Weierstraß-Institut

für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Reproducible research through persistently linked and visualized data

Bastian Drees[1], Angelina Kraft[1], Thomas Koprucki[2]

submitted: September 27, 2017

[1] Technische Informationsbibliothek (TIB)
Welfengarten 1B
30167 Hannover
Germany
E-Mail: bastian.drees@tib.eu
angelina.kraft@tib.eu

[2] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: thomas.koprucki@wias-berlin.de

No. 2430
Berlin 2017

2010 *Mathematics Subject Classification.* 68P20, 68U35, 65Yxx.

Key words and phrases. Research data, repositories, visualization, persistent identifier.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Reproducible research through persistently linked and visualized data

Bastian Drees, Angelina Kraft, Thomas Koprucki

Abstract

The demand of reproducible results in the numerical simulation of opto-electronic devices or more general in mathematical modeling and simulation requires the (long-term) accessibility of data and software that were used to generate those results. Moreover, to present those results in a comprehensible manner data visualizations such as videos are useful. Persistent identifier can be used to ensure the permanent connection of these different digital objects thereby preserving all information in the right context. Here we give an overview over the state-of-the art of data preservation, data and software citation and illustrate the benefits and opportunities of enhancing publications with visual simulation data by showing a use case from opto-electronics.

1 Introduction

During the research process in mathematical modeling and simulation (MMS) starting with a research question and ending with published results, a lot of output is generated in a variety of different forms using different kinds of input and methods. This includes mathematical models, research software, code, raw and analyzed research data, visualizations, and so on. These materials are in many cases crucial for reproducible and comprehensible research results. Thus, they should be made available for both editors and reviewers as well as the reader of the respective scientific publication (cf. Fig 1). The demand for submitting comprehensible and reproducible results has been addressed in the field of optoelectronic device simulation by Piprek (2017) and is reflected by implementing general quality guidelines[1] for contributions in the NUSOD 2017 special issue in the journal *Optical and Organic Electronics*. In addition to other requirements, these guidelines expect the author to „prepare supplemental information that enables the reader to exactly reproduce your results by using the same software (models, complete parameter list, input files, etc.)“.[2]

In order to ensure reproducibility in numerical simulations a responsible treatment of mathematical models, software, source code and research data is necessary. To this end researchers should follow some best practices in scientific computing (Wilson et al (2014); or at least good enough practices: Wilson et al (2017)). Moreover, appropriate solutions that guarantee those resources to be findable and long-term accessible are needed. Here, persistent identifiers play a vital role as they remain valid even if the storage location (referred to by e.g. URLs) changes. This is especially important as the information contained in objects like source code, research data, etc. depends strongly on the context, e.g. the underlying mathematical model, description of the simulation studies, parameter values, etc. or, to put it the other way around, the full story can only be understood if all pieces are present. A permanent link between all relevant objects (cf. Fig 1) is therefore a key prerequisite for reproducible research results.

Reproducibility is an essential requirement for all scientific results, but moreover, scholarly publications should also present the underlying data in an understandable way. Data visualizations that capture the dynamics and the general behavior of the system in a large parameter range are a viable alternative for those readers that just want to understand the results without reproducing them themselves.

[1] See `http://www.nusod.org/2017/conf_oqe.html`

[2] Guideline 6

DOI 10.20347/WIAS.PREPRINT.2430

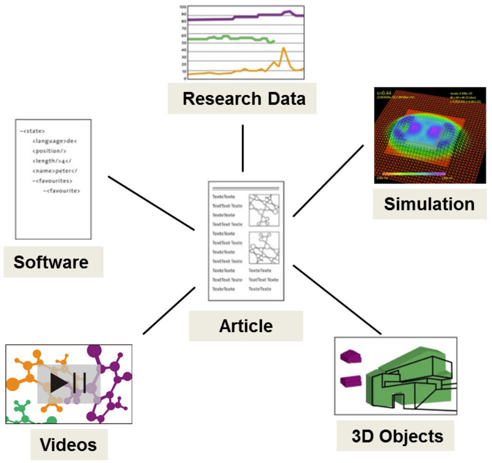

Figure 1: Mathematical modeling and numerical simulation produces research results in many different forms. Publishing all these materials such as data, software, 3D objects (e.g. device geometries) or videos alongside the article improves both comprehensibility and reproducibility of the work presented.

While visualisations of simulation results are quite common nowadays, there still exists no common standard on how to handle these visual data. In most cases the visualisations are kept on the private hard drive of the researcher who publishes only snapshots of the visualisations as figures in scientific articles. However, a growing number of scientists also publishes the visualisations as supplemental videos accompanying the respective article. This allows the reader to get more information on how the research results were found and how to interpret the data. Nonetheless, the way supplemental videos are published is in many aspects far from ideal.

Most publishers do not have the suitable infrastructure for hosting research data, software or scientific videos. Therefore, supplemental videos are often hosted on private webpages, platforms like YouTube or only available on request. These solutions are neither sustainable, nor are the videos findable or citable. Furthermore, from a readers perspective it is often unclear if and how videos may be reused as no legal licence is associated with the video. Finally, even if all the before said is fulfilled it is in general not possible to cite just part of a video, e.g. that single segment corresponding to a snapshot shown in the articles figure.

The aim of the paper is to give an overview over the state-of-the art of data preservation, data and software citation in Section 2. Furthermore, we illustrate in Section 3 the benefits of publishing videos of simulation results by comparing the current version of Kantner et al (2016) with an enhanced version using the AV-Portal[3] as an underlying infrastructure for the visual simulation data used therein.

2 Data and software citation principles

With the new digital research landscape, we experience a paradigm shift in the way knowledge is shared and scientists are confronted with new services around data to support their research. Infrastructure providers such as academic libraries provide new digital curation and publication services for research data (Kraft et al (2017)) and other content such as audio-visual media (Neumann and Plank (2014)). In 2016, the 'FAIR' Guiding Principles for scientific data management and stewardship were

[3] https://av.tib.eu

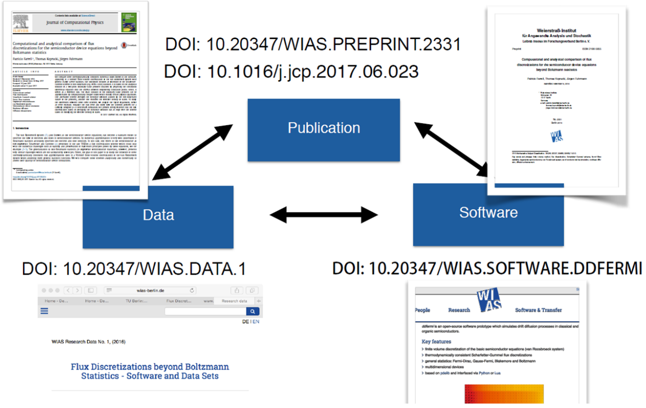
DOI: 10.20347/WIAS.PREPRINT.2331
DOI: 10.1016/j.jcp.2017.06.023
Publication
Data
Software
DOI: 10.20347/WIAS.DATA.1
DOI: 10.20347/WIAS.SOFTWARE.DDFERMI
WIAS Research Data No. 1, (2016)
Flux Discretizations beyond Boltzmann Statistics - Software and Data Sets

Figure 2: Publishing software, data and article enables readers to reproduce research results. All objects are assigned a DOI thereby guaranteeing a persistent link between all of them.

published by the FORCE 11, a group of researchers, publishers and research funders – their goal is to make research data “Findable, Accessible, Interoperable and Reusable” (Wilkinson et al (2016)). The four principles provide a guideline for data producers and data publishers and data repositories to enhance the reusability of scientific results, with a strong emphasis on supporting the ability of machines to automatically find and (re)use the data. Nowadays, the available infrastructures for the management of digital objects such as research data or software are either discipline-specific or generic, and often include a range of different metadata standards and technical protocols. A good overview of available data repositories is provided by the discovery portal r3data.org. A major challenge addressed by the FAIR principles is the difficulty in exchanging and analyzing research data of different types (e.g. from different disciplines), with the goal to achieve a more similar representation for data sets and their associated metadata.

With emphasis on the first principle to make data ‘findable’, the use of unique and persistent identifiers (PID) has become a central aspect of proper management and access. DOI (Digital Object Identifiers = DOI® names) have been originally used for scientific publications as the core technology to refer to the electronic version of an article. A DOI consists of a unique character string which identifies the object itself and not the place where it is located. If the object is moved and the location (URL) has changed, the only requirement is to update the URL in the underlying central database. This system ensures that the DOI persistently resolves to the location of the object (Paskin (2006)). Several registration agencies provide DOI services and registration worldwide. One of them is DataCite, a registration agency particularly dedicated to services that support the enhanced search and discovery of research content, especially on research data and grey literature. The DOI community and the DataCite consortium are maintaining a network of services for institutions and researchers, for example metadata schemas and documentations which build the context information for the DOI framework and provide detailed citation guidelines for research data (DataCite Metadata Working Group (2016)). Within the DataCite Metadata v4.0 Documentation, mandatory, recommended and optional properties for data citation are given, with the mandatory properties including the PID, the creator with optional name identifier and affiliation sub-properties, the dataset title, the publisher, the publication year and the resource type. The usage of persistent identifiers like DOI also enables the development of services such as event tracking of citations leading to a better recognition of scientific achievements. With PID services at the base, infrastructure providers are supporting researchers to deposit their research output into disciplinary research data repositories, e.g. ICPSR in social sciences, GenBank in life sciences, and Pangaea in geo science. In the frame of the EU Horizon 2020 Project „Novel Materials

Discovery (NOMAD) Laboratory“ in computational material science, the NOMAD repository[4] has been launched. As data become more widely available and better integrated into research workflows, services around data such as PIDs are key to the way infrastructure providers need to support research and business.

Good and responsible scientific practice requires gapless traceability and reproducibility of the scientific statements. For producing the underlying data, e.g. in the numerical simulation of optoelectronic devices open-source and commercial software packages and computer resources ranging from desktop PCs to high performance computers are used that allow to perform extensive simulation studies.

In contrast to research data, which is often static, research software is a dynamic, evolving entity and often dependents on other software or libraries. The citation standards of an appropriate software repository have to take this into account. Once software is an essential prerequisite of the scientific process, appropriate recognition of software developers is demanded. To be traceable (and sustainable), a carefully maintained software repository infrastructure is needed, which not only provides maintained PID and metadata information, but also appropriate citation guidelines. For the citation of software, possibly also of individual modules of it, there are various proposed guidelines (for example, the FORCE11 recommendations by Smith et al (2016)).

Using PID for data and software allows to persistently connect these objects with each other as well as with the corresponding article. Farrell et al (2017) employ this principles as schematically shown in Fig. 2.

3 Videos as a generic representation of simulation data

As pointed out above, it is of great importance that research data obeys the FAIR-principles. In addition, research results should be presented in a comprehensible and understandable manner. To this end, it is necessary to visualize the involved research data sufficiently well. However, numerical simulations are often used to solve complex problems in a multi-dimensional (parameter-)space. For instance device simulations may provide the spatio-temporal evolution of carrier densities in semiconductor devices and also the evolution of further quantities charaterizing the state of the device such as the electrostatic potential or the current densities. Therefore, visualizing these aspects is not always easy.

Ideally, mathematical models as well as source code for their numerical solution and research data are published openly alongside with the research article. This enables the interested reader to reproduce and reuse the research results. However, data visualizations that capture the dynamics and the general behavior of the system in a large parameter range are helpful for those readers that just want to understand the discussed effects or phenomena without reproducing them themselves. An interactive visualization tool that enables the reader to freely choose the parameter values of interest could be a solution. The major challenge here is to meet the subject-specific and individual requirements for data visualization and at the same time, find a solution that is generic enough to enable publication, use and preservation by using standard software.

For most simulations and their results this is not easily possible and interactive data visualizations are not very common, yet. A more common extension of graphical representations of research data in figures are videos. Most publishers allow the publication of supplemental videos but in general they do not provide a suitable platform that allows a good combination of article and video.

[4] See `https://nomad-coe.eu`

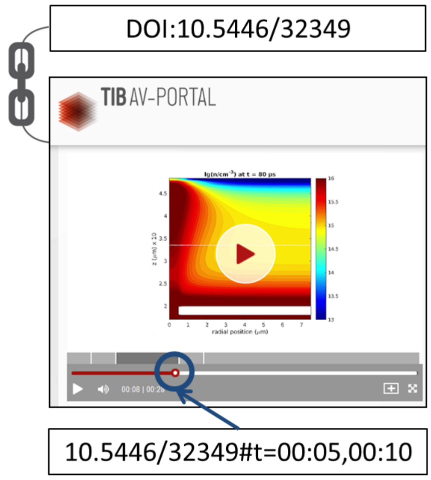

Figure 3: Visual simiulation data as a scientific resource. Videos in the AV-Portal are assigned a DOI (here: DOI:10.5446/32349). Video segments can be cited to the second using a combination of DOI and MFID (here: https://doi.org/10.5446/32349#t=00:05,00:10). Screenshot from Kantner et al (2017a).

3.1 Enhancing publications with videos

In some sense a video is the prolongation of an image in the time dimension. Therefore it allows to illustrate or visualize one dimension more than a figure. When we think about visualization of simulation results, this additional dimension may be of course time, but also any other changing parameter such as voltage or current or the change of material or geometric parameters. Besides providing an additional dimension, videos also enable the author to visualize phenomena which are too complicated or too complex or too full of information to visualize with pictures or describe with words. Examples range from high-resolution numerical simulations that use specialized visualization tools to produce high-quality videos to simple image sequences that are no more than digital flip-books. The latter may consist for example of series of figures that visualize the simulation results at different parameter values. Generating such a simple image sequence, allows to publish large numbers of figures in a single file and captures at the same time the dynamical changes of the system with the changing parameter. Many easy-to-use software solutions are available that allow to create videos from a series of images. Besides the commercial software Adobe Premiere Elements, there are also open source solutions. For instance the free and open source, cross-platform video editor shotcut that runs on all common operating systems. After initially importing the images into a playlist, the resolution and the frame rate can be adjusted. Finally the image sequence can be exported into a video file.

To make full use of the potential of simulation videos a suitable platform is necessary. Such a platform ideally combines state-of-the-art search functionalities, a sustainable infrastructure and the possibility of scholarly video citation. The AV-Portal is such a platform for scientific videos from the realms of science and technology. Automated video analysis, like speech, text and scene recognition, allow for an easy search within the video content. Moreover, all videos in the AV-Portal are assigned a digital object identifier (DOI) that enables the users to cite and link to the article in accordance with scientific standards (cf. Fig. 3). The combination of DOIs with a media fragment identifier (MFID) even allows to cite video segments to the second (cf. Fig. 3). Interlinking between article and video (and if applicable further supplemental resources) via DOI ensures a persistent connection between the different parts

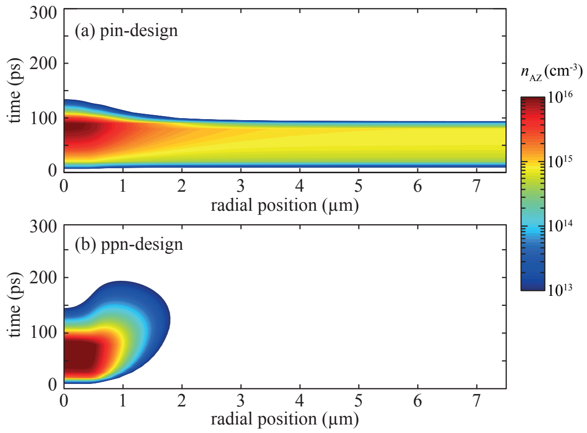

Figure 4: Response of the electron density along the active zone in vertical p-n diodes with laterally oxidized apertures to a periodically pulsed contact voltage (100 ps long bias pules with repetition rate of 1 GHz on top of a constant bias voltage U_0 with 20 ps rise and fall time), see Kantner et al (2016). (a) pin-design, (b) ppn-design. Time is measured relative to the onset of the respective bias pulse. White areas indicate very small electron densities lower than 10^{13} cm^{-3}. The lateral carrier spreading observed for the pin-design is effectively suppressed for the ppn-design.

of a publication. This guarantees that the whole publication, i.e. the article and all related material, remains connected even if URLs are changed or similar (cf. Fig.1).

3.2 An example from opto-electronic device simulation

We consider a use case from opto-electronics and show how the presentation of results from Kantner et al (2016) may benefit from enhancing the publication with the respective video of visualized simulation data. We compare the publication without those enhancements with an enhanced publication using the features offered by the AV-Portal. Finally we propose extensions of the current features to harmonize the citation scheme with the actual physical simulation parameters.

In Kantner et al (2016) the current injection into single quantum dots embedded in vertical pn-diodes featuring oxide apertures has been studied using three-dimensional simulations of the carrier transport. The simulations revealed that the experimentally observed parasitic excitation of QDs located up to several micrometers away form the aperture can be explained by a rapid lateral spreading of the carriers after passing the oxide aperture, see Fig. 4a. Guided by these findings an improved device design has been proposed by Kantner et al (2016) which effectively suppresses the unintended current spreading and allows for an efficient electrical pumping of sub-micron sized regions, see Fig. 4b.

For the analysis in continuous wave (CW) operation mode (controlled by a bias voltage U) color-coded two-dimensional maps of the spatial distribution of the carrier density and the carrier flow for *one* specifically selected value of the bias voltage have been shown suitable for supporting the scientific argumentation. Here, a video showing the evolution of the carrier density together with the current flow for the a suitable bias sweep would be beneficial for the reader to gain insights in the mechanisms behind the observed phenomena.

For the pulsed operation mode the *temporal evolution* of the electron density in the active zone along a *line scan* (1D) across the two-dimensional density map has been shown, see Fig. 4, for demonstrating the excellent current funneling for the improved device design. Here, the interpretation of this plots would considerably benefit from a visualization of spatio-temporal evolution of the full two-dimensional

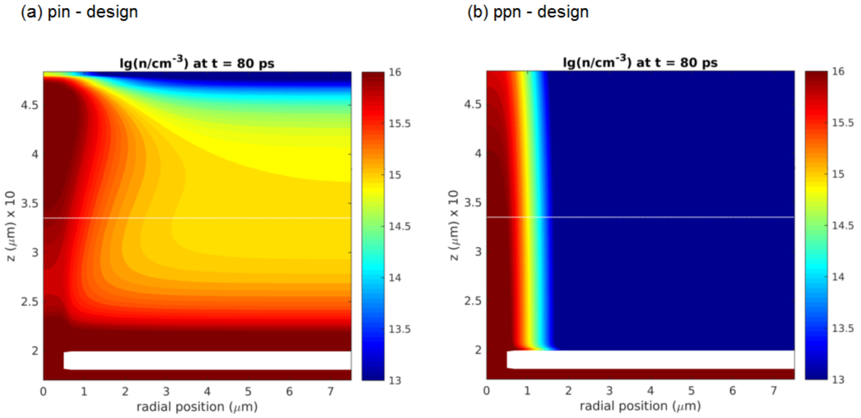

Figure 5: Same results as in figure 4 with the spacial dimension z added at a constant time $t = 80$ps. Snapshots from the videos of (a) p-i-n-design (https://doi.org/10.5446/32349), (b) p-p-n-design (https://doi.org/10.5446/32350).

carrier density. This is shown for the pin-design in Kantner et al (2017a) and for the ppn-design in Kantner et al (2017b).

Figure 5 shows snapshots of these videos at $t = 80$ ps. This single frame may be cited in the video with https://doi.org/10.5446/32349#t=00:08,00:09 for the pin-design and with https://doi.org/10.5446/32350#t=00:08,00:09 for the p-p-n-design. Similarly the results for 50 ps $<$ $t <$ 100 ps can be cited with https://doi.org/10.5446/32349#t=00:05,00:10 for the pin-design and with https://doi.org/10.5446/32350#t=00:05,00:10 for the ppn-design.

The introduction of domain-specific extensions (Refer frames by parameters, e.g. applied voltage, or physical simulation time, e.g. ps in our case) could be a useful next step for authors interested in enhancing their publications by visualizations.

4 Conclusion

The field of optoelectronic device simulation is constantly growing and an increasing amount of research articles are submitted every year (cf. Piprek (2017)). However, in order to publish reproducible and comprehensible research results it is necessary to publish the whole story, that is the underlying mathematical model, software and source code, parameter values, simulation data and suitable vizualisations of simulation data. Therefore, publications should consist of a combination of the text describing the research and its results as well as the underlying material while ensuring all materials being permanently link to each other (cf. Fig. 1). Additionally, those materials should be findable, accessible, interoperable and reusable, i.e. they should obey the FAIR-principles. Because the different object types require very different infrastructures and platforms, a possible solution is to store them separately in suitable repositories and interlink them by using persistent identifiers, e.g. digital object identifier (DOI).

Similar arguments also apply to the treatment of visualizations of simulation data. Often numerical simulations of optoelectronic devices treat multidimensional, dynamic problems that are difficult to communicate by figures or plots alone. Here a sequence of images or a initially produced video from simulation results can enhance the comprehensibility of the presented results. Using the TIB AV-Portal as a platform for videos accompanying numerical simulations ensures a persistent linkage between article and video via DOIs and allows for citation to the second via a combination of DOI and MFID. Moreover, their curation and long-term preservation is in the mission of the TIB.

Even though it requires additional work to publish these objects appropriately and sustainably it will increase the chance of being published as well as enhance the reception and visibility of the research.

References

DataCite Metadata Working Group (2016) Datacite metadata schema for the publication and citation of research data. version 4.0. DOI 10.5438/0012, URL https://doi.org/10.5438/0012

Farrell P, Koprucki T, Fuhrmann J (2017) Computational and analytical comparison of flux discretizations for the semiconductor device equations beyond Boltzmann statistics. Journal of Computational Physics 346:497 – 513, DOI https://doi.org/10.1016/j.jcp.2017.06.023

Kantner M, Bandelow U, Koprucki T, Schulze JH, Strittmatter A, Wünsche HJ (2016) Efficient Current Injection Into Single Quantum Dots Through Oxide–Confined p-n-Diodes. IEEE Transactions on Electron Devices 63(5):2036–2042, DOI 10.1109/TED.2016.2538561, available: http://dx.doi.org/10.1109/TED.2016.2538561

Kantner M, Bandelow U, Koprucki T, Schulze JH, Strittmatter A, Wünsche HJ (2017a) Spatio-temporal evolution of carrier densities in oxide-confined p-n-diodes. - p-i-n-design. DOI 10.5446/32349, URL https://doi.org/10.5446/32349

Kantner M, Bandelow U, Koprucki T, Schulze JH, Strittmatter A, Wünsche HJ (2017b) Spatio-temporal evolution of carrier densities in oxide-confined p-n-diodes. - p-p-n-design. DOI 10.5446/32350, URL https://doi.org/10.5446/32350

Kraft A, Dreyer B, Löwe P, Ziedorn F (2017) 14 years of PID services at the German National Library of Science and Technology (TIB): Connected frameworks, research data and lessons learned from a national research library perspective. Data Science Journal 16(36):1–10, DOI 10.5334/dsj-2017-036, URL https://doi.org/10.5334/dsj-2017-036

Neumann J, Plank M (2014) TIB's portal for audiovisual media. IFLA Journal 40(1):17–23, DOI 10.1177/0340035214526531, URL https://doi.org/10.1177/0340035214526531.

Paskin N (2006) Digital object identifiers for scientific data. Data Science Journal 4:12–20, DOI 10.2481/dsj.4.12, URL https://doi.org/10.2481/dsj.4.12

Piprek J (2017) NUSOD Blog: How to get your simulation paper accepted. URL https://nusod.wordpress.com/2017/01/04/how-to-get-your-simulation-paper-accepted/

Smith A, Katz D, Niemeyer K, FORCE11 Software Citation Working Group (2016) Software citation principles. PeerJ Computer Science 2:e86, DOI 10.7717/peerj-cs.86, URL https://dx.doi.org/10.7717/peerj-cs.86

Wilkinson MD et al (2016) The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(160018), DOI 10.1038/sdata.2016.18, URL https://dx.doi.org/10.1038/sdata.2016.18

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SHD, Huff KD, Mitchell IM, Plumbley MD, Waugh B, White EP, Wilson P (2014) Best practices for scientific computing. PLOS Biology 12(1):1–7, DOI 10.1371/journal.pbio.1001745, URL https://doi.org/10.1371/journal.pbio.1001745

Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L, Teal TK (2017) Good enough practices in scientific computing. PLOS Computational Biology 13(6):1–20, DOI 10.1371/journal.pcbi.1005510, URL https://doi.org/10.1371/journal.pcbi.1005510