Close connections between open science and open-source software

Youhua Chen1, Xueke Lu2, You-Fang Chen3
1, Department of Renewable Resources, University of Alberta, Edmonton, T6G 2H1, Canada, Email: youhuach@gmail.com
2, Department of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS, UK
3, School of Software, Harbin Normal University, Harbin, Heilongjiang Province, China

ABSTRACT

Open science is increasingly gaining attention in recent years. In this mini-review, we briefly discuss and summarize the reasons of introducing open science into academic publications for scientists. We argue that open-source software (like R and Python software) can be the universal and important platforms for doing open science because of their appealing features: open source, easy-reading document, commonly used in various scientific disciplines like statistics, chemistry and biology. At last, the challenges and future perspectives of performing open science are discussed.

KEYWORDS: open source, open access, data reuse, reproductivity, knowledge integration

INTRODUCTION

In this cyber era, we have access to an unprecedented amount of data and information online every day (Reichman et al. 2011). Scientists have to learn and absorb new knowledge in an effective way so as to keep their research advances not fallen behind others. Traditional ways, like joining seminars, conferences and/or workshops allow researchers from relative disciplines to communicate in a face-to-face way and facilitate collaborations. However, these traditional research routines would become less effective when open-access journals is becoming much more prevalent. At current time, there are many good-reputation open-access publishers, for instance, Biomed Central (http://www.biomedcentral.com), PeerJ (https://peerj.com) and Public Library of Science (http://www.plos.org).

However, even though the main text and supplemental materials from open-access journals can be well presented, there is still much scientific information hidden behind a paper: for example, some technical backgrounds, programming codes or experimental skills. As such, it would be of great help to fellows to follow and learn better from a paper if all these things are fully open, in addition to the open text. Such a kind of demands becomes an emerging issue in contemporary scientific researches and asks for the development of open science.

About open science

As similar above, one general definition of open science might be that all the contents, methods, techniques and data are not hidden from readers but instead, they are well organized and available for the readers to understand, reproduce and reuse. Thus, no scientific privacies are allowed for doing open science. The concept of open science has been proposed in the last several years and has attracted increasing attention in the field

of ecology and evolution (Reichman et al. 2011, Gilbert et al. 2012, Wolkovich et al. 2012). There are growing debates and discussion about its feasibility and applicability.

Why should we do open science?

In conventional research activities, scientists can not sufficiently deliver some important tricks, patents and methods to the readers in their publications due to limited access and publication restriction. One direct consequence of the incomplete information is that readers and other scientists cannot easily follow, re-examine or be inspired by the results of the papers. As a consequence, the associated research themes become dominated by these scientists as they uniquely hold some important tricks. Moreover, since the details are not fully open to public, researchers tend to not trust the results showed in these papers.

The benefits and reasons for doing open science by making the associated original data, experimental videos, and programming scripts become available for the public can be multifaceted. Here, before discussing relevant advantages of doing open science, we define a fully open paper as the one with open data, programs, analytical scripts and all other materials in the text.

First, it is a great chance for fellow researchers to reuse, reexamine and produce new insights from fully open papers (Duke and Porter 2013, Piwowar and Vision 2013, Vision and Piwowar 2013). The direct benefit is that people will cite the open papers (Piwowar et al. 2007, Calver and Breadley 2010). For most scientists, citation is the most important indicator of his/her academic influence and importance in the relevant research field (Garfield 1970, Bornmann and Daniel 2005, Hirsch 2005, Editorial 2008).

Second, it facilitates benign research cooperation and competition, thus accelerates research progresses (Woelfle et al. 2011). Because researchers can easily verify and modify the associated documents to fulfill their own research goals under the framework of open science, the competition among the researchers can be very benign and beneficial among themselves. Moreover, if any researchers in this open-science game feel inferior, they can easily find advanced colleagues to work synergistically. Thus, open science can promote win-win achievements for competing and collaborating scientists, simulate global participation and share of knowledge, and reduce barriers of knowledge dissemination (Evans and Reimer 2009). For example, in the case of open source software like R, different people can work on the same R code and use the codes from others to make up their own R packages. Different researchers can communicate much easier because the codes are totally open. Communication can be much harder when the software is commercial and not open, customers can only rely on limited and sometimes not-free technical helps. As consequences, benign research cooperation is hard between two researchers if any of them doesn't have this commercial software.

Third, open science allows researchers from different countries and ages to work together (Evans and Reimer 2009). Like open source codes, people can work on them to address their own specific questions. This is extremely necessary and handful for junior or developing-country researchers as they don't have many sources and funds to support and develop their studies. Open science definitely can be very helpful to assist the growth of young scientists by improving their logics and skills in the researches by looking at the open documents associated with the open-science papers.

Fourth, open science is not conceptual but very practical at the current time. As seen, programming codes, original data, and experimental videos are increasingly documented in some online databases or as the supplemental materials, for example, most journals (like Nature, Science and PloS journals) have supplementary material sections for the authors of the paper to release raw data or show the deduction of mathematical equations. The Dryad online depository (http://datadryad.org/) can allow authors to deposit the data and figures and some journals are now asking authors to submit their data to Dryad depository. Nowadays, sharing research data is a vital step in scientific activities (White et al. 2013). Moreover, many journals have the policies to require authors to make their data become available for the public (Duke and Porter 2013, Vines et al. 2013). Moreover, many universities increasingly support the publication of open-source paper and some university libraries (e.g. Cornell University) have established preprint library (http://arxiv.org/). These efforts are actually important steps for open science. However, for publishing fully open papers, some more things are mandated to do maybe. Authors have to provide detailed deduction of the formulas and provide self-explanatory programming codes for simulating and calculating their results. They are encouraged to inform the readers how they can obtain the results (including tables and figures), not only their results and implications. Through these open initiatives, different researchers can reduce their knowledge gaps greatly and thus improve research novelty, rigorousness and efficiency.

Open-source software for doing open science

There are a suit of open source software (http://www.opensource.org), and we will focus on two examples: Python (Van Rossum 1995, Python Software Foundation 2013) and R software (R Development Core Team 2011). In specific, R software has many appealing features for allowing researchers to do open science. First, it is totally open and free. The base of the software and all the affiliated packages are free and can be re-distributed. Researchers can redistribute and reuse these packages to produce their own packages or programs. The only requirement it should obey is that these new packages should be open source and follow a uniform standard: GNU General Public License (GPL) (https://gnu.org/licenses/gpl.html).

There are some ongoing activities to support open-source software to perform open science. For example, GitHub (https://github.com/) is proved to be a good repository of restoring source codes for biological researches (Ram 2013). There are a lot of R source codes which are still under development by the authors placed in GitHub website. Recently, rOpenSci (http://ropensci.org/) is released as a new in-progress collaborative project aiming to effectively and synergistically connect various online databases through R computing platform.

Python is also widely used in computational biology (Bassi 2007). Biopython is a collection of useful tools for performing bioinformatics and computational biology analyses under the Python environment (http://biopython.org). Similar to Python, as one script language, the running speed of Python is also relatively slow (similar as R). However, the release of PyPy (http://pypy.org) can help solve the computational slowness issue. PyPy is a fast and compliant implementation of Python language. Many Python programs can run using PyPy to speed up the computational time without modifications. Nowadays, it has growing voice to make PyPy to be compatible to the well-known

Python package: Numpy (one of the fundamental packages in scientific computing under Python environment). In the near future, Python has the promise to be one of the platforms for performing open science.

The challenges for doing open science

It requires some time for most of the scientists to gradually recognize open science, just like open-access journals and sharing their original data (Tenopir et al. 2011). Many influential researchers tend to avoid publishing their papers in open-access journals, many of which are usually in low publishing quality. Then, it can be forecasted that these scientists might not like to research mode exhibited by open science.

The reasons for that many scientists prefer conventional research behaviors can be numerous too. As mentioned a bit above, some important techniques, tricks and programs have commercial values. Therefore, researchers and institutes can gain profits from these things and of course, they are kept confident from other colleagues and the pubic. At another hand, these things can be valuable for their owners to continue publishing and dominating a specific research field. Consequently, they don't want others to get access these things other than themselves.

Nowadays, no single scientific field can be developed without the assistance of other related disciplines. Thus, it is an unavoidable process to do multidisciplinary researches to address questions for different research fields using similar scientific philosophy and technologies. For example, researches in biological science have to use electronic devices built by physical sciences and engineering to collect experimental data and post-experiment data analyses heavily rely on statistical science.

Finally, there are many field-specific vocabularies and jargons in different scientific fields. For performing open science under the multidisciplinary framework, these specialized vocabularies and jargons should be unified so as to promote the dissemination of open science.

Future perspectives

We believe that, open science is an unpreventable trend for future research since it offers an ultimate solution to minimize the time lag to distribute the research advances among countries and researchers for the abovementioned reasons. By opening every aspect of a paper to the broad audience, it can promote benign research competition and cooperation, provide more chances for young and developing-country researchers, and allow junior scientists to grow in a fast and effective way.

In the coming future, open science should be more emphasized so as to simulate citizens to engage into scientific activities better. Citizen science (Irwin 1995, Silvertown 2009, Hand 2010) is now recognized and appreciated in recent years too. Thus, it might be an important but open challenge to effectively combine citizen and open science together so as to promote scientific innovation and accelerate knowledge dissemination. We design an ambitious goal for open-citizen science in 21st century: every person can be a amateur scientist!

Conclusions

Information era allows scientists to finish data-mining-related researches within a quick time. The close connection between open-source software, open-access journals,

open data and open science can better facilitate the dissemination of scientific discoveries and fancy results. In the coming future, along with the development of citizen science and open-access journals, it is expected that open science should have better prospects.

Conflict of interest

The authors declare they have no competing interests.

ACKNOWLEDGEMENTS

This work is supported by China Scholarship Council (to Youhua Chen and Xueke Lu). We thank Jian Zhang for providing comments and language correction.

REFERENCES

Bassi, S. 2007. A primer on python for life science researchers. PLoS computational biology 3:e199.

Bornmann, L., and H. Daniel. 2005. Does the h-index for ranking of scientistis really work? Scientometrics 65:391–392.

Calver, M., and J. Breadley. 2010. Patterns of citations of open access and non-open access conservation biology journal papers and book chapters. Conservation Biology 24:872–880.

Duke, C., and J. Porter. 2013. The ethics of data and reuse in biology. BioScience 63:483–489.

Editorial. 2008. The importance of being cited. Nature Geoscience 1:563.

Evans, J., and J. Reimer. 2009. Open access and global participation in science. Science 323:1025.

Garfield, E. 1970. Citation indexing for studying science. Nature 227:669–671.

Gilbert, K., R. Andrew, D. Bock, M. Franklin, N. Kane, J. Moore, B. Moyers, S. Renaut, D. Rennison, T. Veen, and T. Vines. 2012. Recommendations for utilizing and reporting population genetic anlyses: the reproducibility of genetic clustering using the program STRUCTURE. Molecular Ecology 21:4925–4930.

Hand, E. 2010. Citizen science: people power. Nature 466:685–687.

Hay, S., D. George, C. Moyes, and J. Brownstain. 2013. Big data opportunities for global infectious disease surveillance. Plos Medicine 10:e1001413.

Hirsch, J. 2005. An index to quantify an individual's scientific research output. PNAS 102:16569–16572.

Irwin, A. 1995. Citizen Science: A Study of People, Expertise and Sustainable Development. . Routledge, New York.

Piwowar, H., R. Day, and D. Fridsma. 2007. Sharing detailed research data is associated with increased citation rate. PLoS ONE 2:e308.

Piwowar, H., and T. Vision. 2013. Data reuse and the open data citation advantage. PeerJ PrePrints 1:e1.

Python Software Foundation. 2013. Python Language Reference, version 2.7. Available at http://www.python.org.

R Development Core Team. 2011. R: A Language and Environment for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. R Foundation for Statistical Computing, Vienna, Austria.

Ram, K. 2013. Git can facilitate greater reproducibility and icnreased transparency in science. Source Code for Biology and Medicine 8:7.

Reichman, O., M. Jones, and M. Schildhauer. 2011. Challenges and opportunities of open data in ecology. Science 331:703–705.

Van Rossum, G. 1995. Python tutorial, Technical report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam.

Silvertown, J. 2009. A new dawn for citizen science. Trends in Ecology and Evolution 24:467–471.

Tenopir, C., S. Allard, K. Douglass, A. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. 2011. Data sharing by scientists: practices and perceptions. PLoS ONE 6:e21101.

Vines, T., R. Andrew, D. Bock, M. Franklin, K. Gilbert, N. Kane, J. Mooer, B. Moyers, S. Renaut, D. Rennison, T. Veen, and S. Yeaman. 2013. Mandated data archiving grately improves access to research data. FASEB Journal:doi: 10.1096/fj.12–218164 fj.12–218164.

Vision, T., and H. Piwowar. 2013. Data reuse and scholarly reward: understanding practice and building infrastructure. PeerJ PrePrints 1:e14v1.

White, E., E. Baldridge, Z. Brym, K. Locey, D. McClinn, and S. Supp. 2013. Nine simple ways to make it easier to (re)use your data. PeerJ PrePrints 1:e7.

Woelfle, M., P. Olliaro, and M. Todd. 2011. Open science is a research accelerator. Nature Chemistry 3:745–748.

250 Wolkovich, E., J. Regetz, and M. O'Connor. 2012. Advances in global change research
251 require open science by individual researchers. Global Change Biology 18:2102–
252 2110.

253