# Digital Open Science – Teaching digital tools for reproducible and transparent research

Ulf Toelch[1,2] and Dirk Ostwald[3,4,5]

1: QUEST Center, Berlin Institute of Health, Berlin, Germany

2: Biological Psychology und Cognitive Neuroscience, Freie Universität Berlin, Berlin, Germany

3: Computational Cognitive Neuroscience, Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany

4: Center for Cognitive Neuroscience Berlin, Freie Universität Berlin, Berlin, Germany

5: Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

## The need for open science education.

In the past years, a reproducibility crisis has shaken many scientific disciplines [1–3]. In our own field, psychology, up to 32% of results fail a replication test [4]. Even though a multifaceted problem, one of the core problems of published research is an incomplete representation of a scientific project [5]. For example, given the ever increasing complexity of quantitative scientific inquiry, traditional article methods and results sections are not sufficient to fully understand the whole analysis process from raw data to final figure in the publication [6]. As a response to these problems, journals [7], funding agencies [8,9], and policy makers have introduced more or less strict regulations how hypotheses, materials, data, and procedures should be made available. These new requirements have a direct impact particularly on the work of young researchers (from MSc level to postdoc). They have to ensure that their research practices are in line with these new requirements to be able to publish their work or apply for funding. Moreover, as we understand open science as an integral part of quality management, it is also relevant for young researchers aspiring a career in an industry context. As a result, new and established digital tools have evolved into a novel “toolbox” for digital open science. In our view, however, the diffusion of these innovations is somewhat hindered as these are seldom part of a curriculum in the life sciences (for notable exception see [10,11]). An EU report recently marked early career education in open science as highly desirable, but “training opportunities for open access and open data are not yet widely offered” [12]. For this purpose, we have developed and taught a course at late MSc, early PhD level that covers the necessary tools to ensure that young researchers embrace these techniques and meet the quality standards

of a reproducible and transparent research process [13]. We designed this course for students coming from disciplines that need to process digital data from experiments or simulations or transform analog data to a digital format and then apply computational methods to the data by means of (custom) software. Course participants were expected to have only minimal programming experience and the course was designed without restrictions to any particular programming language. In the following, we will outline the curriculum for an introduction to open science (approx. 60 hours with 15-20 hours lectures and tutorials). Additionally, we will highlight tools and techniques that are supplementary and could serve as material for more advanced classes. All materials for this course including a reading list and presentations are available as an OSF project (https://osf.io/X6892/).

**Course overview**

A scientific project (Table 1) entails many steps that are often only partly reported in a standard scientific publication. As of 2015, only 13 % of publications publish their raw data [14] and even fewer publish analysis code [15]. The reasons for this are numerous, such as unclear data protection issues, pending patents, or lack of technical know-how [16]. This course was designed to particularly eliminate the latter, i.e. technical barriers. To introduce students to the open science toolbox, we developed a narrative for the course that involved a planned replication of an existing study. Students needed to bring (and read) a paper to the course from their field of expertise with a single, ideally simple experiment that should be replicated. Importantly, our course consisted of two parts. The first part conveyed the theoretical background on open science and comprised an introduction to the available digital tools in interactive teaching sessions. The second part was dedicated to projects in which students applied these tools to a research project of their choice, and resulted in a symposium at which students presented their projects. In the following, we will briefly outline each session. A detailed description of the background and exercises can be found in the associated OSF project.

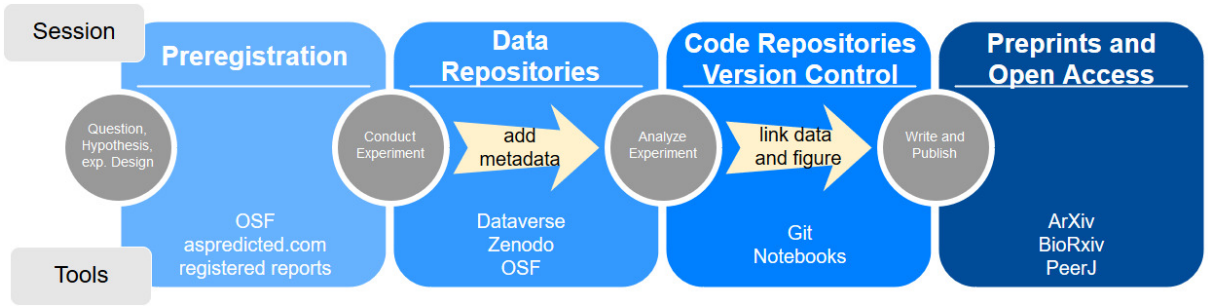


Figure 1. Outline of interactive teaching sessions and associated tools.

**Interactive Teaching Sessions**

*(1) Introduction.* In the first session, we recapitulated the steps involved in a scientific project including an introduction to the open science framework and an outline of the expected active contributions of the students.

*(2) Preregistration.* Preregistration is a summary of the research rationale, hypotheses with predictions, methods, and, in an extended version, also an analysis plan [17]. We introduced students to the minimum requirements for a preregistration and the difference between preregistrations that are either submitted directly to a journal and submissions to a suitable platform like OSF or aspredicted.com. We discussed how preregistration potentially prevents p-hacking [18] and HARKing [19], two major threats for the reproducibility of science. Here, we emphasized the difference between confirmatory and exploratory analysis and particularly why exploratory analysis is not per se forbidden, but can be a major driver of scientific discovery [20].

*(3) Data Repositories.* Once data has been collected, it has to be saved in a digital format in an appropriate repository. We highlighted different types of general repositories (OSF, Dateverse, Zenodo) vs. specialized repositories for a particular data structure (e.g. proteins [21] or MRI data [22]). Importantly, all repositories should ideally adhere to the "FAIR" guiding principles: Findability, Accessibility, Interoperability, and Reusability [23]. This includes unique digital object identifiers, but also metadata that allows humans and machines to understand the format of the data. Additionally, we also covered differences between licences for the shared data and legal and ethical issues that are often associated with data from human subjects.

*(4) Code repositories, version control, and notebooks.* Raw data is seldom reported in a paper, but rather derived statistics and figures of aggregated data. Annotated analysis scripts make the transition from raw data to figure transparent and can be shared via specialized code repositories. One major advantage of such repositories is a version control system that keeps track of changes to the code and allows for collaborative work on the scripts. In particular, we introduce students to Git [24] as a version control system that connects to online repositories like Github and Gitlab [25–27]. Importantly, we then integrated analysis code with annotating text into notebooks (livescript for MATLAB, Jupyter for Python, and RMarkdown for R) to illustrate how to write a fully reproducible analysis in a manuscript format.

*(5) Open Access.* The final step in the scientific process is the publication of the results (plus data, and steps that led to the results) in a journal. In this session, we give an introduction to the

publication process and explain differences in open access formats. We highlight different monetization schemes of publishers (pay for view vs. pay for publication). Here, university or funding agency specific regulations on who is paying for open access at an institution should be discussed. A short note on predatory open access journals should be made [28]. We cover the differences between preprint servers, green, and gold access and how they fit into the current publication landscape [29–31]. This session should enable young researchers to discuss which journals may be suitable with their supervisors under consideration of publications costs.

*(6) Chances and limitations of open science.* The last session was dedicated to the chances and limitations of open science and directly linked to the discussion on open access publishing. Students collected advantages of open science and arguments under which circumstances there are limitations to open science in small groups and presented them. Important topics included privacy concerns for patient and participant data and patents. Students further listed the fear of being scooped before their own publication and a reuse of data they collected without proper attribution. We collected these on a white board and used this for an engaging discussion. This turned into an individual guide: how to convince my supervisor that open science is a *win-win* situation.

**Project work**

In a last session, students presented their plans for an open science project that they would conduct over the next couple of weeks. Importantly, these projects were connected directly to current or past research projects in our department. The projects were roughly separable into two categories. Students either came up with a feasible idea for a small replication experiment that they conducted and analyzed. Others worked on an already collected (and potentially published) data set. Importantly, students actively used a selection of the aforementioned open science tools in these projects. Examples are data set projects in which students transformed already existing neuroimaging data sets into a novel community standard for data sharing (the Brain Imaging Data Structure (BIDS, [32]) and prepared data set papers [33] . Another project involved an extended replication of an experiment and making resulting data for this available on OSF (osf.io/jaxdp). Yet another project extended an analysis on an existing data set from a bachelor thesis and used Git as a version control system for programming an analysis on participants' choice data which is now developed into a master thesis. All these projects enabled students to use open science tools in a scientific context and integrate open science practices in their workflows. This setup is, in our opinion, feasible in almost every environment as all labs have unpublished data sets, undocumented analyses, or are interested in a small replication

study. This directly connects teaching to the scientific process enabling students to contribute to knowledge creation in a meaningful way and hence increases motivation [34–36]. We concluded the course by a colloquium where students gave short presentations on their projects and reflected on challenges they encountered during the project.

**Concluding remarks**

The course described here offers a first introduction to open science for early career researchers and we hope that ideas and parts from this course will find their way in many higher education curricula. We acknowledge that starting new open science practices will initially increase the work and effort needed to complete a full research cycle from question to publication. This may put a burden on young researcher who already struggle with the other obstacles in their new career. It is in our view thus important that supervisors and PIs are supportive and allow for additional time to implement these practices. We want to stress that this will result in a *win win* situation as open science practices are an important part of quality management in science. PhD students, for example, often leave the lab after obtaining their degree. If a PhD student, however, established a standardized way to analyze data from an experiment, the whole lab will benefit from a code repository that contains a commented and reproducible version of this analysis technique.

*Challenges and improvements.* Most of our course focusses on the technical side of open science and familiarization with tools that are already in use for example in software development. Future extension could include already existing repositories for open materials like technical devices. For example, repositories complete plans how to build devices are already exist (e.g. openbehavior.com). In our view, it is also important to link these techniques to already existing courses on philosophy of science and statistical inference to integrate them seamlessly in the scientific process. A subsequent evaluation of the course by the students (conducted by the education quality assessment team at Freie Universität) revealed that students particularly liked the project work. Critical points involved the low number of sessions which sometimes led to only superficial or too rapid coverage of material. More time particularly in the version control notebook sessions would thus be beneficial.

Future courses need to keep up with the changes in the field. In our view, major milestones will be the development of a reproducible document format that will allow for the seamless integration of multiple programming languages into one coherent manuscript file that will retain the ease of annotation and commenting functions in current word processors. Another exciting new development is Docker, a software container platform. Docker allows for

the creation of computing environments that contain all dependencies and a snapshot of the software versioned at the date of analysis. This will potentially prevent backwards compatibility and *runs on my machine* problems.

We have not established open science practices rigorously if at all in all of our own past research projects. Rigor and highest quality standards for knowledge creation, however, mandate a change in our thinking. This course, hopefully, provides some material to teach young researchers the necessary skills to implement open science practices and increase the sustainability and quality of modern science even beyond the prevailing high standards.

**Acknowledgements**

We thank Christina Riesenweber for discussions on open access and providing slides on open access. This project was funded by the Freie Universität Berlin through a teaching award.

**References**

1. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging? Ann N Y Acad Sci. 2010;1191: 133–155. doi:10.1111/j.1749-6632.2010.05446.x

2. Camerer CF, Dreber A, Forsell E, Ho T-H, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016; aaf0918. doi:10.1126/science.aaf0918

3. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov. 2011;10: 712–712. doi:10.1038/nrd3439-c1

4. Collaboration OS. Estimating the reproducibility of psychological science. Science. 2015;349: aac4716. doi:10.1126/science.aac4716

5. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Sert NP du, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1: s41562-016-0021-016. doi:10.1038/s41562-016-0021

6. Roche DG, Kruuk LEB, Lanfear R, Binning SA. Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLOS Biol. 2015;13: e1002295. doi:10.1371/journal.pbio.1002295

7. Stodden V, Guo P, Ma Z. Toward Reproducible Computational Research: An Empirical Analysis of Data and Code Policy Adoption by Journals. PLOS ONE. 2013;8: e67111. doi:10.1371/journal.pone.0067111

8. Baker M. Dutch agency launches first grants programme dedicated to replication. Nat News. doi:10.1038/nature.2016.20287

9. Collins FS, Tabak LA. NIH plans to enhance reproducibility. Nature. 2014;505: 612–613.

10. Schmidt B, Orth A, Franck G, Kuchma I, Knoth P, Carvalho J. Stepping up Open Science Training for European Research. Publications. 2016;4: 16. doi:10.3390/publications4020016

11. Teal TK, Cranston KA, Lapp H, White E, Wilson G, Ram K, et al. Data Carpentry: Workshops to Increase Data Literacy for Researchers. Int J Digit Curation. 2015;10: 135–143. doi:10.2218/ijdc.v10i1.351

12. Providing researchers with the skills and competencies they need to practise Open Science - Report of the Working Group on Education and Skills under Open Science [Internet]. [cited 19 Oct 2017]. Available: https://www.rri-tools.eu/-/providing-researchers-with-the-skills-and-competencies-they-need-to-practise-open-science-report-of-the-working-group-on-education-and-skills-under-op

13. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348: 1422–1425. doi:10.1126/science.aab2374

14. Womack RP. Research Data in Core Journals in Biology, Chemistry, Mathematics, and Physics. PLOS ONE. 2015;10: e0143460. doi:10.1371/journal.pone.0143460

15. Morin A, Urban J, Adams PD, Foster I, Sali A, Baker D, et al. Shining Light into Black Boxes. Science. 2012;336: 159–160. doi:10.1126/science.1218263

16. Howe A, Howe M, Kaleita AL, Raman DR. Imagining tomorrow’s university in an era of open science. F1000Research. 2017;6. doi:10.12688/f1000research.11232.2

17. Nosek BA, Lakens D. Registered Reports. Soc Psychol. 2014;45: 137–141. doi:10.1027/1864-9335/a000192

18. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The Extent and Consequences of P-Hacking in Science. PLOS Biol. 2015;13: e1002106. doi:10.1371/journal.pbio.1002106

19. Kerr NL. HARKing: Hypothesizing After the Results are Known. Personal Soc Psychol Rev. 1998;2: 196–217. doi:10.1207/s15327957pspr0203_4

20. Tukey JW. We Need Both Exploratory and Confirmatory. Am Stat. 1980;34: 23–25. doi:10.1080/00031305.1980.10482706

21. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Mol Biol. 2003;10: 980–980. doi:10.1038/nsb1203-980

22. Poldrack RA, Gorgolewski KJ. OpenfMRI: Open sharing of task fMRI data. NeuroImage. 2017;144, Part B: 259–261. doi:10.1016/j.neuroimage.2015.05.073

23. 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. doi:10.1038/sdata.2016.18

24. Perkel J. Democratic databases: science on GitHub. Nat News. 2016;538: 127. doi:10.1038/538127a

25. Eglen SJ, Marwick B, Halchenko YO, Hanke M, Sufi S, Gleeson P, et al. Toward standard practices for sharing computer code and programs in neuroscience. Nat Neurosci. 2017;20: 770–773. doi:10.1038/nn.4550

26. Marwick B. Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. J Archaeol Method Theory. 2017;24: 424–450. doi:10.1007/s10816-015-9272-9

27. Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, Leprevost F da V, et al. Ten Simple Rules for Taking Advantage of Git and GitHub. PLOS Comput Biol. 2016;12: e1004947. doi:10.1371/journal.pcbi.1004947

28. Shen C, Björk B-C. 'Predatory' open access: a longitudinal study of article volumes and market characteristics. BMC Med. 2015;13: 230. doi:10.1186/s12916-015-0469-2

29. Björk B-C, Laakso M, Welling P, Paetau P. Anatomy of green open access. J Assoc Inf Sci Technol. 2014;65: 237–250. doi:10.1002/asi.22963

30. Harnad S, Brody T, Vallières F, Carr L, Hitchcock S, Gingras Y, et al. The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update. Ser Rev. 2008;34: 36–40. doi:10.1080/00987913.2008.10765150

31. Laakso M, Welling P, Bukvova H, Nyman L, Björk B-C, Hedlund T. The Development of Open Access Journal Publishing from 1993 to 2009. PLOS ONE. 2011;6: e20961. doi:10.1371/journal.pone.0020961

32. Gorgolewski KJ, Auer T, Calhoun VD, Craddock RC, Das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci Data. 2016;3: sdata201644. doi:10.1038/sdata.2016.44

33. Press release archive: About NPG [Internet]. [cited 19 Oct 2017]. Available: http://www.nature.com/press_releases/scientific-data.html

34. Brew A. Teaching and Research: New relationships and their implications for inquiry-based teaching and learning in higher education. High Educ Res Dev. 2003;22: 3–18. doi:10.1080/0729436032000056571

35. Healey M. Linking Research and Teaching to Benefit Student Learning. J Geogr High Educ. 2005;29: 183–201. doi:10.1080/03098260500130387

36. Robertson J. Beyond the 'research/teaching nexus': exploring the complexity of academic experience. Stud High Educ. 2007;32: 541–556. doi:10.1080/03075070701476043