Open Data Policy of the Quantum Nanoscience Department, TU Delft

Authors: Anton Akhmerov, Gary Steele

The proposal below was discussed and accepted at the Faculty Meeting of the Quantum Nanoscience Department in Delft University of Technology on September 18th, 2018.

Policy

At this time, we identify two relevant “levels” of publishing data:

- Level 0: Uploading of the numerical data shown in figures in a format that is readable by others

- Level 1: Publication of the raw data and scripts for the full data processing chain that produce the final plots shown in the paper

The department policy is that all QN Principal Investigators (PIs) commit to publishing the data accompanying every paper at least at level 0, simultaneously with the paper being published in a journal. PIs are encouraged to look towards implementing level 1 data publication.

Additionally, the authors of the document offer the following step-by-step guide to publishing research data, and offer personal support in advising how to prepare data for publication.

How to prepare data for Level 0 data publication

To publish data compliant with “level 0”, one should first have the data extracted in a format that is readable to others.

Ideally, one should aim to export your data in a common format such as plaintext “dat” files, for example space, comma, or tab separated.

It is also acceptable to upload the data in another format, as long as that data file is accompanied by an instruction on how to load the data. For example, one could include a piece of matlab or python code that would load the data for the reader. As long as it is documented properly how to load the data, this is sufficient.

How to prepare data for Level 1 data publication

For level 1 data publication, prepare a copy of the raw data files that were recorded to the computer, and that serve as input for the publication figures. To this, include scripts that contain code to process this data and output the plots used in the paper.

For this, it is useful to choose data processing tools that are open source or at the very least publicly available. A common choice in the department is python.

It is good if the code itself is readable without having to install additional software. Commercial software platforms that do not provide free “reader” functionally should be avoided if possible. If your data processing is performed by software that does not support scripting capabilities, the full instructions of the steps of how to perform the data processing, referenced to the menu lists of a given software version, could be included as a substitute for a script.

A “readme.txt” file should also be provided that describes the contents of your dataset, specifies which version of the libraries / software are used, and describes how to run the code and what each script and data file corresponds to.

Uploading the data to a data repository

For both Level 0 and Level 1 open data, the next step is to upload the data to a repository. These repositories guarantee the level of data persistence required to issue a “Digital Object Identifier” (DOI) that can be then included as a reference in your manuscript.

TU Delft is a member of the 4TU data center, which provides a repository that allows TUD employees to upload datasets after filling in an upload form. Another option available worldwide is Zenodo, created by CERN and the EU OpenAIRE project.

An example checklist for level 0 data publication (using Zenodo)

- Go over all the plots in the manuscript and the supplementary.
- Save the processed data for each plot in a CSV file named like: “fig_s5_panel_a.csv”
- Zip all the files together
- Go to https://zenodo.org/deposit, upload the zip and fill out the upload form
- Cite the data in your manuscript. If you reserved the DOI in advance, you may cite the dataset before making it publicly accessible.

An example checklist for level 1 data publication

1. Collect all the raw data that needs to be processed to produce the figures.

    a. A good suggestion is to make a separate subfolder for each figure, or each figure panel
2. Collect the scripts that process this data, and produce the plots as PDF files. These PDF files of the output figures should ideally have names like “fig2.pdf” or “fig2a.pdf”, etc.
3. Verify that running the scripts without human intervention outputs the plots as they appear in the publication (excluding any manual post-processing, like adding labels).
4. Create a text file called “readme.txt” in which you include:
    - A list the software required to run the scripts
    - A brief summary of what each of the raw datasets are
    - A brief summary of what is done in the scripts
5. **Optionally** ask a person not involved in working on the manuscript to run the files and verify that the instructions are clear, and the result is as expected.
6. ZIP all the files together and upload

To make your scripts as useful as possible to others, it is best to make code well commented so that it is easy to follow.

Relevant documents

- ERC now runs a pilot, but their guidelines suggest that very soon they will apply strict standards, possibly even exceeding what we suggest in this document.
- NWO requires a data management section in proposals, and a data management plan. The data management section is informational, the plan requires approval. NWO does not specify how they evaluate the data management plan and what the minimal requirements are right now.
- TU Delft offers a guideline on open science and open data, but does not have an open data policy. **UPDATE**: in June 2018 TU Delft approved the open data policy framework. The policy document describes specific obligations of PIs.

Additional considerations

- Level 0 data publishing will likely become the norm in coming years, and also will likely become a formal requirement for many funding agencies and some journals within 5 years.

- The benefit of publishing the data is an increased likelihood of it being used directly in follow-up research. Now, when it is relatively rare, this also gives you an increased visibility: it’s good for you if people in other groups use your code because it means they know about you and your group.

- Publishing the data also motivates researchers to be more systematic in their data processing. While initially it is an investment of labor, in the long run it saves time due to the increased automation.

- Published datasets and code are also a useful resource for people in the field, including future generations of PhD students in your own group, to learn exactly how to do the processing needed for the physics you do.

- It is our opinion that it should the responsibility of the PIs in the department to ensure that the data from their group is published. We are happy to help support groups in the department in learning how to make this process efficient.