Data management

Data (mis)management in practice

	Data acquisition	Analysis	First submission	Review	Second submission	Publication
Raw data	Data arrives in cumbersome and proprietary format.	Gets converted to format of choice. Original files (and conversion settings) are lost.		Leads a quiet life on the HPC cluster, until the project expires and the data has to be urgently retrieved.	Ends its days on an external hard drive on the researcher's desk.	"Data available upon request".
Meta data	In researcher's lab journal.	Hard-coded in various analysis scripts.	Mailed back and forth between collaborators in ever-changing (but nicely colored) Excel sheets.		Reformatted and included as PDF in the supplementary.	

FAIR

Strive to make your data **FAIR** – Findable, Accessible, Interoperable, and Reusable *for both machines and humans*.

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards


www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*#

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Supporting discovery through good data management
Good data management is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process. Unfortunately, the existing digital ecosystem surrounding scholarly data publication prevents us from extracting maximum benefit from our research investments (e.g., ref. 1). Partially in response to this, science funders, publishers and governmental agencies are beginning to require data management and stewardship plans for data generated in publicly funded experiments. Beyond proper collection, annotation, and archival, data stewardship includes the notion of 'long-term care' of valuable digital assets, with the goal that they should be discovered and re-used for downstream investigations, either alone, or in combination with newly generated data. The outcomes from good data management and stewardship, therefore, are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies. What constitutes 'good data management' is, however, largely undefined, and is generally left as a decision for the data or repository owner. Therefore, bringing some clarity around the goals and desiderata of good data management and stewardship, and defining simple guideposts to inform those who publish and/or preserve scholarly data, would be of great utility.

This article describes four foundational principles—Findability, Accessibility, Interoperability, and Reusability—that serve to guide data producers and publishers as they navigate around these obstacles, thereby helping to maximize the added-value gained by contemporary, formal scholarly digital publishing. Importantly, it is our intent that the principles apply not only to 'data' in the conventional sense, but also to the algorithms, tools, and workflows that led to that data. All scholarly digital research objects[2]—from data to analytical pipelines—benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability.

There are numerous and diverse stakeholders who stand to benefit from overcoming these obstacles: researchers wanting to share, get credit, and reuse each other's data and interpretations; professional data publishers offering their services; software and tool-builders providing data analysis and processing services such as reusable workflows; funding agencies (private and public) increasingly

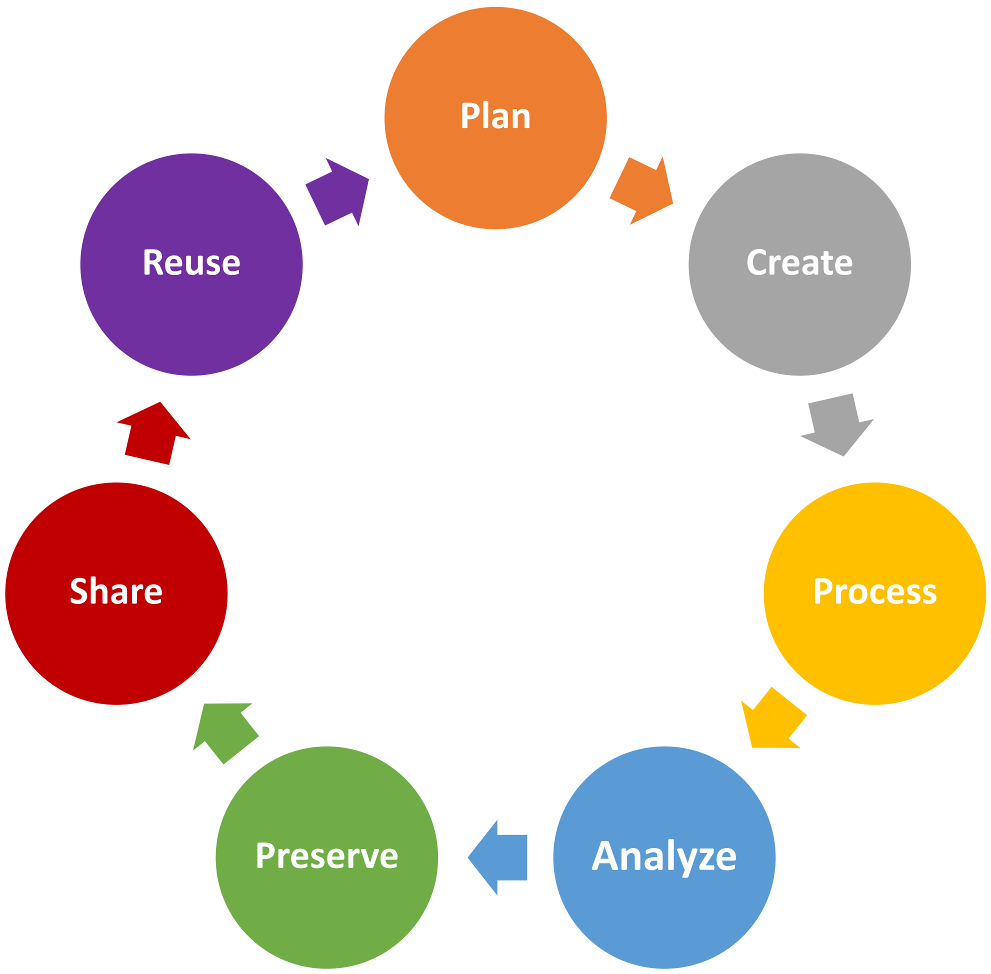
Correspondence and requests for materials should be addressed to B.M. (email: barend.mons@dtls.nl).
#A full list of authors and their affiliations appears at the end of the paper.

SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18 1

Wilkinson, Mark et al. "The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data 3, 160018 (2016) doi:10.1038/sdata.2016.18

Data management plan

- Check the requirements of your funding agency and field of research.
- List the types of data that you expect to produce.
- Decide what data require archiving, and determine how much storage space you will need (short and long term).
- Provide metadata that allows others to understand, cite and reuse your data files.
- Make clear how and when your data can be shared with scientists outside your group.
- If your research involves sensitive data, explain any legal and ethical restrictions on data access and reuse.
- Look for suitable data repositories used by your research community.
- Check what data format and structure the chosen repository might request.



Life cycle for scientific data

Adapted from Nature 555, 403-405 (2018), doi: 10.1038/d41586-018-03071-1

Pair up and discuss!

- Does your group have a data management plan in place?
- Do you know "your" repositories and how to submit data to them?

Data acquisition and deposit

- Find the right repository for your data, and strive towards uploading data to its final destination already at the beginning of a project.
- Structure metadata in the format needed by the repository already as the experiments are being performed.
- Stick to non-proprietary and widely used file formats.

- Only 12% of articles from NIH funded research mention data deposited in international repositories
- Estimated 200000+ “invisible” data sets / year

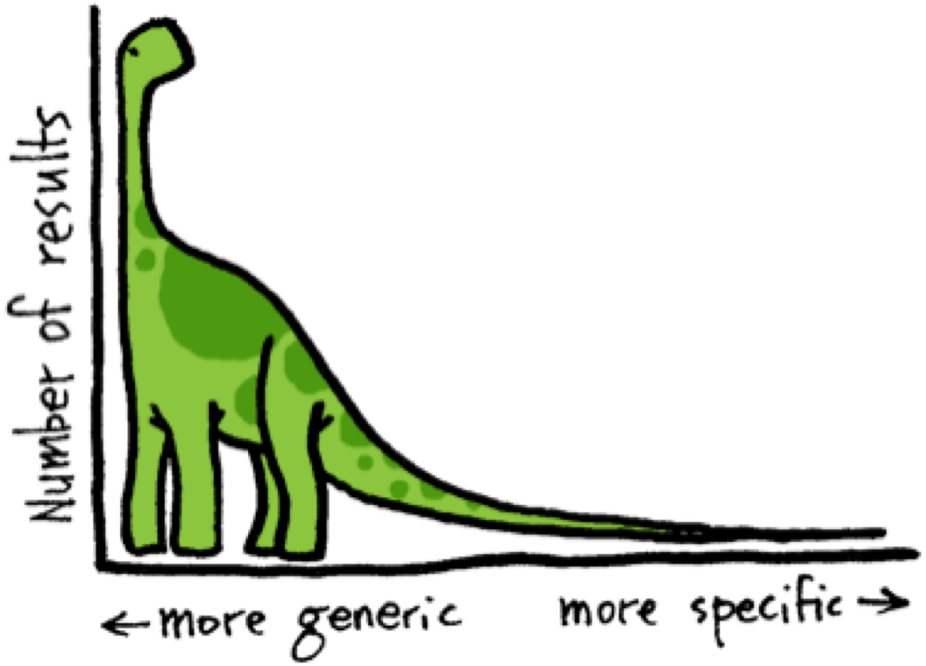
Read et al. (2015) PLoS ONE 10(7) doi:10.1371/journal.pone.0132735

Scientific Data (Springer Nature) maintains a list of recommended repositories at www.nature.com/sdata/policies/repositories.

Dedicated repositories:
e.g. SRA, GEO, GenBank, UniProt etc.

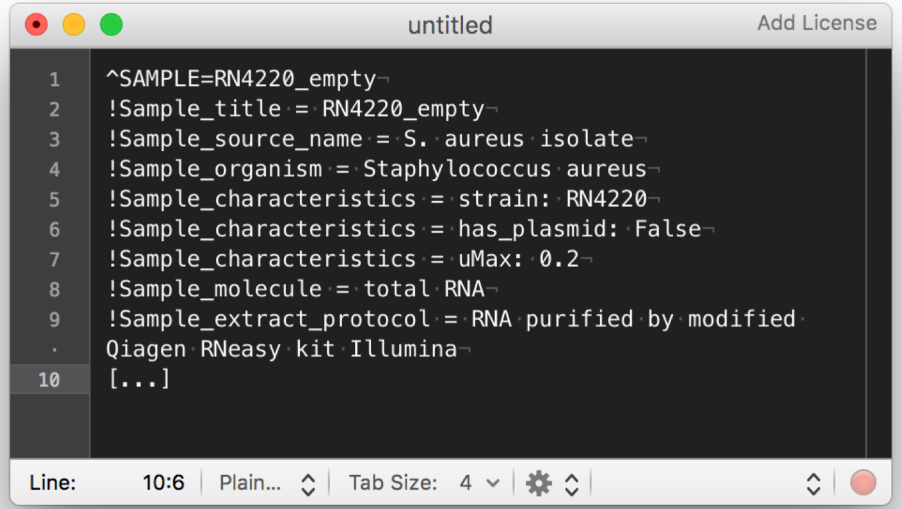
Generalist ("long-tail data") repositories:
Research data that doesn’t fit in structured data repositories, e.g. Data Dryad, Figshare, Zenodo.

Each dataset can be assigned a Digital Object Identifier (doi);
a persistent identifier used to uniquely identify objects.

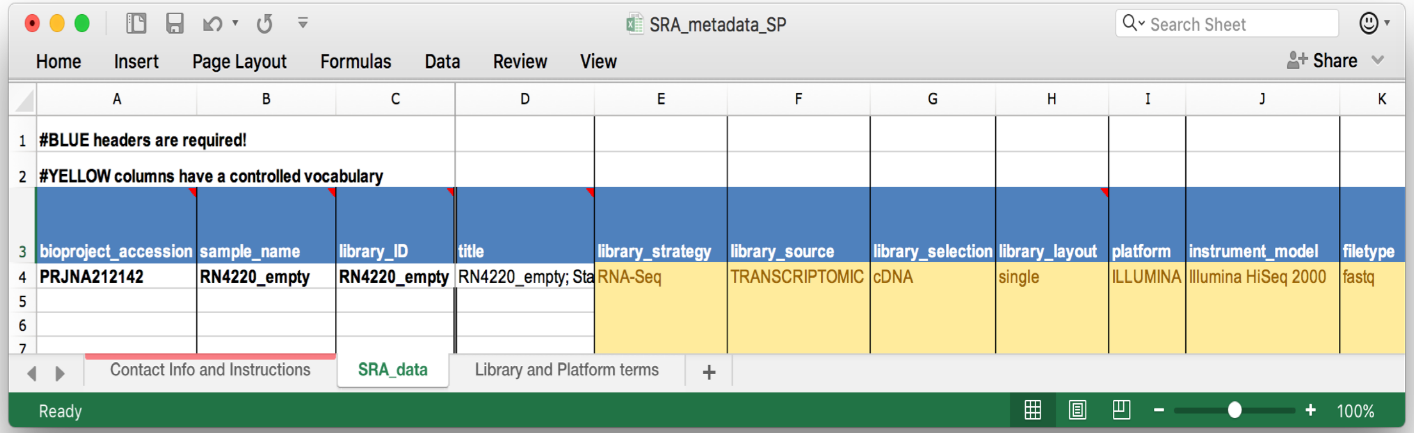


Data acquisition and deposit

- Find the right repository for your data, and strive towards uploading data to its final destination already at the beginning of a project.
- Structure metadata in the format needed by the repository already as the experiments are being performed.
- Stick to non-proprietary and widely used file formats.



GEO (Gene Expression Omnibus) uses text files in SOFT format.



SRA (Sequence Read Archive) uses a template Excel sheet for metadata.

Data acquisition and deposit

- Find the right repository for your data, and strive towards uploading data to its final destination already at the beginning of a project.
- Structure metadata in the format needed by the repository already as the experiments are being performed.
- Stick to non-proprietary and widely used file formats.

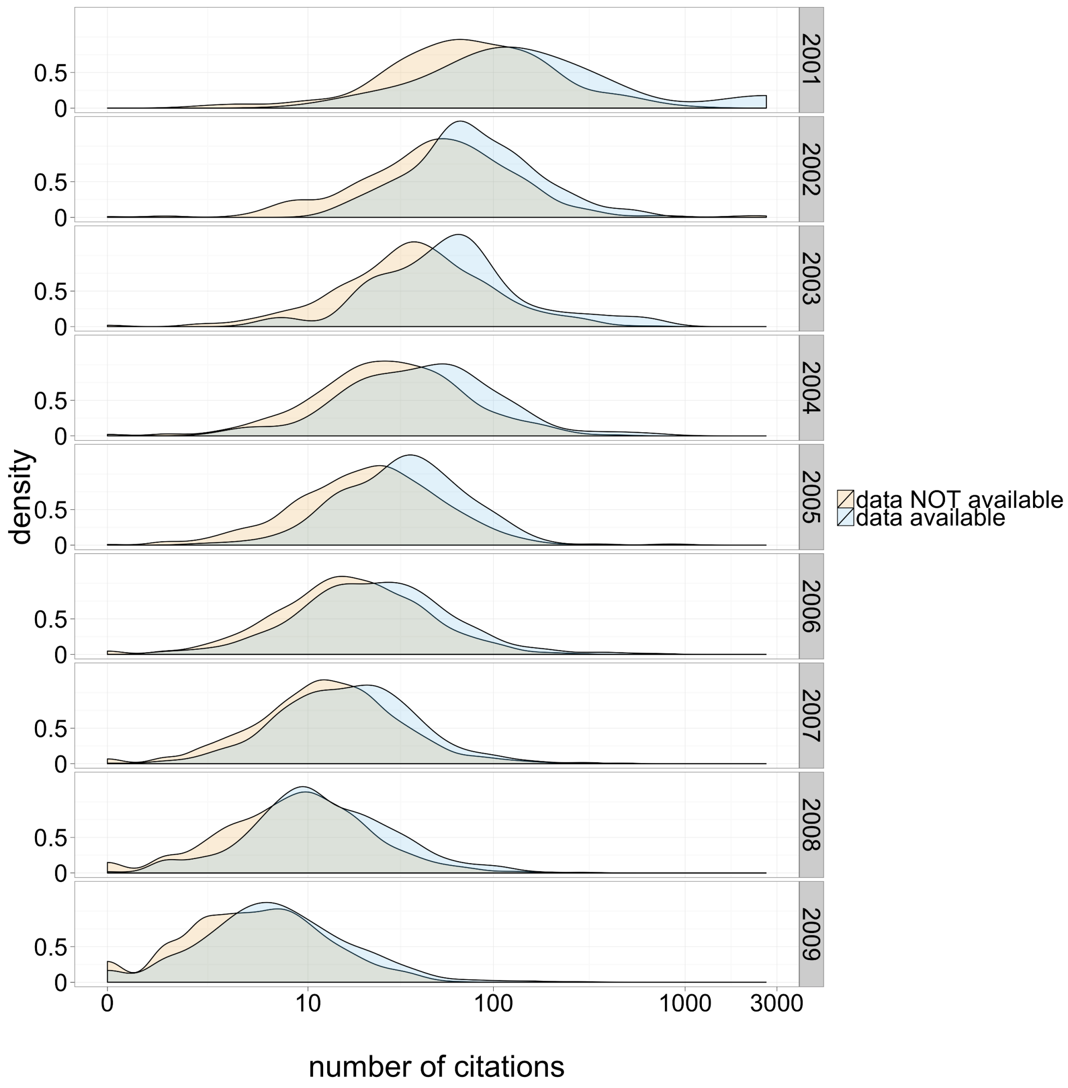
-	+
Binary	Text-based
Proprietary	Open
New kid on the block	Old as the hills
Compressed/encrypted	Uncompressed/unencrypted
Platform dependent	Interoperable
Complex	Simple

	🤢	🙄	😀
Raster graphic	wmf, psd	bmp, gif	tiff, png, jpeg
Vector graphic	ai, eps	pdf	svg
Document	doc	docx, tex	odt, utf-8, md
Archive	rar	7z	zip, tar, gz
Tabular data	xls, rds, mat	xlsx, ods	csv

Data sharing

From 10,555 studies with gene expression microarray data:

- Studies that shared data received 9% more citations (after accounting for other covariates).
- Data reuse by other researchers continued for >6 years.
- A very conservative estimate found that 20% of the datasets deposited between 2003 and 2007 had been reused at least once by third parties.



Piwowar and Vision (2013), Data reuse and the open data citation advantage, PeerJ 1:e175, doi:10.7717/peerj.175

Data sharing – Open access

- Democracy and transparency
  - Publicly funded research data should be accessible to all free of charge.
  - Published results and conclusions should be possible to check by others.
- Research
  - Enables others to combine data, address new questions, and develop new analytical methods.
  - Reduce duplication and waste.
- Innovation and utilization outside research
  - Public authorities, companies, and individuals outside academia can make use of the data.
- Citation
  - Citation of data will be a merit for the researcher that produced it.

Data sharing – Ontologies

lauroyl-CoA

dodecanoyl-CoA

C12:0-CoA

lauroyl coenzyme A

coenzyme A, S-dodecanoate

dodecanoyl coenzyme A

C12:0 coenzyme A

dodecanoic acid coenzyme A

lauroylic acid CoA

Dodecanethioic acid, S-ester with coenzyme A

Coenzyme A, S-laurate (7CI,8CI)

12:0, lauroyl-CoA

1-undecanecarboxylic acid CoA

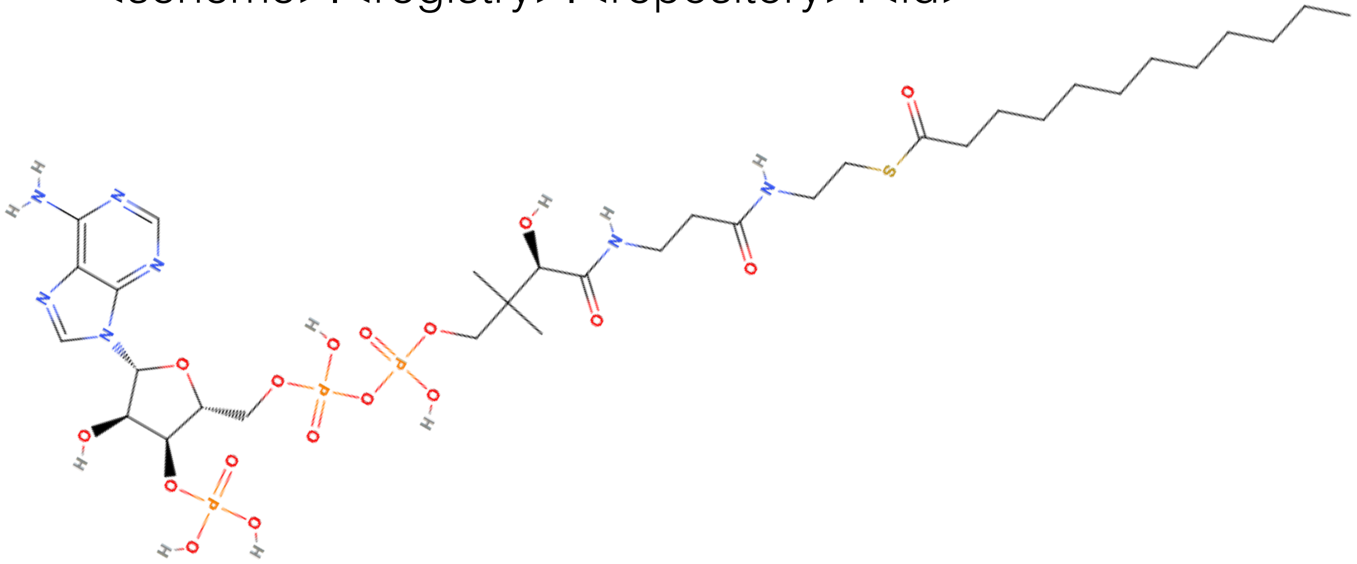
vulvic acid CoA

3'-phosphoadenosine 5'-(3-{(3R)-4-[(3-{[2-(dodecanoylsulfanyl)ethyl]amino}-3-oxopropyl)amino]-3-hydroxy-2,2-dimethyl-4-oxobutyl} dihydrogen diphosphate)

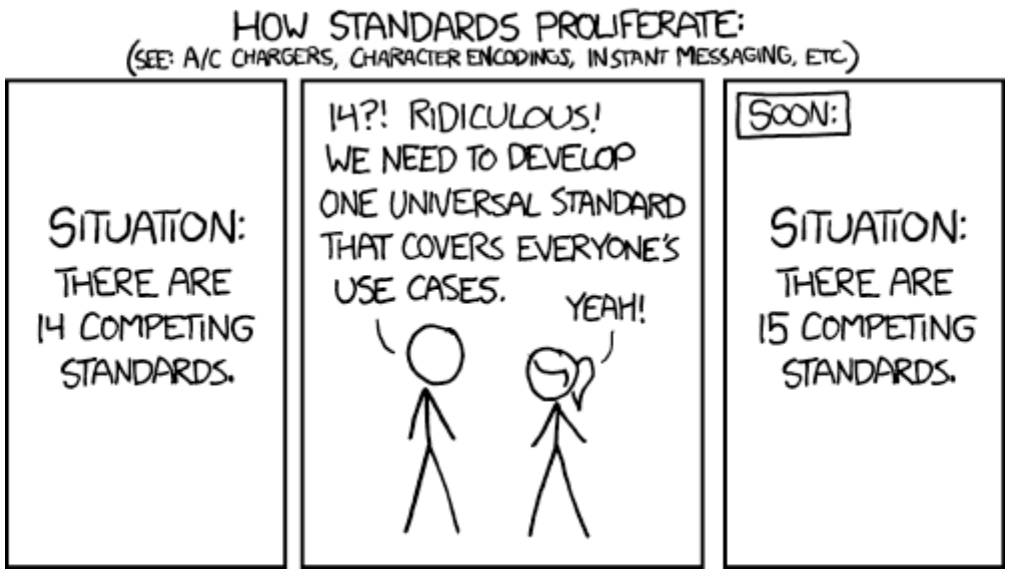
Who am I?

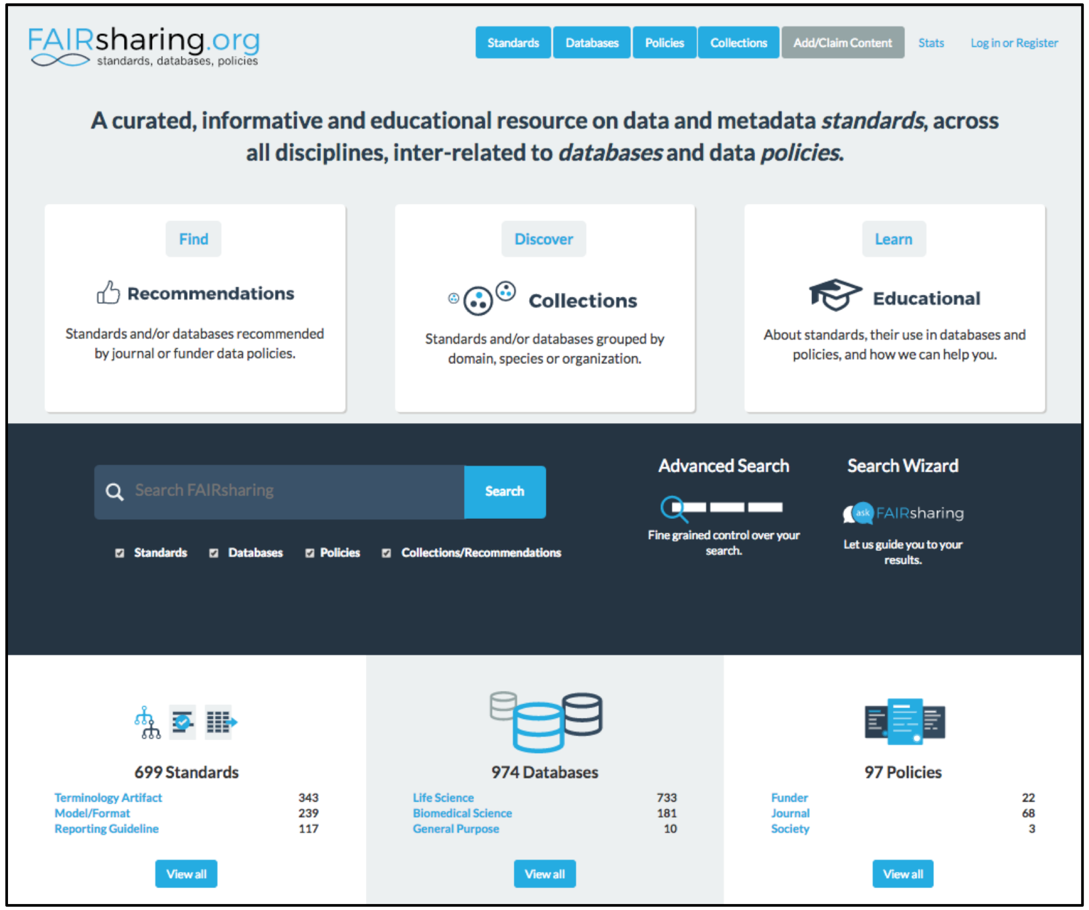
urn.miriam.chebi:15521 of course!

<scheme>.<registry>.<repository>:<id>



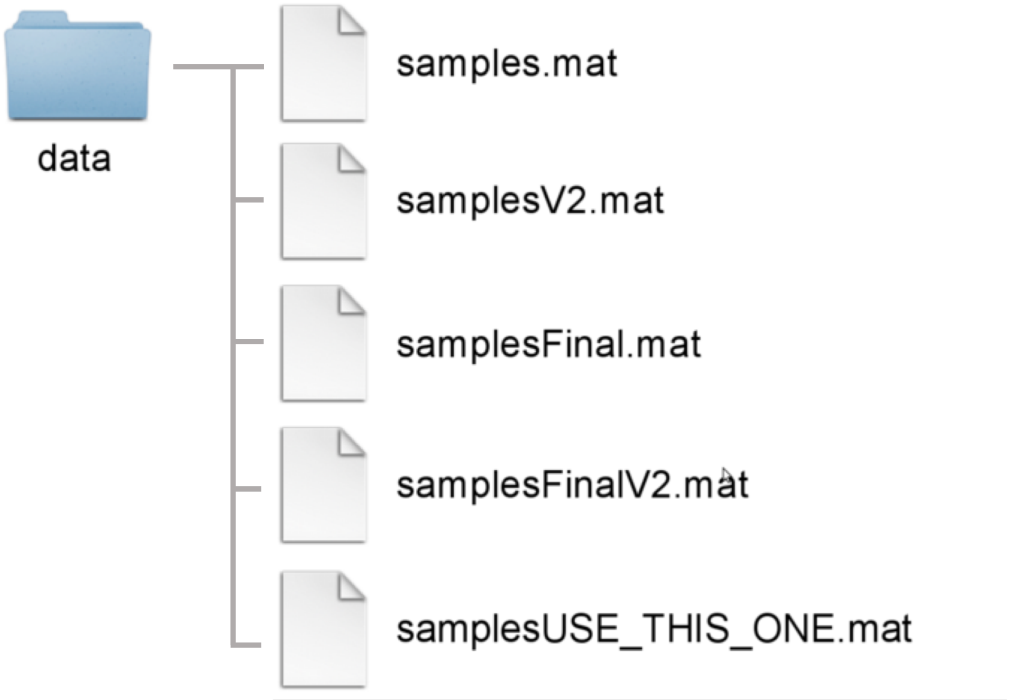
Data sharing – Ontologies





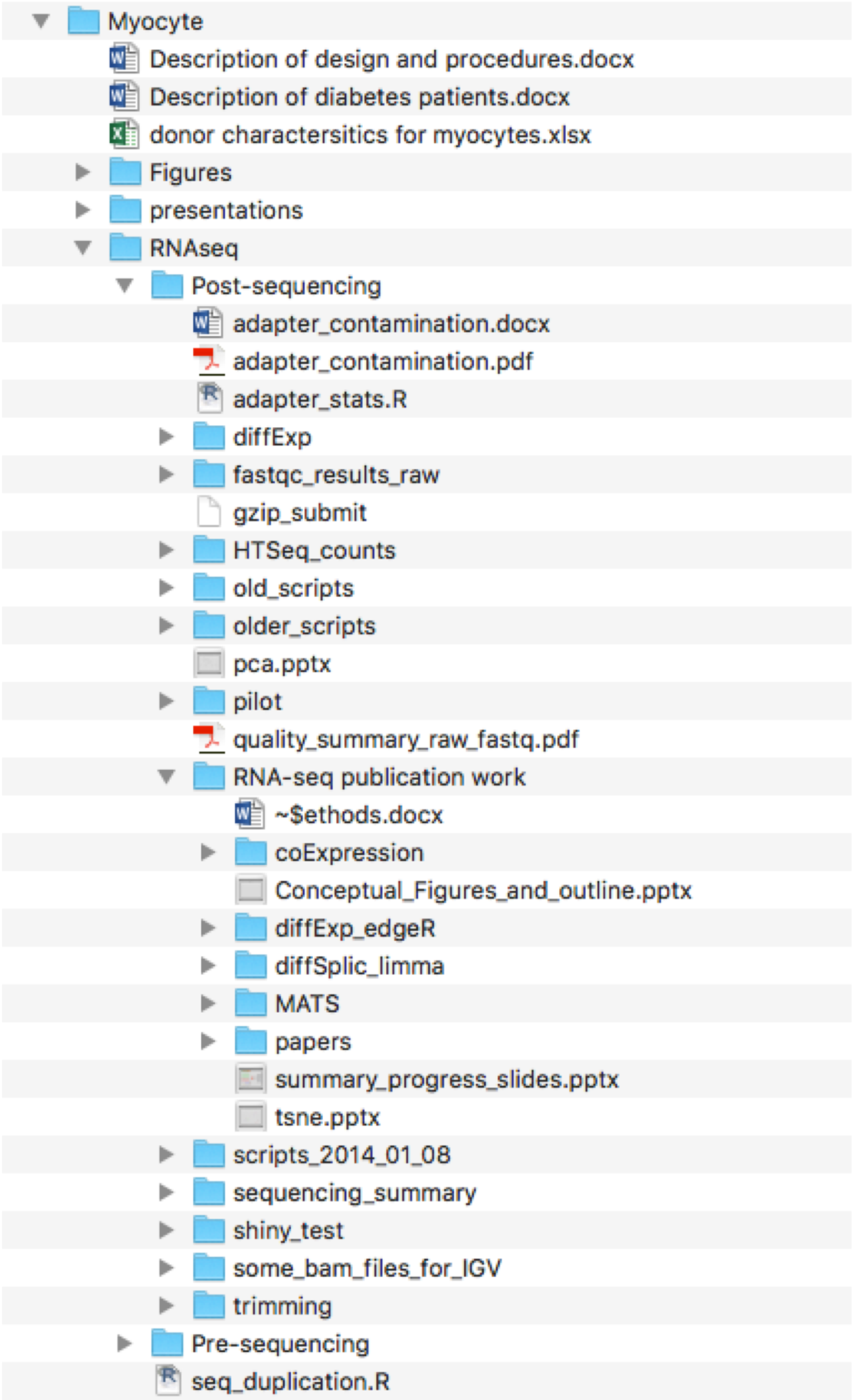
Project organization

The project directory



The first step towards working reproducible: Get organized!

Divide your work into distinct projects and keep all files needed to go from raw data to final results in a dedicated directory with relevant subdirectories.



Pair up and discuss!

- Do you organize your work in distinct projects?
- How do you organize your files in this context?
- Are you happy with the way you work today?

The project directory

```
project
|- doc/                 documentation for the study
|
|- data/                raw and primary data, essentially all input files, never edit!
|   |- raw_external/
|   |- raw_internal/
|   |- meta/
|
|- code/                all code needed to go from input files to final results
|- notebooks/           notebooks that document your day-to-day work
|
|- intermediate/        output files from different analysis steps, can be deleted
|- scratch/             temporary files that can be safely deleted or lost
|- logs/                logs from the different analysis steps
|
|- results/             output from workflows and analyses
|   |- figures/
|   |- tables/
|   |- reports/
|
|- Snakefile            project workflow, carries out analysis contained in code/
|- config.yml           configuration of the project workflow
|- environment.yml      software dependencies list, used to create a project environment
|- Dockerfile           recipe to create a project container
```

Working in projects

RStudio

ATOM

Sublime Text

git

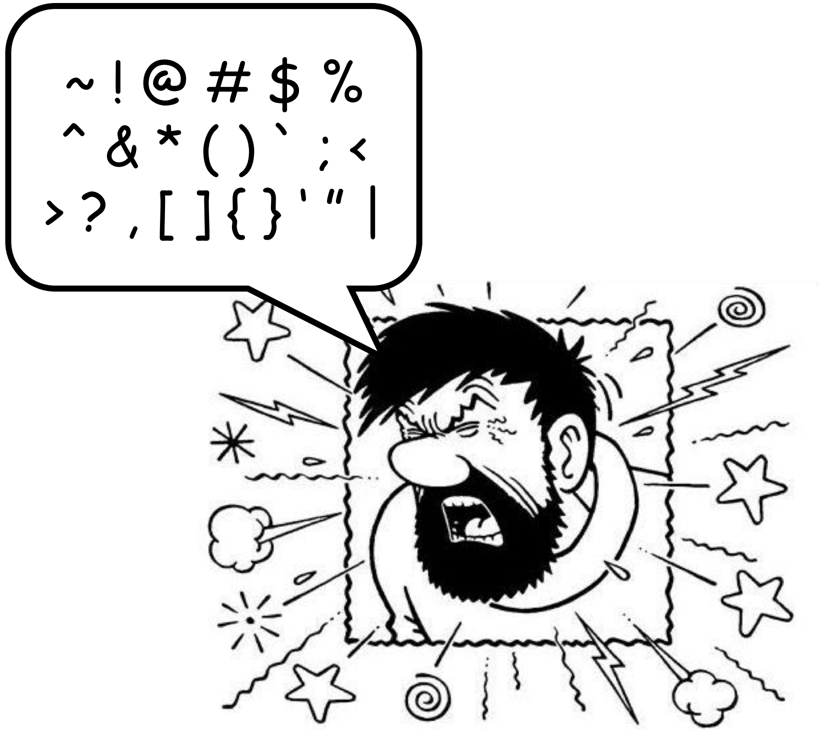
Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLoS Comput Biol 5(7): e1000424.
http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1000424

https://github.com/NBISweden/project_template

File naming system

Machine readable

- Avoid special characters, e.g.: `~!@#$%^&*()`;<>?,[]{}‘”|`
- Avoid spaces, alternatives:
  - `file_name.txt`
  - `file-name.txt`
  - `filename.txt`
  - `FileName.txt`



Human readable

- Know the content of a file without opening it, e.g.:
  `SRR1234.hg19.sorted.trimmed.bam`

Control file ordering

- Use dates if appropriate
- Use 01, 02, rather than 1, 2

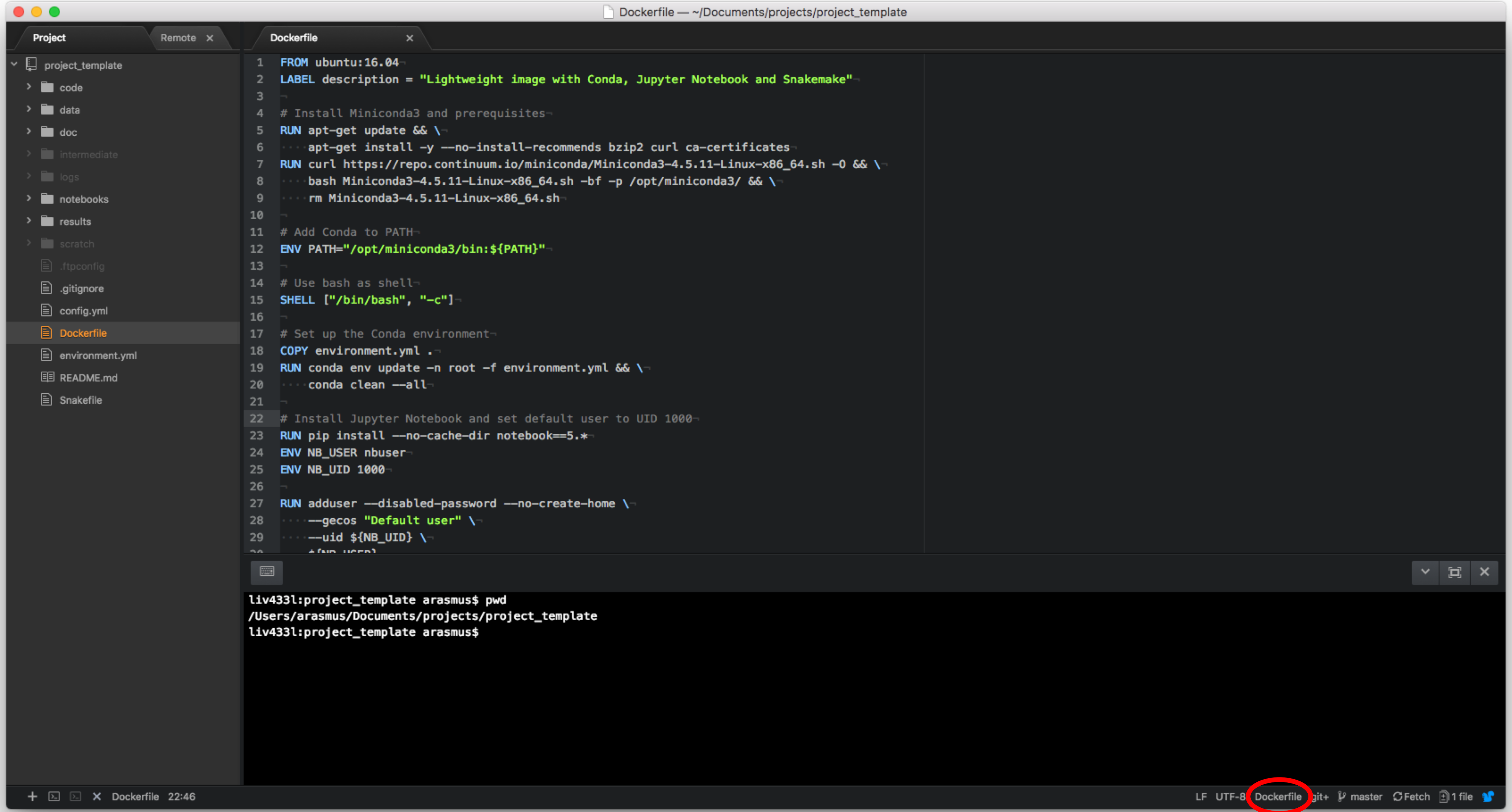
Bad examples:

- reproducible%20research.pptx
- suppl fig 10.png
- Supplementary Figure 9.png

Good examples:

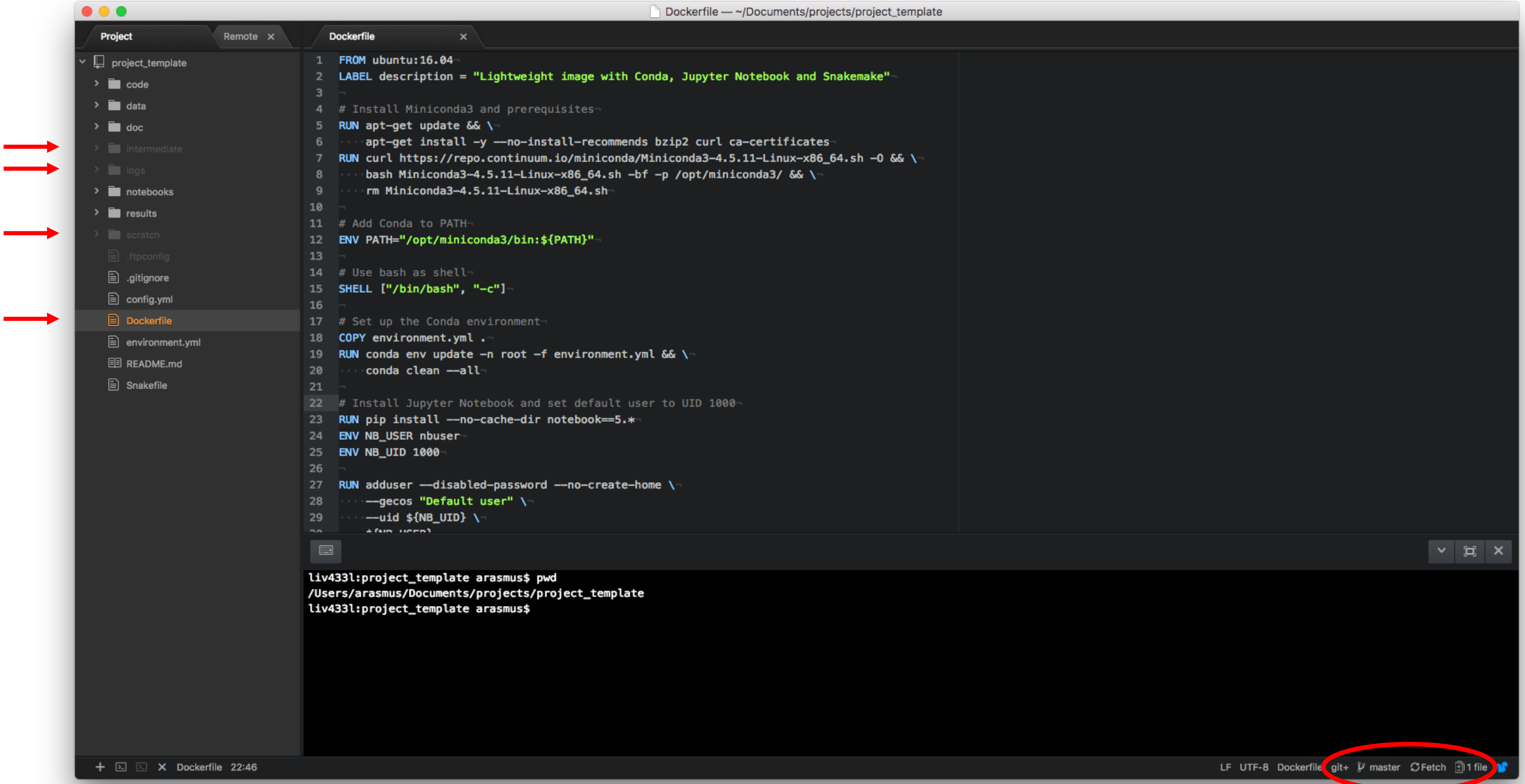
- 2018-11-28_Gothenburg_Reproducible_research.pptx
- suppl_fig_09_barplot_alignment_stats.png
- suppl_fig_10_samples_PCA_count_data.png

A project in Atom



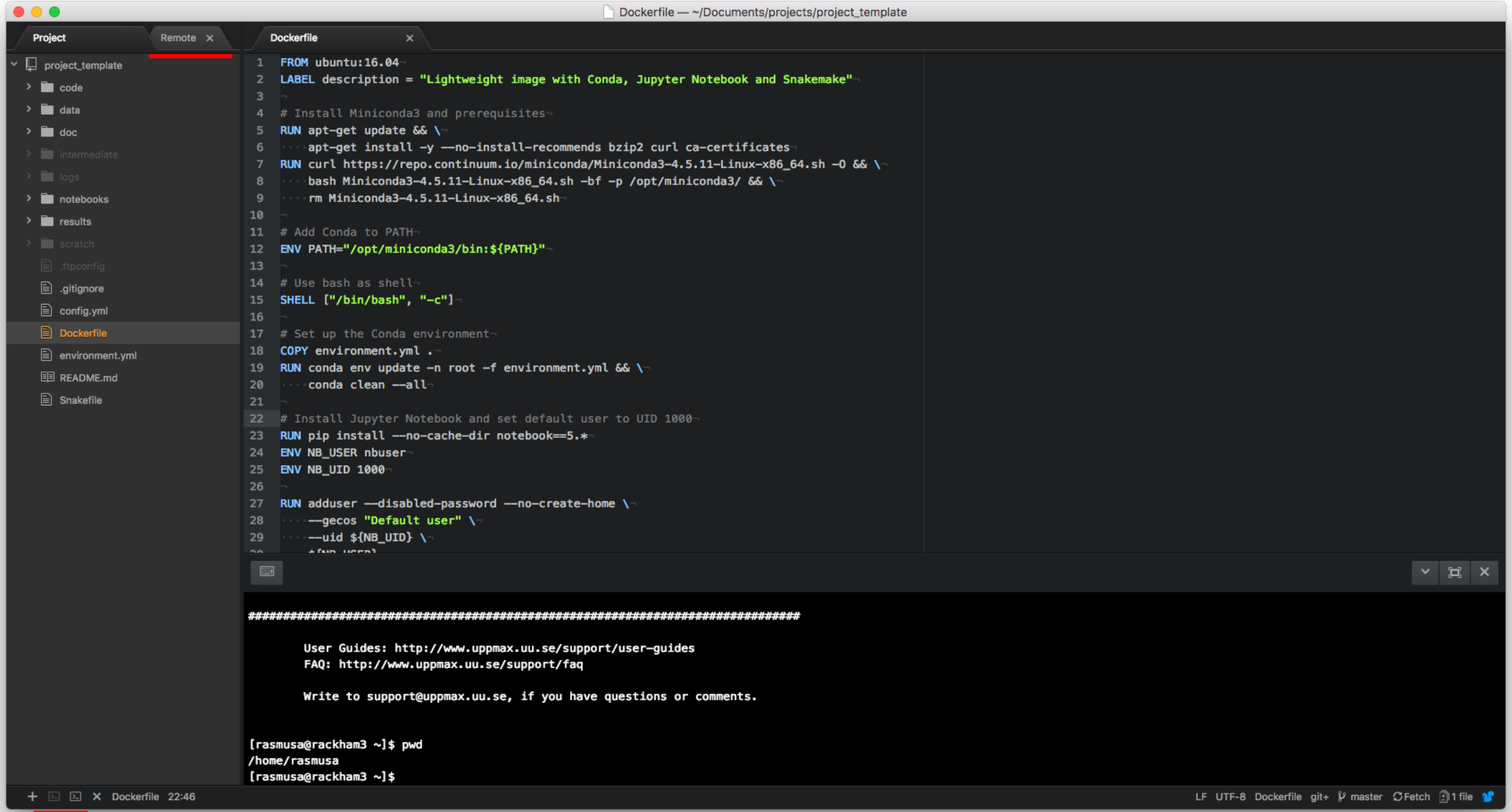
Syntax highlighting, indentation, and autocomplete

A project in Atom



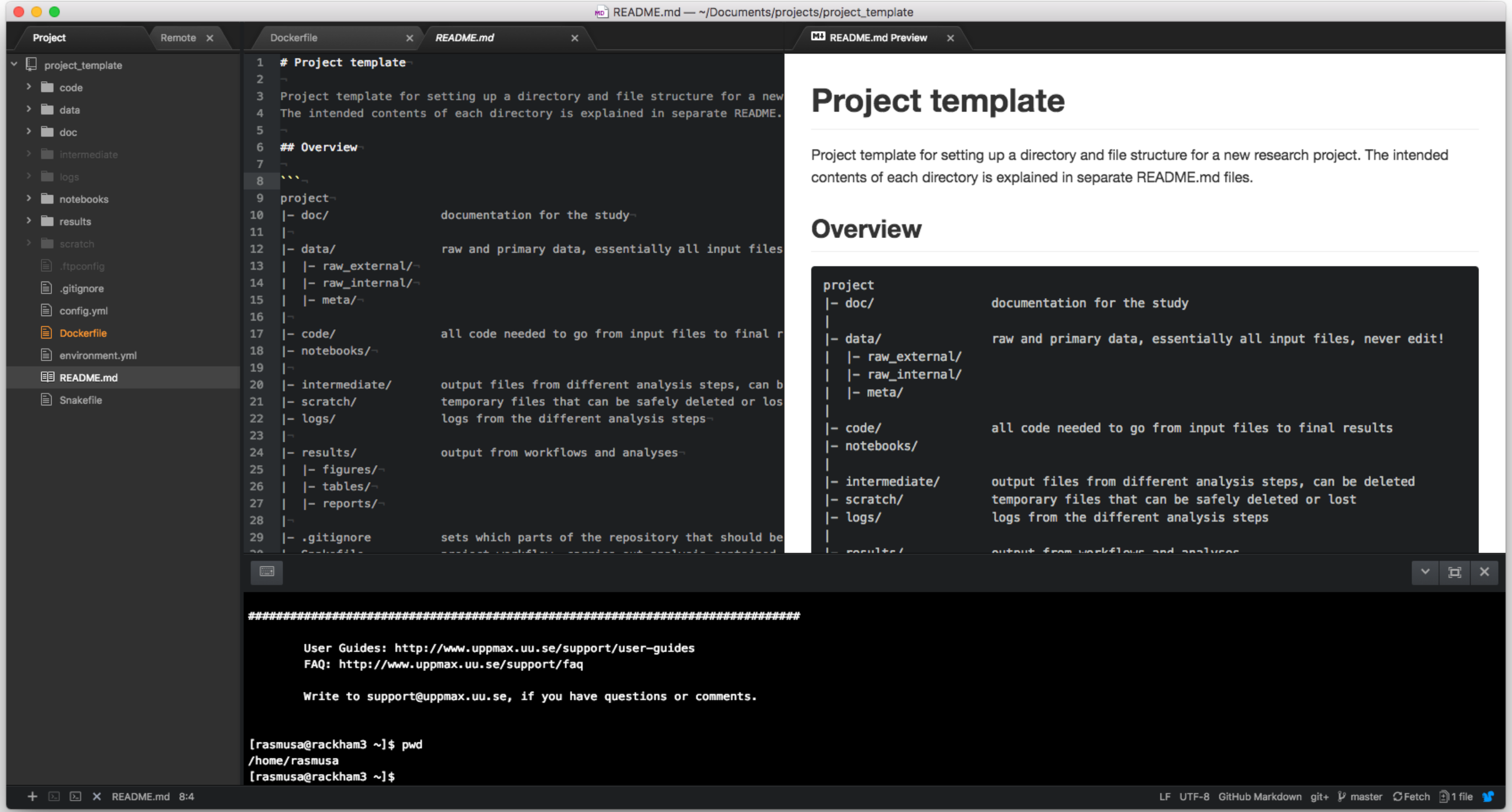
Integrated version control with Git

A project in Atom



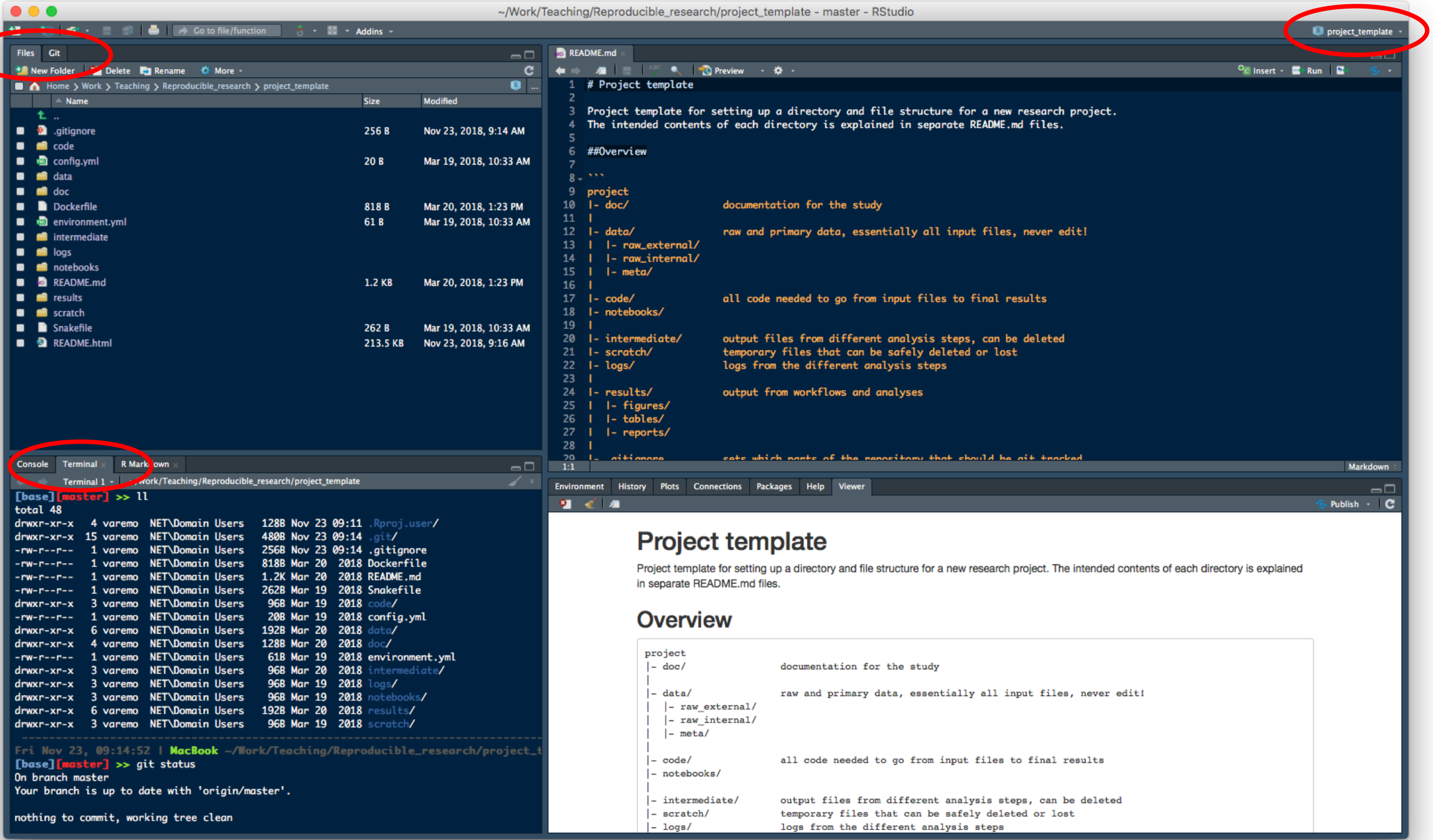
Automatically sync files between local/remote

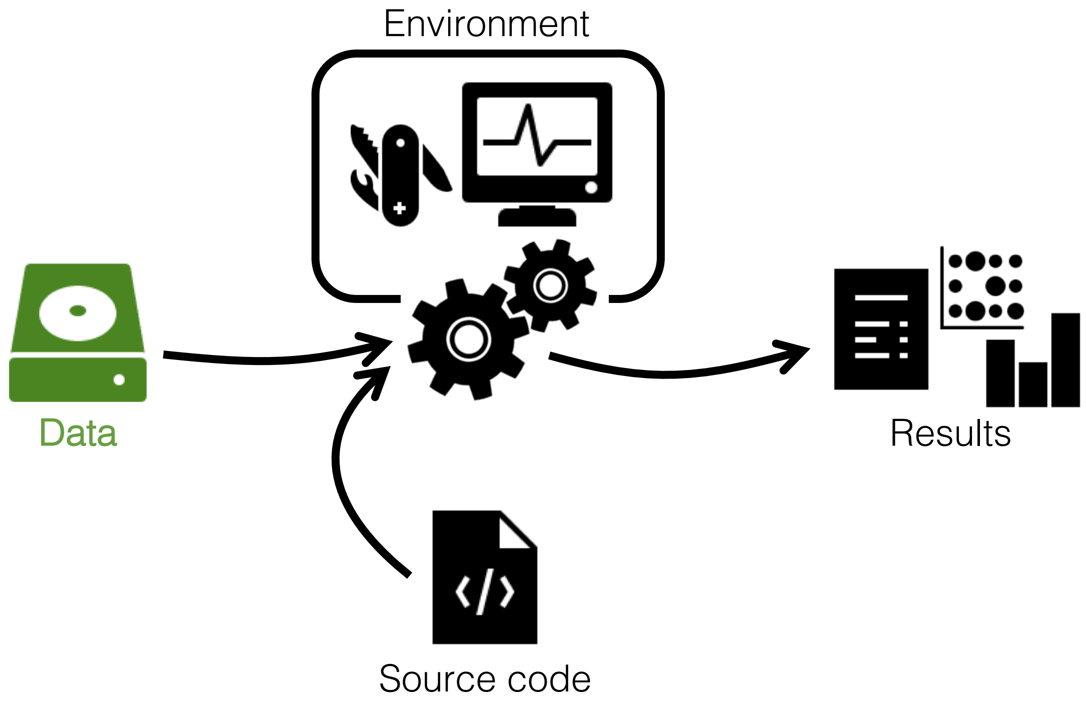
A project in Atom



Tons of plugins, e.g. for viewing different file formats

A project in RStudio






```
project
|- doc/
|
|- data/
|   |- raw_external/
|   |- raw_internal/
|   |- meta/
|
|- code/
|- notebooks/
|
|- intermediate/
|- scratch/
|- logs/
|
|- results/
|   |- figures/
|   |- tables/
|   |- reports/
|
|- Snakefile
|- config.yml
|- environment.yml
|- Dockerfile
```