798 databases listed at the time of writing).

Table 1: A list of general-purpose and domain-specific data repositories (in alphabetical order).

Name	Description	Domain	Website
Cell Image Library	public repository of reviewed and annotated images, videos, and animations of cells from a variety of organisms	biological imaging	http://www.cellimagelibrary.org
Coherent X-ray Imaging Data Bank	open repository for X-ray images	macromolecular structures	http://www.cxidb.org/id-2.html
Crystallography Open Database	open-access collection of crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding biopolymers	macromolecular structures	http://www.crystallography.net
DataOne	a framework and infrastructure for Earth observational data	environmental and ecological data	https://www.dataone.org
Dryad	a resource that makes the data underlying scientific publications discoverable, freely reusable, and citable	general-purpose	http://datadryad.org
Figshare	a repository where users can make all of their research outputs available in a citable, shareable and discoverable manner	general-purpose	https://figshare.com
GenBank	the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences	sequence and omics data	http://www.ncbi.nlm.nih.gov/genbank/
GEOSS portal	a portal for Earth science data	environmental and ecological data	www.geoportal.org
Global Biodiversity Information Facility	a repository containing data about all types of life on Earth, published according to common data standards	environmental and ecological data	http://www.gbif.org
JCB Data Viewer	a platform to view, analyze and share image data associated with articles published in The Journal of Cell Biology	biological imaging	http://jcb-dataviewer.rupress.org
Morphbank	an image database documenting a range of specimen-based research, including comparative anatomy and taxonomy	biological imaging	http://www.morphbank.net
Movebank	an online database of animal tracking data	environmental and ecological data	https://www.movebank.org
NERC data centers	seven centers for: marine, atmospheric, Earth observation, solar and space physics, terrestrial and	environmental and ecological data	http://www.nerc.ac.uk/research/sites/data/

	freshwater, geoscience, and polar and cryosphere data		
NeuroVault	a repository for statistical maps, parcellations, and atlases produced by MRI and PET studies	neuroimaging data	http://neurovault.org
NIH 3D Print Exchange	a repository with models for 3D printers and tools to create and share 3D-printable models related to biomedical science	3D-printable models	http://3dprint.nih.gov
Open Energy Information	a crowdsourced collection of information, data and discussions around multiple aspects of energy	engineering	http://en.openei.org
Open Science Framework	a research and workflow management tool and open repository; allows for integration with several external tools like Dropbox, Github, and Zotero	general-purpose	https://osf.io
OpenfMRI	a project dedicated to the free and open sharing of functional magnetic resonance imaging (fMRI) datasets	neuroimaging data	https://openfmri.org
OpenTrials	a project to locate, match, and share all publicly accessible data and documents on all trials conducted on all medicines and other treatments	health data	http://opentrials.net
PANGAEA	a repository for geospatial data	environmental and ecological data	https://www.pangaea.de
PRIDE	an archive of protein expression data as determined by mass spectrometry	sequence and omics data	http://www.ebi.ac.uk/pride/archive/
Protein Data Bank	a databank for 3D protein structures	macromolecular structures	http://www.rcsb.org/pdb/home/home.do
The Knowledge Network for Biocomplexity	an international repository intended to facilitate ecological and environmental research	environmental and ecological data	https://knb.ecoinformatics.org
Uniprot	a comprehensive resource for protein sequence and functional annotation data	sequence and omics data	http://www.uniprot.org
Worldwide Protein Data Bank	a publicly available repository of macromolecular structural data	macromolecular structures	http://www.wwpdb.org
Zenodo	a repository that supports a wide variety of content including publications, presentations, images, software (integration with GitHub), and data	general-purpose	https://zenodo.org

Submitting data: points to consider

The following section highlights some key aspects to keep in mind when submitting research data.

- Research materials in a broad sense (essentially any research output such as figures, posters, code, presentations, and media) are best deposited in general-purpose repositories. Domain-specific data on the other hand, are best submitted to a domain-specific repository (see **Table 1**). Recent surveys have shown that the majority of researchers still prefer to share data as supplementary material to an article, but this is certainly not an optimal solution, because it is essentially a very static representation of data (often also formatted in document rather than data mark-up formats, such as PDF) and therefore does not allow for dynamic inspection and re-use of the data. It may also not represent a long-term data storage solution.
- If researchers wish to **publish data sets** through a data article, they can target appropriate data journals. Rather than presenting any analysis, results, or conclusions on the data, such a data article focuses on detailed descriptions of these data, and presents arguments about the value of the data for future (re-)analysis. Notable examples of data journals are: GigaScience (BioMed Central, <http://gigascience.biomedcentral.com>), Scientific Data (Nature Publishing Group, <http://www.nature.com/sdata/>) and Data in Brief (Elsevier, <http://www.journals.elsevier.com/data-in-brief/>). A data journal will not normally host data itself but will instead recommend a suitable repository where the data set should be deposited, and then link to it.
- When targeting a particular **journal** to publish their research, scientists should check for any **policies on data**. In fact, journals are increasingly requiring authors to deposit the data underlying their articles in a recognized repository, to complement or even replace any in-house facility for supplementary materials. For example, Public Library of Science (PLOS) recommends repositories it recognizes as “trusted within their respective communities” and also points to re3data as a more general source.
- The following questions can assist a researcher in choosing the right repository for their data:
 - *Is the repository well known?*
Is it community-recognized (e.g., listed in the re3data registry)? Some repositories are certified, meaning that they have passed a check in terms of reliable and long-term access to the data collections they host, but one should keep in mind that some good repositories are not compliant yet, and this might remain the case for some time.
 - *Will the repository accept my data?*
With the obvious exception of general-purpose repositories, most online databases accept data sets that relate to a specific research topic or domain, typically also formatted in a specific way. Three key aspects therefore need to be taken into account: (1) the data must be of a specific data type (e.g., microarrays, or biological imaging); (2) the data must be submitted in a specific data format (most likely an open, standard format instead of proprietary ones); (3) specific legal terms and conditions need to be satisfied (e.g., informed consent forms must be collected for health data).
- Use a recognized **waiver or license** that is appropriate for data. The OpenDefinition project lists conformant licenses (both for content and data): <http://opendefinition.org/licenses/>. Importantly, licenses non-conformant to the open definition are also reported: <http://opendefinition.org/licenses/nonconformant/>. As a general rule, it is important to remember that the use of licenses which limit commercial re-use or limit the production of derivative works by excluding use for specific purposes is discouraged. This because these

licenses can make it quite a bit harder to effectively re-use datasets, and could also prevent (tangential) commercial activities that could be used to support data preservation in the long-term⁵.

- Share the **metadata** along with the data. As Gray has put it: “Data is incomprehensible and hence useless unless there is a detailed and clear description of how and when it was gathered, and how the derived data was produced” [23]. Clear metadata make it easier to understand if data are appropriate for a project; without clear metadata data sets can be overlooked or even go unused. Worse yet, such data sets may be misinterpreted. The recently released FAIR (Findability, Accessibility, Interoperability, and Reusability) guidelines are a good starting point to check for efficient metadata reporting [24].

- Whenever possible, use **standard file formats**.

This applies for both data and metadata file formats. The Biosharing registry lists a comprehensive collection of standards for the life sciences (<https://biosharing.org/standards/>) (663 standards at the time of writing). To ensure that both data and metadata are reported accurately and compliant with community-established standards, use (semantic) validation tools, whenever available.

Open Source: Sustainable Software for Sustainable Science

Open source refers to software that is made available under a license that permits anyone to use, change, improve, or derive from existing source code, and sometimes even to distribute the software⁶. The case for open source code is straightforward: the code researchers write and use to analyze data is a vital part of the scientific research cycle, and, similar to data, is not only necessary to reproduce and interpret the results and corresponding conclusions, but can also be used to answer novel research questions. Therefore, if researchers write code as a means to obtain results from data, then this code should be released as well [8]. Clear arrangements for the storage and preservation of the code should be made, instructions need to be provided that will allow the code to be compiled and run without issue, and the code should be accompanied by a description of the core functionalities and hard- and software requirements for its use. This in turn means that source code alone is not sufficient: the software environment needs to be described too, including for instance, any linked libraries, any runtime environments or virtual machines, The open source container engine Docker is intended to provide an efficient solution for computational reproducibility (see www.docker.com) [25,26].⁷

Researchers sometimes prefer not to share code because of a lack of complete and clear documentation. While documentation is undoubtedly essential for code validation and re-use, as a general rule, sharing undocumented code is preferable to not sharing code at all [27]. Another concern that might stop researchers from sharing their code is the fear that they will have to provide full user support afterwards. One solution to this problem is to setup a simple online mailing list (for example through Google), and point all users to ask questions through it. In this way, answers are searchable on the web and available to other users who might have the same issue/question. In fact, this system utilizes a core property of open source code, in that a community can come into being around useful code. This community can then maintain, support, and update this code even in the absence of the original author.

⁵ see also the Panton Principles for Open Data at: <http://pantonprinciples.org>

⁶ see the full Open Source definition at the Open Source Initiative webpage: <https://opensource.org/docs/osd>

⁷ see also: <http://goo.gl/oba1qN>

It should however, be noted that many of the issues with code quality and sharing can actually be addressed by following simple best practices in code organization and planning. For instance, a key tool that all research programmers should incorporate into their workflow is the use of a Version Control System (VCS) such as git [28] or subversion (SVN). A VCS provides a way for taking snapshots of evolving code that allow tracking of changes, and for reverting these if necessary (*e.g.*, after making a change that ends up breaking the functionality of the code). A rapidly growing community of scientists use the Github platform (<https://github.com>), which is a freely available implementation of the git system, to contribute to collaborative projects, and to review and test code in a transparent and efficient way [29]. Interestingly, GitHub also promises to be a useful tool in assessing part of a researcher's impact. For example, a repository can be forked (which means there will be adaptations of the code), starred (showing appreciation for the work), pull requests can happen (which show public engagement with the work and the degree of potential collaboration), as well as downloads (which may signal software installations or code use).

Another interesting way to make code available is by integrating it with tools that enable data interrogation and interactive visualization. This approach, known as literate programming [30], seamlessly integrates analysis code, visualization plots, and explanations in the form of narrative text. There are a number of tools available to support this style of research, including Jupyter (for R, Python and Julia, <http://jupyter.org>), R Markdown (for R, <http://rmarkdown.rstudio.com>), and matlabweb (for MATLAB, <https://www.ctan.org/pkg/matlabweb>). With these tools, researchers can create code files (in the case of Jupyter these are called Notebooks⁸) that can be then shared on Github, in turn allowing other people to directly run these integrated code files through their browser, without having to install any additional software.

Resources for open source

- The **Software Sustainability Institute** provides further guidance on the benefits and methods of software preservation, including guidance on code repositories (<http://goo.gl/CE1OLY>).
- Another comparative list of source code hosting facilities is maintained on Wikipedia (<https://goo.gl/KfMfPu>).
- The **Open Source Initiative** (OSI, <http://opensource.org/>) is an organization dedicated to promoting open source software. Amongst the resources made available by the organization, a list of open source licenses is available at <https://opensource.org/licenses>. For each license, a full description is reported, together with terms and conditions of use.
- The **NumFOCUS** is a nonprofit organization that supports and promotes world-class, innovative, open source scientific software (<http://www.numfocus.org>). The mission of NumFOCUS is to promote sustainable high-level programming languages, open code development, and reproducible scientific research. A list of sponsored projects is available at <http://goo.gl/VQgw0M>. Amongst these:
 - The **IPython** (Interactive Python, <http://ipython.org>), with the Jupyter Notebook available at <https://jupyter.org>, and a gallery of interactive Notebooks available at <https://goo.gl/z3HgwH>.
 - The **rOpenSci** (R Open Science, <http://ropensci.org>), which promotes the open source R statistical environment for transparent and reproducible research. A

⁸ a gallery of Notebooks is available at <http://nb.bianp.net>

list of open source rOpenSci packages is available at <http://ropensci.org/packages/>. Some of these packages enable communication with widely used repositories, such as Figshare and Dryad.

- The **Software Carpentry** (<http://software-carpentry.org>) is a training organization that runs workshops and lessons to teach scientists basic computing skills. All educational materials are developed collaboratively online on GitHub (<https://github.com/swcarpentry>), and are distributed under the open CC-BY license.
- The **Data Carpentry** (<http://www.datacarpentry.org>) is a sister organization to Software Carpentry, and aims to teach basic concepts, skills and tools for working more effectively with data. Again, lessons and workshop materials are available online and distributed under the open CC-BY license.

Open Access: The Right to Knowledge

Open access is a term coined for the first time at the Open Access Budapest Initiative, and it refers to an unrestricted online access to scholarly research, primarily intended for scholarly journal articles⁹. The case for open access has been extensively reported in the literature [14,16,31–33]¹⁰. Essentially, advocates of open access want full access to, and use of, published scientific articles, moved by the core argument that publicly funded universities and granting bodies have a moral duty to make academic research output available on the web at no charge. Usage data from PubMedCentral (the online repository of the US National Institutes of Health) show that 25% of the daily unique users are from universities, 17% from companies, 40% of users are individual citizens, and the remaining 18% are from government or other categories (UNESCO, 2012).

To answer this call for open access to scientific publications, a variety of full open access journals have been launched in recent years, BioMed Central and PLOS are just two examples of publishers whose journals are all open access (see resources below for a comprehensive list). However, researchers may actively opt against open access journals as a possible venue for their research output. This reluctance is often related to the fact that the highest impact factors remain associated with subscription-based journals, and these are therefore more prestigious dissemination devices. However, as Sydney Brenner wrote twenty years ago, “Before we develop a pseudoscience of citation analysis, we should remind ourselves that what matters absolutely is the scientific content of a paper and that nothing will substitute for either knowing it or reading it” [34]. In the long term, it should be irrelevant where researchers publish their findings. What is important is that to speed up scientific progress, discovery and impact, research should be shared and made available without delay for others to use and to build upon.

Because citation rates and journal impact factors have become key evaluation criteria in funding decisions and research staff appointments and promotions, and because scientists are inherently rather conservative in their adoption of new approaches and tools, researchers should keep in mind that there still remain ways to make their work open, while still publishing in traditional subscription-based journals. Authors can make their work available on the web by posting preprints prior to formal peer-review and journal publication. This methodology is very well established in domains with lengthy peer-review cycles such as physics, astronomy, computer science, and

⁹ see the full definition of open access at the Budapest Open Access Initiative:

<http://www.budapestopenaccessinitiative.org/read>

¹⁰ see also: <http://www.nature.com/openresearch/about-open-access/benefits-for-authors/>

mathematics [35,36], with a very large amount of articles posted on the special-purpose arXiv repository every day. The overall use of preprints in the life sciences however, is still not significant, although a modest increase has been observed with the launch of PeerJ PrePrints and BioRxiv [37]. A list of all available preprint servers is given in **Table 2**.

Submitting unpublished work to a preprint server at (or even before) the time of submission brings two broad scientific benefits. First, it achieves free and immediate dissemination of the scientific results, and can solicit a wider input from the community that constitutes prompt feedback for possible improvement to the authors. Second, because preprints have DOIs assigned, these can be referenced even before the work is published in a journal. An interesting side-effect is that the DOI comes with a timestamp in the preprint archive, which can be important for priority claims.

Table 2: List of preprint servers.

Preprint server	Discipline	Webpage
arXiv	physics, mathematics, computer science, statistics, quantitative biology and finance	http://au.arXiv.org
bioRxiv	biology, genomics, biochemistry, bioinformatics, biophysics, life sciences at large	http://biorxiv.org
CERN Document Server	high-energy physics	https://cds.cern.ch/collection/Preprints
PeerJ Preprints	biology, medicine, computer science	https://peerj.com/archives-preprints/

Resources for open access

- **The Budapest Open Access Initiative** - <http://www.budapestopenaccessinitiative.org>
The BOAI has taken place in 2001 in response to the growing demand to make research free and available to anyone. BOAI is an active and diverse coalition that has issued guidelines on open access policies, licensing, infrastructures, and advocacy (<http://goo.gl/d2lVx0>).
- **Open Access Button** - <https://www.openaccessbutton.org/>
The goal of the OA Button project is to find the number of research outputs that are behind a subscription paywall. When looking for a research article, and not being able to access it, users can mark the article using the OA Button browser bookmarklet. When users bookmark those restricted items, the system automatically connects to CORE and Google Scholar and searches for an available open access version of the same research output, and links it back to the user.
- **Open Knowledge Maps** - <http://openknowledgemaps.org>
Launched very recently, Open Knowledge Maps is a large-scale system of open, interactive and interlinked knowledge maps spanning all fields of research (currently based on the PLOS library). *Figure 2* shows an example of a knowledge map for the query “cell migration”.

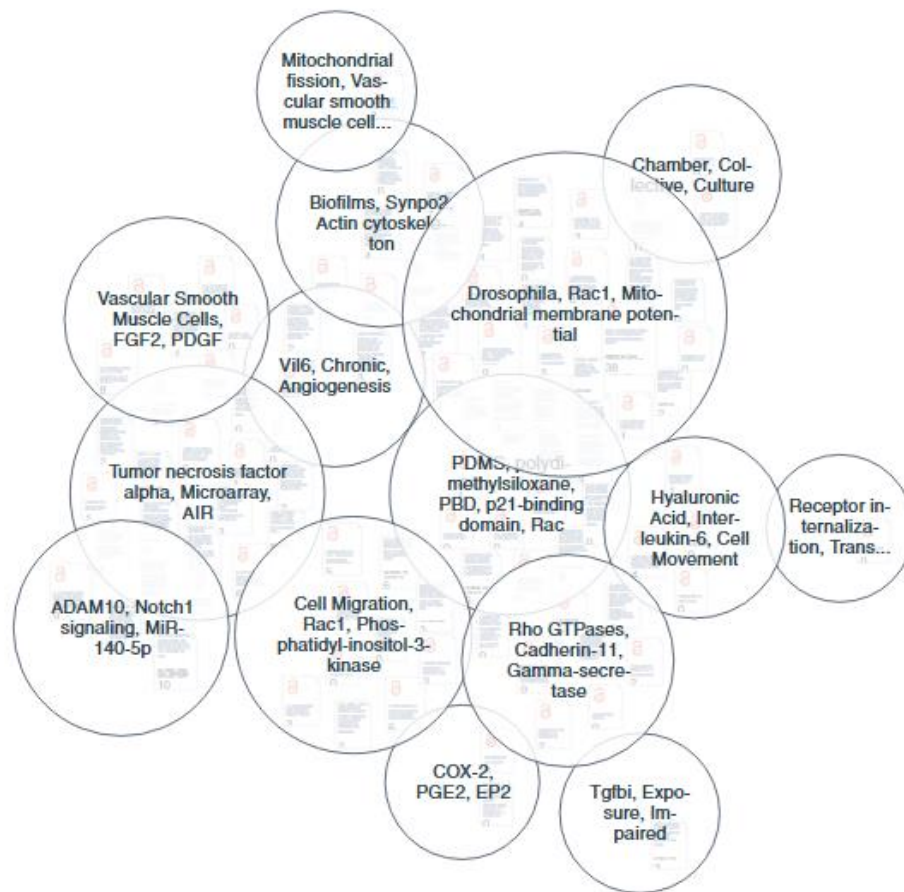


Figure 2: The Open Knowledge Map for the query “cell migration”.

- **Open Access Week** - <http://www.openaccessweek.org/>
Every year in October, the scholarly community celebrates the International Open Access Week with events around the world. These events can be registered on the Open Access Week website, which also contains plenty of open access advocacy material.
- **Open Access Tracking Project** - <http://bit.ly/o-a-t-p>
The OATP uses social tagging to capture new developments on open access to research. The OATP offers an updated catalogue of OA-related news and comments, and furthermore organizes knowledge of the open access field by tag or subtopic. The project also lists resources for open data, open educational resources, and anything related to open science. Researchers interested in the latest open access developments can also subscribe to the daily news feed.
- **SPARC Europe Open Access Diary** - <http://sparceurope.org/oadiaryeurope/>
To capture open access developments in Europe, the SPARC Europe Open Access Diary collects data from Europe from the OATP and then highlights the most important news in an interactive map, such as open access funders’ policies, presentations, and other news related to the movement.
- **SHERPA/RoMEO** - <http://www.sherpa.ac.uk/romeo/index.php>

RoMEO is part of SHERPA Services based at the University of Nottingham, and it allows authors to check policies from over 2,100 journals, *i.e.*, if public archiving of papers published in these journals is permitted, and to which level (pre-print and/or post-print and/or publisher's version/PDF).

Another list of academic journals by preprint policy is maintained on Wikipedia (<https://goo.gl/RFmlBw>).

- **Directory of Open Access Journals (DOAJ)** - <https://doaj.org>
The DOAJ is an online directory that indexes and provides access to high quality, open access, peer-reviewed journals. At the time of writing, the directory contains more than 10 000 open access journals covering all areas of science, technology, medicine, social science and humanities, and therefore constitutes a useful resource to guide researchers in their choice of open access journal.
- **Cofactor Journal Selector** - <http://cofactorscience.com/journal-selector>
The Cofactor UK company has developed an online tool, the journal selector, which allows researchers to look for journals that meet certain criteria, including the option for open peer-review and open access.
- **Eigenfactor Project** - <http://www.eigenfactor.org>
This project maintains a full list of no-fee open access journals for all research fields, which is accessible at: <http://www.eigenfactor.org/openaccess/fullfree.php>.
- **Open Access Overview** - <http://bit.ly/oa-overview>
This is an introduction to open access (OA) for those who are new to the concept, created and maintained by Peter Suber. Amongst useful resources, the Open Access Book is available at <http://bit.ly/oa-book>.

Open Peer-Review: Transparent Research Evaluation

An often heard complaint among researchers is that the peer-review system is 'broken'. A considerable number of articles have appeared in various journals that question the process, and how it is employed [38–41]. Most of these articles have raised issues with the consistency of review, its definition, ethics, cost, and the speed of the process [9].

Perhaps the first problem lies in the recognition of who peer-review is for. Peer-review is perhaps the best example of a community-wide way to practice science, and should provide authors with feedback on their work, preferably also with input for improving it. However, in most cases, peer-review also helps journal editors decide which submitted manuscripts not to publish. Furthermore, in most cases, the authors do not know the identity of their reviewers, and, with very few exceptions, these pre-publication reviews are discarded as soon as articles are published. This is unfortunate, as a lot of valuable context and insight goes to waste through this discarding. Another important aspect to consider is that traditional peer-review gives very few incentives (or none at all) to the reviewers, who are not credited for the considerable amount of time and energy they spend in performing manuscript reviews.

Another flaw of the current peer-review system seems to be associated with the number of retractions of articles that journals announce every year. In 2014 and 2015, Springer and IEEE retracted over 100 published fraudulent articles from several journals [42,43]. Similarly, the Retraction Watch (<http://retractionwatch.com>) reports on these issues in other journals. Although it is not easy to evaluate the amount of published scientific papers containing incorrect conclusions, the number of retractions may provide information on the problems associated with traditional peer-

review. In 2012, Grieneisen and Zhang surveyed 42 of the largest bibliographic databases for major scholarly fields and publisher websites [44]. They found that the number of retractions has increased considerably after 2001. Retractions happen more often in fields such as medicine, life sciences and chemistry than in fields such as mathematics, physics, engineering and the social sciences. According to the study, the main cause of retraction is publishing misconduct (such as plagiarism and authorship or copyright issues), followed by incorrect use of data or incorrect data interpretation, and research misconduct (*e.g.*, the use of fraudulent or fabricated data).

To address the abovementioned issues, open peer-review models are emerging, in many cases to complement traditional models. For example, BioMed Central's GigaScience, all the journals in BioMed Central's medical series, and the journal F1000Research all publish reviewer reports, either as part of the pre-publication review process, or subsequent to publication. This last case is referred to as open post-publication peer-review: after a first editorial quality check, submitted manuscripts are published online, peer-review is then carried out openly (reports and names are published alongside the article), and the authors are invited to publish a revised version of the article, together with their response to the reviewers (see *Figure 3*).

Another form of open review comes from comments on blogs or third party sites, independent of any formal peer-review that may have already occurred on the article. Amongst other platforms, PubMed Commons was launched in 2013 as an initiative to enable signed post-publication commenting on articles indexed by PubMed. It is worth noting that this platform is not related to any specific journal or publisher, and as such constitutes a forum for public scientific discourse.

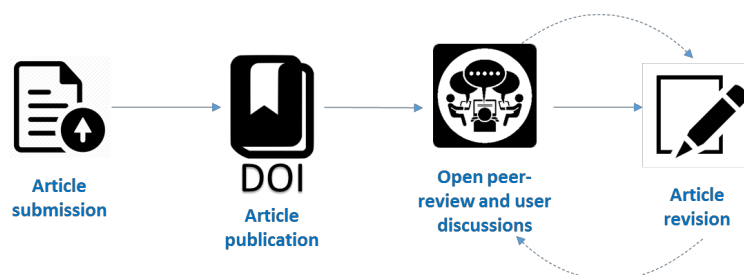


Figure 3: A schematic overview of a post-publication open-review model.

Importantly, studies have shown that open peer-review can produce reviews of higher quality, with better verified claims, and more constructive criticisms, when compared to closed review [21,45]. Of course, one should keep in mind that open and transparent peer-review does not come without risks: especially young, early-career researchers might fear that by signing critical and thorough reviews they could become a target for retaliation at a sensitive point in their career. In this sense, the traditional closed process provides, in theory, a sort of protection for the reviewer.

Table 3 lists publishing platforms and journals with an open peer-review policy, either as part of a pre- or a post-publication process.

Table 3: A list of platforms and journals (in alphabetical order) that support an open peer-review policy.

Platform/Journal	Open peer-review type	Website	Comments
Copernicus	post-publication	http://publications.copernicus.org	manuscripts are first published as discussion papers, then undergo an interactive public discussion, and are wherefore revised and published

F1000Research	post-publication	http://f1000research.com	referees are selected and invited, and their reports and names are published alongside the article, together with the authors' responses and comments from registered users
GigaScience, BioMed Central¹¹	pre-publication	http://gigascience.biomedcentral.com	anonymous peer review is <i>not</i> an option; final reviewer reports are online, distributed under a CC-BY license
Nature Communications, NPG	pre-publication	www.nature.com/ncomms/	as of January 2016 [46], an opt-out open review is available: authors can have the review history published along with their manuscript, unless they ask not to
PeerJ	pre-publication	https://peerj.com	optional open peer-review: referees are encouraged, but not required, to disclose their names; an <i>all-or-nothing</i> option then publishes the complete review history of the paper
Publons	pre- and post-publication	https://publons.com	reviewers can sign up and record their history of peer reviews, both before and after publication
PubMed Commons	post-publication	https://www.ncbi.nlm.nih.gov/pubmedcommons/	authors of publications in PubMed can post public comments on published papers
PubPeer	post-publication	https://pubpeer.com	a non-profit organization that allows authors to openly comment on published research papers
Royal Society Open Access	pre-publication	http://rsos.royalsocietypublishing.org	if both authors and referees agree on an open peer-review model, then signed referee reports are made public online; in-between scenarios are also possible

Miscellaneous Resources for Open Science

This section provides a list of general resources for open science, from e-learning platforms, over conferences and events, to open science coalitions and opportunities for early-career researchers' fellowships.

1. **CO**nnecting **RE**positories (CORE) - <https://core.ac.uk>

¹¹ many more BioMed Central journals implement an open peer-review model, for a full list see: <https://www.biomedcentral.com/journals>

CORE is the largest search engine of open access research outputs. At the time of writing, CORE indexes 861 repositories and contains more than 26 million records.

2. **Open Knowledge Foundation** - <https://okfn.org>
Open Knowledge International is a worldwide, non-profit network of people passionate about openness. This network uses advocacy, technology, and training to unlock information, and enables people to work with the network to create and share knowledge. Researchers can get involved through chapters and local groups: <https://okfn.org/network/>.
3. **Right to Research Coalition (R2RC)** - <http://www.righttoresearch.org>
The R2RC is a student and early career researcher organization that aims to promote open access, based on the belief that no student should be denied access to the articles they need because their institution cannot afford the often (too) high cost of subscription. Amongst other resources, a database of speakers on open access, open data, and open education is maintained by R2RC (see <http://www.righttoresearch.org/resources/Speakersdatabase/index.shtml>), which constitutes a useful resource in case researchers would like to invite speakers at their institutions to hear more about specific aspects of open science.
4. **OpenCon** - <http://www.opencon2016.org>
OpenCon is an annual conference for students and early career researchers who are interested in the promotion of open access, open data, and open educational resources. The conference offers scholarship opportunities for applicants, and a live stream for people not attending. Students can join the coalition to be educated about open access and to promote open access in their institution or research field.
5. **Facilitate Open Science Training for European Research (FOSTER)** - <https://www.fosteropenscience.eu>
The FOSTER project maintains an e-learning platform that brings together a variety of training resources for those who need to know more about open science, or who need to develop strategies and skills for implementing open science practices in their daily workflows. A nice overview of available resources is given through the open science taxonomy, as shown in *Figure 4*.
6. **The Future of Research Communication and e-Scholarship (FORCE11)** - <https://www.force11.org>
FORCE11 is a community of scholars, librarians, archivists, publishers, and research funders that has grown to bring a change in modern scholarly communications through the effective use of information technology.
7. **Research Data Alliance (RDA)** - <https://rd-alliance.org>
The RDA promotes the development and adoption of infrastructure for data sharing and data-driven research, in order to accelerate the growth of a cohesive data community that integrates contributors across domain, research, national, geographic, and generational boundaries. RDA Plenaries are held every six months in different places around the world. Interested early career researchers can apply for a scholarship to support their participation (both for USA and Europe: <https://rd-alliance.org/get-involved/studentearly-career-programms>).
8. **Mozilla Science Lab** - <https://www.mozillascience.org>
The Mozilla Science Lab facilitates learning about open source and open data, and furthermore offers fellowships for early-career researchers.
In particular, the Mozilla Fellowship for Science enables early-career researchers to spend ten months to work on open, web-enabled research and to further open science as mentors within the community (see <https://www.mozillascience.org/fellows>).
9. **Center for Open Science** - <https://cos.io>

The COS is a non-profit technology company that provides free and open services to increase inclusivity and transparency of research. Amongst the interesting resources are the “Transparency and Openness Promotion” (TOP) Guidelines [13] (<https://cos.io/top/>), and a curated list of public open science projects (<https://osf.io/explore/activity/#newPublicProjects>).

10. **Opening Science** - <http://www.openingscience.org>

A project for the free sharing of open science resources, Opening Science has published the Open Book: “Opening Science – The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing” (<http://www.openingscience.org/get-the-book/>).

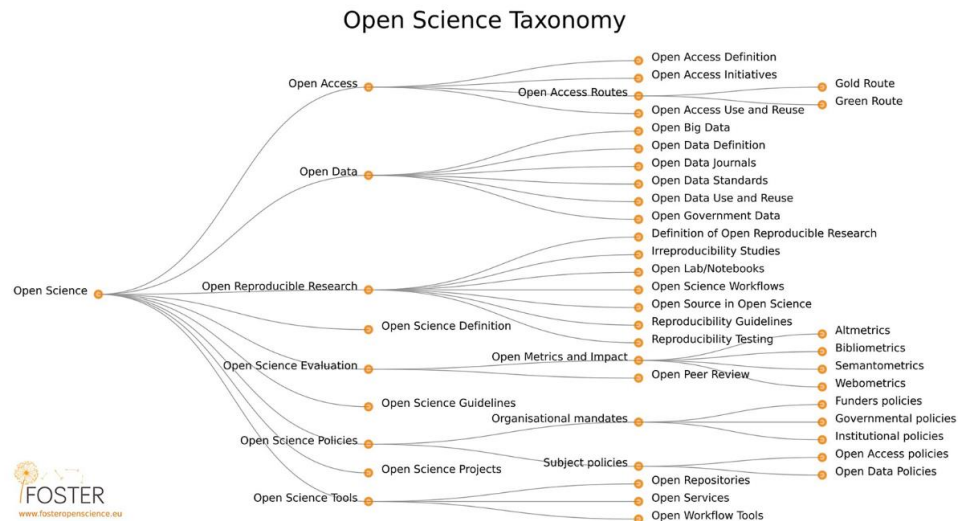


Figure 4: The FOSTER Open Science taxonomy.

Image credit: Knoth, Petr; Pontika, Nancy (2015): Open Science Taxonomy. figshare. - <https://dx.doi.org/10.6084/m9.figshare.1508606.v3> - Retrieved: 12 11, May 19, 2016 (GMT)

What Can You Do?

If you are willing to engage in open science, there are a number of practices and resolutions that can be adopted without too much effort; **Box 1** lists the key ones.

Box 1: Practices and resolutions to adopt in order to engage in open science.

1. When possible, **use and cite existing public data**.
2. Whenever feasible, **share your research data** through trusted repositories. General-purpose repositories and domain-specific ones are available on the web. Make sure you share relevant **metadata** as well, as these are essential for data interpretation and reproduction.
3. If you use software code as part of your research cycle, **release the code** and the environment needed to run it. Specify the open source license you intend to use, and link the readers to a stable repository that hosts the code.
4. Post **free copies of your research articles** online. The majority of journals allow researchers to do so, sometimes after an embargo period of 6-12 months.
5. Post **preprints** of your research manuscripts online, ideally at the same time of official submission to a journal.
6. When possible, choose an **open access journal** as venue for your scientific articles. Keep in mind that subscription journals also offer an open access solution, upon payment of extra fees.

Conclusions

The next scientific revolution is underway. Modern science is undergoing profound structural changes enabled by the advent of digital technology and communications, and these shifts are occurring on multiple levels of the scientific process at once. If we want to speed up scientific progress, we must engage in open science practices, and make our research output freely available to the scientific community, and to the public at large.

However, scientists are inherently quite conservative in their adoption of new approaches. Novel methods often struggle to be accepted until their superiority is confirmed, and found overwhelming. As a result, a wide community of researchers is currently awaiting evidence-based benefits of open science practices before adopting them. From an optimistic viewpoint, this situation provides a perfect occasion for individuals to show initiative and take immediate action, potentially yielding a first-mover advantage. At the same time, adherence to open science often relies on the complete support of colleagues, supervisors, research leaders, and host institutions, especially for early-career researchers. In this respect, training academics early in their career is crucial: graduate programs should incorporate open science into their existing curricula. A key topic to be included in such curricula is training on publishing practices, such as author rights, appropriate citation practices, and open access publishing. Institutions and funding agencies could together

provide skills training on data and code deposition, self-archiving of articles, and modern scientific computing, and could moreover consider mandates and policy requirements for open science practices. With appropriate training and support, early-career researchers will thus be able to pursue open science to the point that it becomes the default *modus operandi* for all academic research.

This paper has presented an inventory of resources and practical tips to conduct science in the open. Of course, the availability of resources in the scientific community is essential, but not sufficient: scientists' commitment is crucial, both at the individual and at the collective level. Only with commitment and wide participation will we be able to unleash the potential of open research practices, and reap the profound benefits of the increased scientific progress that can be brought about by open collaboration and ready exchange of ideas and data between and beyond disciplines and sectors.

Competing interests

PM is a Research Data Alliance early career fellow, an OpenCon alumna, and is funded through an EC H2020 project that aims to create an open data exchange ecosystem.

Acknowledgments

The authors acknowledge funding from the European Union's Horizon 2020 Programme under Grant Agreement 634107 (PHC32-2014).

References

1. Nielsen M. An informal definition of OpenScience | The OpenScience Project [Internet]. [cited 2016 May 17]. Available from: <http://www.openscience.org/blog/?p=454>
2. Watson M. When will “open science” become simply “science”? *Genome Biol.* 2015;16:101.
3. Nielsen M. *Reinventing Discovery: The New Era of Networked Science*. Princeton, N.J: Princeton University Press; 2011.
4. Hannay T. A new kind of science? *Nat. Phys.* 2011;7:742–742.
5. Khabsa M, Giles CL. The Number of Scholarly Documents on the Public Web. *PLOS ONE*. 2014;9:e93949.
6. Womack RP. Research Data in Core Journals in Biology, Chemistry, Mathematics, and Physics. *PLOS ONE*. 2015;10:e0143460.
7. Begley CG, Ioannidis JPA. Reproducibility in science: improving the standard for basic and preclinical research. *Circ. Res.* 2015;116:116–26.
8. Ince DC, Hatton L, Graham-Cumming J. The case for open computer programs. *Nature*. 2012;482:485–8.
9. Björk B-C, Solomon D. The publishing delay in scholarly peer-reviewed journals. *J. Informetr.* 2013;7:914–23.
10. Molloy JC. The Open Knowledge Foundation: Open Data Means Better Science. *PLoS Biol.* 2011;9:e1001195.
11. Wolkovich EM, Regetz J, O’Connor MI. Advances in global change research require open science by individual researchers. *Glob. Change Biol.* 2012;18:2102–10.
12. Opening Science – The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing | openingscience.org [Internet]. [cited 2016 May 20]. Available from: <http://www.openingscience.org/get-the-book/>
13. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. *Science*. 2015;348:1422–5.
14. McKiernan E, Bourne PE, Brown CT, Buck S, Kenall A, Lin J, et al. The benefits of open research: How sharing can help researchers succeed [Internet]. 2016 [cited 2016 May 14]. Available from: https://figshare.com/articles/The_open_research_value_proposition_How_sharing_can_help_researchers_succeed/1619902
15. Tracz V, Lawrence R. Towards an open science publishing platform. *F1000Research*. 2016;5:130.
16. Tennant JP, Waldner F, Jacques DC, Masuzzo P, Collister LB, Hartgerink CHJ. The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000Research*. 2016;5:632.
17. Gorgolewski KJ, Poldrack R. A practical guide for improving transparency and reproducibility in neuroimaging research. *bioRxiv*. 2016;039354.
18. Big Data, Little Data, No Data [Internet]. MIT Press. [cited 2016 May 18]. Available from: <https://mitpress.mit.edu/big-data>
19. Piwowar HA, Day RS, Fridsma DB. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE*. 2007;2:e308.
20. Piwowar HA, Vision TJ. Data reuse and the open data citation advantage. *PeerJ*. 2013;1:e175.
21. Wicherts JM, Bakker M, Molenaar D. Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLoS ONE*. 2011;6:e26828.

22. Kidwell MC, Lazarević LB, Baranski E, Hardwicke TE, Piechowski S, Falkenberg L-S, et al. Badges to Acknowledge Open Practices: A Simple, Low-Cost, Effective Method for Increasing Transparency. *PLOS Biol.* 2016;14:e1002456.
23. Gray J, Szalay AS, Thakar AR, Stoughton C, vandenBerg J. Online Scientific Data Curation, Publication, and Archiving. *arXiv:cs/0208012*. 2002;103–7.
24. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data.* 2016;3:160018.
25. Boettiger C. An introduction to Docker for reproducible research, with examples from the R environment. *ACM SIGOPS Oper. Syst. Rev.* 2015;49:71–9.
26. Cito J, Ferme V, Gall HC. Using Docker Containers to Improve Reproducibility in Software and Web Engineering Research. In: Bozzon A, Cudre-Maroux P, Pautasso C, editors. *Web Eng.* [Internet]. Springer International Publishing; 2016 [cited 2016 Jul 13]. p. 609–12. Available from: http://link.springer.com/chapter/10.1007/978-3-319-38791-8_58
27. Barnes N. Publish your computer code: it is good enough. *Nature.* 2010;467:753.
28. Blischak JD, Davenport ER, Wilson G. A Quick Introduction to Version Control with Git and GitHub. *PLOS Comput Biol.* 2016;12:e1004668.
29. Ram K. Git can facilitate greater reproducibility and increased transparency in science. *Source Code Biol. Med.* 2013;8:7.
30. Knuth DE. *Literate Programming*. First Edition edition. Stanford, Calif.: Center for the Study of Language and Inf; 1992.
31. Harnad S, Brody T. Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *-Lib Mag.* [Internet]. 2004 [cited 2016 May 21];10. Available from: <http://eprints.soton.ac.uk/260207/>
32. Gwinn M. The Access Principle: The Case for Open Access to Research and Scholarship. *Emerg. Infect. Dis.* 2006;12:1473.
33. Wang X, Liu C, Mao W, Fang Z. The open access advantage considering citation, article usage and social media attention. *Scientometrics.* 2015;103:555–64.
34. Brenner S. Loose end. *Curr. Biol.* 1995;5:568.
35. Brown C. The E-evolution of preprints in the scholarly communication of physicists and astronomers. *J. Am. Soc. Inf. Sci. Technol.* 2001;52:187–200.
36. Lariviere V, Sugimoto CR, Macaluso B, Mилоjević S, Cronin B, Thelwall M. arXiv e-prints and the journal of record: An analysis of roles and relationships. *ArXiv13063261 Cs* [Internet]. 2013 [cited 2016 May 23]; Available from: <http://arxiv.org/abs/1306.3261>
37. Berg JM, Bhalla N, Bourne PE, Chalfie M, Drubin DG, Fraser JS, et al. Preprints for the life sciences. *Science.* 2016;352:899–901.
38. McCook A. Is peer review broken? Submissions are up, reviewers are overtaxed, and authors are lodging complaint after complaint about the process at top-tier journals. What’s wrong with peer review? *The Scientist.* 2006;20:26–35.
39. Schwartz SJ, Zamboanga BL. The Peer-Review and Editorial System: Ways to Fix Something That Might Be Broken. *Perspect. Psychol. Sci.* 2009;4:54–61.
40. Berquist TH. Peer Review: Is the Process Broken? *Am. J. Roentgenol.* 2012;199:243–243.
41. Missimer T. Journal Paper Peer-Review: A Broken System? *Groundwater.* 2015;53:347–347.
42. Van Noorden R. Publishers withdraw more than 120 gibberish papers. *Nature* [Internet]. 2014 [cited 2016 May 23]; Available from: <http://www.nature.com/doi/10.1038/nature.2014.14763>
43. Retraction of articles from Springer journals [Internet]. www.springer.com. [cited 2016 May 23]. Available from: <http://www.springer.com/gp/about-springer/media/statements/retraction-of-articles-from-springer-journals/735218>

44. Grieneisen ML, Zhang M. A Comprehensive Survey of Retracted Articles from the Scholarly Literature. PLOS ONE. 2012;7:e44118.
45. Walsh E, Rooney M, Appleby L, Wilkinson G. Open peer review: a randomised controlled trial. Br. J. Psychiatry J. Ment. Sci. 2000;176:47–51.
46. Transparent peer review at Nature Communications. Nat. Commun. 2015;6:10277.