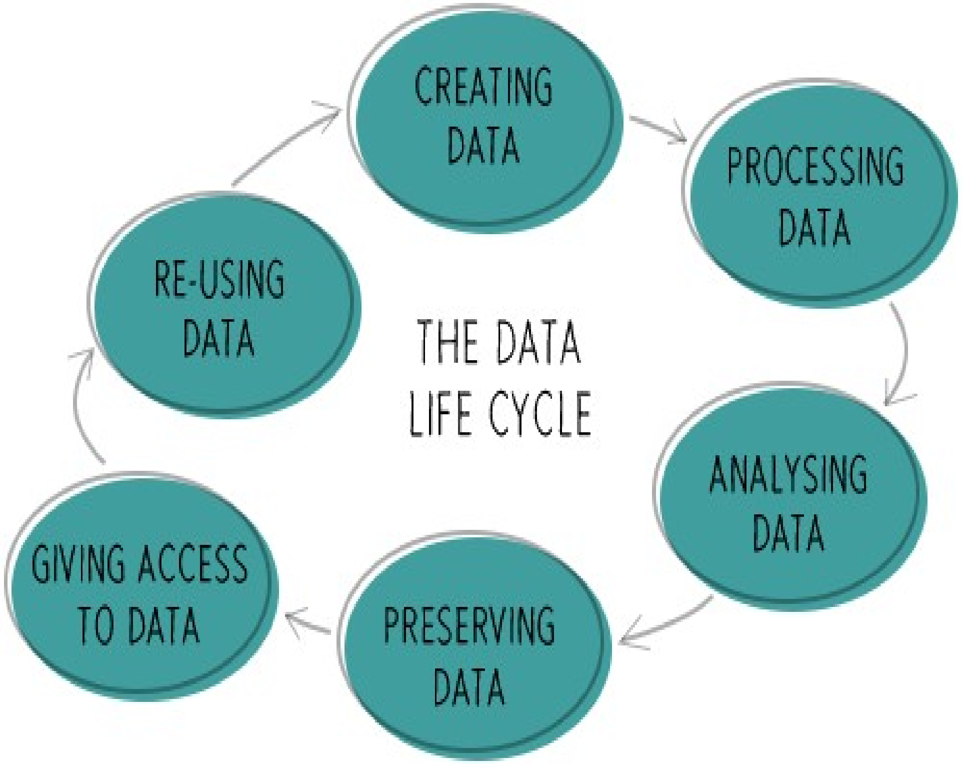
Research Data Management

Why do we meet?

- EPFL and funding agencies require a better way to deal with research data.
- Plenty of information in https://researchdata.epfl.ch

What is RDM?


CREATING DATA
PROCESSING DATA
ANALYSING DATA
PRESERVING DATA
GIVING ACCESS TO DATA
RE-USING DATA
THE DATA LIFE CYCLE

What is RDM?

- Research part: creating, processing analyzing data.
- Data management: storing, documenting, sharing, reusing data.

What is wrong?

- We loose time: many calculations could be avoided.
- Data exists but:
  - We don’t have access to it.
  - We don’t know enough about it.
  - We don’t have the tools to treat it (proprietary format).
  - ...
- Other people cannot reproduce the research we publish.
  - Computational details section is not enough.
  - Input geometries, all input parameters, software version…
- ...

What should we do?

1) Use open source software and format.

2) Describe extensively your data (metadata). It is never enough!

3) Publish your data.

4) Write data management plan (DMP) for each project.

Data should be FAIR

- Findable
- Accessible
- Interoperable
- Re-Usable

https://www.force11.org/group/fairgroup/fairprinciples

Open Source

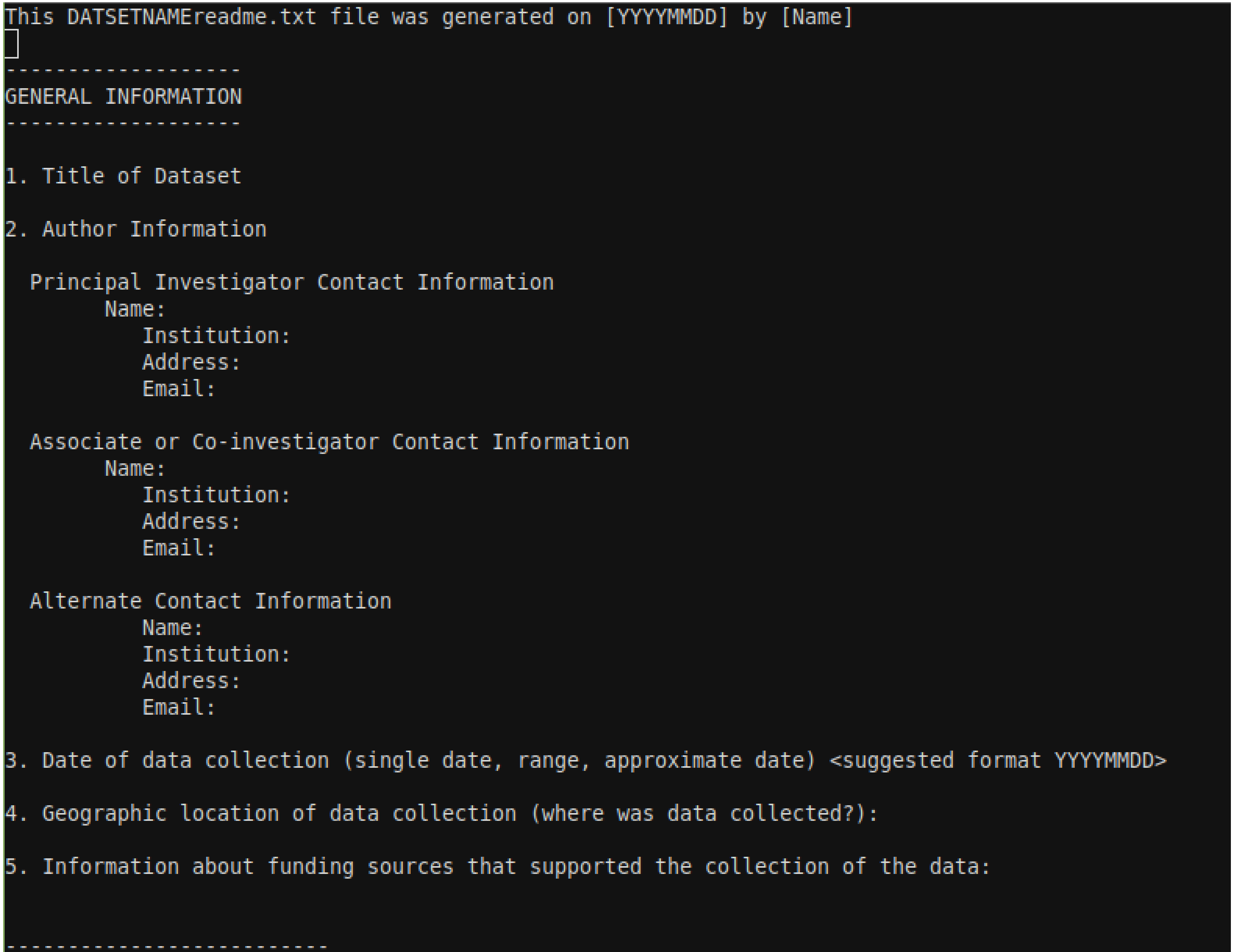
- Software:
  - If you write code or scripts: Learn git; make your source code opensource. People will cite you, they will use your code and improve it.
  - Funding agencies start to require the usage of opensource software.
  - It is just better for everyone!
- Format:
  - Data set: CSV, HDF5…
  - Code: plain text format (try using open source languages, e.g. Python instead of Matlab)
  - Images: .tif .png .svg
  - Movies: .mp4 .mj2 .avi .mkv
  - Text: .pdf .txt .odt .tex .md… (Do NOT use Apple or Windows format!)

From the List of EPFL Recommended File Formats: https://researchdata.epfl.ch/

Metadata

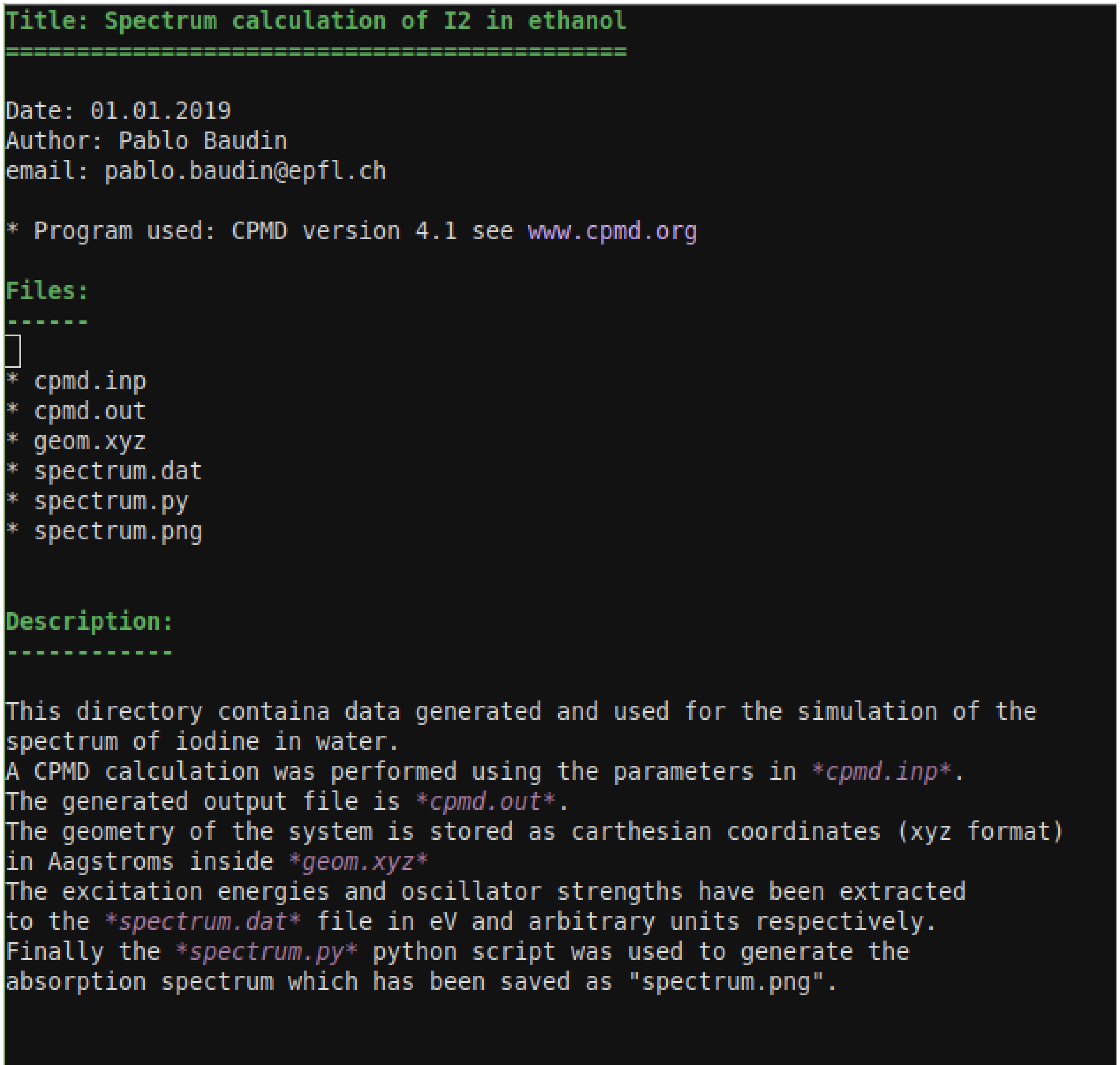
- Good file organization and naming conventions.
- Add README files to describe the content of each folder.
- Guide to “README” style metadata: https://data.research.cornell.edu/content/readme
- To go further:
  - Use tools like HDF5 format.
  - Or AiiDA: “ Infrastructure to manage, preserve, and disseminate the simulations, data, and workflows of modern-day computational science.” Developped at EPFL.

Readme template



```
This DATSETNAMEreadme.txt file was generated on [YYYYMMDD] by [Name]

-------------------
GENERAL INFORMATION
-------------------

1. Title of Dataset

2. Author Information

  Principal Investigator Contact Information
        Name:
           Institution:
           Address:
           Email:

  Associate or Co-investigator Contact Information
        Name:
           Institution:
           Address:
           Email:

  Alternate Contact Information
           Name:
           Institution:
           Address:
           Email:

3. Date of data collection (single date, range, approximate date) <suggested format YYYYMMDD>

4. Geographic location of data collection (where was data collected?):

5. Information about funding sources that supported the collection of the data:


--------------------------
```

Readme example



```
Title: Spectrum calculation of I2 in ethanol
============================================

Date: 01.01.2019
Author: Pablo Baudin
email: pablo.baudin@epfl.ch

* Program used: CPMD version 4.1 see www.cpmd.org

Files:
------

* cpmd.inp
* cpmd.out
* geom.xyz
* spectrum.dat
* spectrum.py
* spectrum.png


Description:
------------

This directory containa data generated and used for the simulation of the
spectrum of iodine in water.
A CPMD calculation was performed using the parameters in *cpmd.inp*.
The generated output file is *cpmd.out*.
The geometry of the system is stored as carthesian coordinates (xyz format)
in Aagstroms inside *geom.xyz*
The excitation energies and oscillator strengths have been extracted
to the *spectrum.dat* file in eV and arbitrary units respectively.
Finally the *spectrum.py* python script was used to generate the
absorption spectrum which has been saved as "spectrum.png".
```

Publish your data

- Make it open: “If other people can’t reproduce it, it’s not science…”
- Give a DOI to your data.
- Data repositories: Zenodo (a free and open digital archive built by CERN)
- Data journals:
  - Journal of Physical and Chemical Research Data (AIP)
  - Scientifc Data (Springer Nature Group)
  - Data in Brief (Elsevier)
- Others will use my data and publish before me!
  - The goal is to produce the best possible science as a community, not for personal glory.
  - Most of the time you will just be helping yourself…
  - They will have to cite your work!

Data Management Plan

The DMP describes:

- Strategies to:
  - Create, store, share, maintain, archive and preserve data throughout their life cycle.
- Which data are going to be produced.
- How each type of data will be:
  - Organized, classified, archived, shared, distributed, secured, preserved.

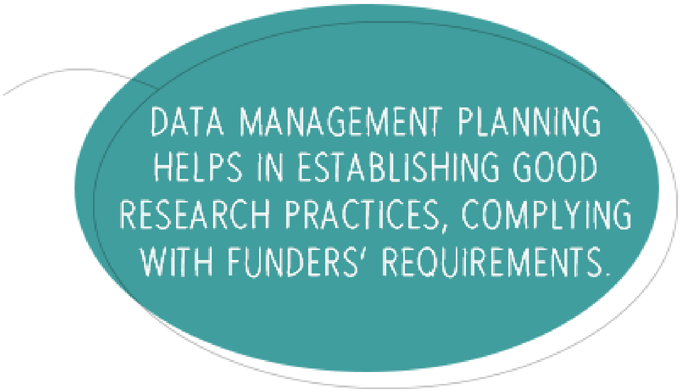
Why a DMP?

- **Plan**: future needs (hardware, software, HR, …)
- **Science**: better research reproductibility
- **Data reuse**: better use of public funds
- **Transparency**: public funded research available
- **Openness**: social impact of your research
- **Visibility**: citations, collabor., career
- **Compliancy** ...

Copy from a presentation of the EPFL Research Data Library Team.

Data Management Plan

- There are lots of guidelines and tools on how to write DMP.
- Sometimes we have to fulfill requirements from funding agencies.
- The DMP is a document that should be updated along with the the projects (we cannot know everything in advance).



DATA MANAGEMENT PLANNING HELPS IN ESTABLISHING GOOD RESEARCH PRACTICES, COMPLYING WITH FUNDERS' REQUIREMENTS.

Summary

- Use OpenSource as much as possible!
- Organize your data:
  - Meaningful file hierarchy.
  - Meaningful naming conventions.
  - OpenSource format (plain text, .odt, hdf5...)
  - **Add metadata (even just README files)!**
- Publish your data!
- **Start writing DMP** for all your projects (even incomplete) and update it regularly.
- Should we agree on a general format for REAMDE files and DMP?