Concepts and services for the homogenization and management of file structures in collaborative neuroscientific projects

Thorsten Arendt[1], Achilleas Koutsou[2], Deepti Mittal[3], Keisuke Sehara[4], Rike-Benjamin Schuppner[4], Matthew Larkum[4], Thomas Wachtler[2] and Julien Colomb[4]

[1]Philipps-Universität Marburg
[2]Ludwig-Maximilians-Universität München
[3]Ruprecht-Karls-Universität Heidelberg
[4]Humboldt-Universität zu Berlin

With the GIN-Tonic tool, we provide researchers with a default file organization and file sharing system for research projects in order to facilitate research collaboration and lab management. In contrast to software developers, researchers mostly do not organize files according to a common standard. While data managers propose to design and follow such an organization, they fail at providing clear recommendations or examples to researchers; and there is no time specifically assigned to this task in the researcher's work. We believe that providing researchers with a commonly accepted folder tree structure template could make a huge difference in promoting data management and facilitating research collaboration. This paper presents the results of an initial survey run in three neuroscientific collaborative research centres in Germany (CRC 1315, CRC 1158, CRC/TRR 135), including a presentation of a new folder structure and its technical implementation in the GIN-Tonic application.

1 Introduction

Every day, researchers spend time doing file management on their computers (creating, downloading, naming, moving, saving, copying, reviewing, navigating, searching for, sharing, and deleting files and folders). While many different initiatives and tools have tried to improve file management (using tags, databases and search algorithms), the use of a folder tree structure appeared to be unavoidable and necessary [1]. In addition, both proponents of reproducible research and data management experts recommend researchers the use of an appropriate folder organizational structure [2, 3].

Das hier beschriebene Poster ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI: https://doi.org/10.11588/heidok.00029718 veröffentlicht.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13756 (CC BY-SA 4.0)

However, only few actually provide examples or attempt to bring uniformity in such structures (see [4] and [5] for exceptions).

As data managers of different institutions working with neurobiologists, we teamed up with the NFDI Neuroscience community to develop a new strategy to support researchers in data management. We hypothesize that implementing a homogeneous directory structure using a template could help researchers collaborate on their projects, and manage data and files better.

In a first step, we collected the feedback from 51 neuroscientists presented with two initial template drafts (see figure in the upper-left part of the corresponding poster), analysed their responses, and built an updated template (see figure in the right part of the corresponding poster and `https://doi.org/10.5281/zenodo.4410128` to download the template(s)). The new template takes three levels of projects organization into account (experiment, project, and laboratory), while staying fairly simple and flexible. This extended abstract ends by outlining the GIN-Tonic application, that brings some technical solution (based on the git sub module technology) to add flexibility and ease of use to the template.

2 The Survey

Two template structures

In order to obtain a practical flair in comments and feedback received, we provided two templates, both having a similar number of folders, but organized differently. The templates were obtained by analysing current research work flows discussed during a number of interviews with researchers of the involved consortia. The *5_top* template represented a more hierarchical structure, while the *9_top* template represented a more flat structure (see figure in the upper-left part of the corresponding poster). Then, we asked researchers to browse the folder tree while asking them to place or find specific files, hoping researchers will make themselves familiar with the template before giving us their feedback.

No clear preference for one or the other template

We ran the survey on the involved CRCs during autumn 2020 (using the lime survey application) and finally got 51 responses. In the meantime, we prepared the analysis code using synthetic data. The analysis was run on the final data to produce a reproducible report. In general, researchers reacted very positively to our project. Surprisingly, about half of the participants preferred one template, while the other half preferred the other one (see figure in the lower-left part of the corresponding poster; participants had to choose one or the other template). This preference was highly correlated with the similarity of the template to the structure they currently use (Pearson Chi-square, p-value = 2.23e-06, no

correction for multiple tests was performed). However, we could not identify any notable effect of career stage or research domain on this preference, or any effect of this preference on willingness to use such a template. In particular, computer scientists did not seem to differ from wet-lab researchers in these aspects.

We asked three questions about where they would save or search for specific documents in the different structures. Researchers were indicating different folders, and few chose the folder we designed for that data, showing that a too detailed folder structure seems to be rather inconvenient. On the other hand, researchers would navigate the repository to find specific files using similar strategies, suggesting that having a structure can be helpful, and may reduce the time to browse for specific information.

Issues

Most participants do see the advantages of having a standardized structure for their files, but were critical about the **cost to benefit ratio** of the process, especially for ongoing projects. Many mentioned that it would only reach full impact if the whole lab would be using it, emphasizing the advantage of such a system for collaborative work. In addition, they mentioned the time saved by not having to create a template for themselves, but only use an existing standard. While they were quite unanimous about using a template for new projects, they were sceptical about making a transition for ongoing projects, as the cost of transition might be too high. More generally, they were worried that learning a new work flow to use the template could be time consuming. In many cases, people mentioned that the **files are organized via experiments**, not at the project level. In particular, people tend to pool data and code in a single experiment folder. This is reminiscent of the two example structures given in the library carpentry course [2]. On the other hand, some **files are organized outside of project folders**, that is in particular places irrespective of the project they belong to. Many researchers reported having a folder for all conference reports or all manuscripts, for instance.

3 The revised Template

Template overview

Data management principles recommend (1) to keep all files related to a project in a single folder (this facilitates sharing of these files with the whole team working on the project, for instance), and (2) to manage data and code differently (this allows different version control systems, as well as independent sharing and reusing of data and code). We finally designed a template that follows these principles, but added some recommendations and technical solutions in order to permit users to have laboratory and experiment-level organization of their files, nevertheless.

The figure in top-right part of the poster shows the folder structure developed after analysing the answers of the survey. The template works mostly on the project level (one unique folder for all files related to one project). The experiment level is taken care by specifying several experiment sub folders when new experiments are started. By sharing specific sub folder independently in a cloud solution, one can reorganize information in cross-project directories that host sub folders coming from different projects. Note that both the creation of experiment sub folders and the creation of laboratory level organization is automated in GIN-Tonic.

Table 1: Our definition of the organisation levels.

Organisation level	Definition
Experiment	The unit of research involving a statistically dependent datasets that are analyzed together. It mostly produces one figure. Different experiments typically involve different methods, or different samples, and could be ending in different publications.
Research Project	The unit of research that address a specific research question. It can mostly be delimited by the team involved and typically produce a unique research paper.
Laboratory	Any organization that involves files from several projects. It can also be for a unique researcher, or for a consortium of laboratories.

The experiment level

We propose to keep data and code in different first level folders, and to create several new folders (new data, analysis, and figure folders) for each new experiment. In addition, some of these new folders may also follow their own templates. For instance, some researchers could use a specific BIDS standard template [6] for some imaging experiments.

The project and laboratory level

We propose to share some sub folders independently (for shared figures, report and conferences, and manuscripts) in order to be able to have them in the project folder or in a different folder structure merging information coming from different projects. One could for instance create a folder containing all manuscripts prepared in the lab (see figure in top-right part of the poster).

4 Sharing and automation

Automation possibilities

The creation of different sub folders for one experiment could easily be automated in your computer language of choice. An automation would make sure that the folder names are kept consistent for each experiment. Working on the laboratory level is more complex. If one wants to have cross-folder organization locally, one can work with *alias* folders, where the user can create short-cuts to specific folders using a different organization. The data exists only once and there is no issue of synchronization.

The expected use case is different though, and we expect some users to have certain files organized by projects and other users having the same files organised by file type. This requires to share sub folders independently and set the different instance of these sub folders to be linked together. This is pretty complex to set up using common cloud technology (DropBox-like) that cannot be easily automated. As explained below, the open source GIN-Tonic tool allows to set it up automatically, using the git sub-module technology.

GIN-Tonic implementation

GIN is the G-Node infrastructure. It is based on gogs, git and git annex technologies and brings non only most of the project management and coordination tools that made the success of open source software development, but also large file support and data publication. It is compatible with the git sub-module technology, where sub folders can be synchronized, shared and published independently of the other sub folders, while looking completely normal on one's computer. GIN brings therefore the possibility to publish sub-modules independently of each others, which will ease the opening of research data. It might also make the use of markdown and LaTeX to write manuscripts a straightforward choice, as these technologies can use git as a version control system. We are building an extension that will facilitate some administrative tasks and automate some complex work flows linked to the use of the template. We could not resist calling it *Tonic*, in reference to the vigor added to the GIN tool. The Tonic tool is still under development, and we are also working on the implementation of the Tonic concept for GitLab-based platforms (that is, using git and git-LFS), called LAB-Tonic.

The Tonic application creates a new project repository, clones the research folder structure (with some folders being created as sub-modules), and adds a script that will synchronize the repository and its sub-modules on a double click. It also adds sub-modules (*shared_figures*, *manuscripts*, and *report_conf*) to laboratory-wide repositories, automatically allowing laboratory level organization of some files. This means that for example a manuscript draft can be available in the project folder on the student computer, while the same data will be available in the *lab_manuscript* folder on the PI computer, both versions being synchronized with a unique version on the GIN server. Furthermore, Tonic

will also be able to add sub-modules and folders to the parent repository. For each experiment performed, a data sub-module will be created, so that the data can be curated and published independently of the other experiments, while other normal folders will be created (see figure in the upper-right part of the corresponding poster). A synchronization of the computer version will then bring these changes to the local version, showing the new folders ready to be filled with data, code, or figures.

5 Conclusions

With this project, we hope to provide the research community with a useful project folder structure template. A follow-up survey will tell us how wide it could be applied, and whether domain specific templates may be needed. The template will get its full power when used inside the GIN-Tonic application. Tonic will indeed automate several administrative tasks, like the production of sub folders upon new experiments, and come with a predefined rule for sharing one's files in the lab.

Acknowledgements

This project has been partially funded by Deutsche Forschungsgemeinschaft (DFG), project numbers 222641018 – CRC/TRR 135 TP INF, 255156212 – CRC 1158 TP Z01, and 327654276 – CRC 1315 TP Z.

Bibliography

[1] Jesse David Dinneen and Charles-Antoine Julien. "The Ubiquitous Digital File: A Review of File Management Research" Journal of the Association for Information Science and Technology 71, no. 1 (January 2020), https://doi.org/10/ghssbm.

[2] Library Carpentry. September 2019.https://librarycarpentry.org/lc-fair-research.

[3] Arnold, Becky, Louise Bowler, Sarah Gibson, Patricia Herterich, Rosie Higman, Anna Krystalli, Alexander Morley, Martin O'Reilly, Kirstie Whitaker, and The Turing Way Community. "The Turing Way: A Handbook for Reproducible Data Science" 2019. https://doi.org/10.5281/zenodo.3233986.

[4] Vuorre, Matti, and Matthew J. C. Crump. "Sharing and Organizing Research Products as R Packages" Behavior Research Methods, September 2020. https://doi.org/10/gg9w4c.

[5] Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. “Good Enough Practices in Scientific Computing“ PLOS Computational Biology 13, no. 6 (June 2017): e1005510. https://doi.org/10/gbkbwp.

[6] Gorgolewski, K., Auer, T., Calhoun, V. et al. “The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments“ Sci Data 3, 160044 (2016). https://doi.org/10.1038/sdata.2016.44.