Active research data management

Caterina Barillari
Scientific IT Services
The data life cycle

Active data management: Annotate, store, backup data while it is produced

Long term preservation: Annotate, store, backup data at the end of a project/publication
Why should we spend time on managing data?

1. Data without context not very useful, same for experimental/computational descriptions without data

2. For researchers:
   - Will you be able to find/reproduce your data in 6 months, 1 year?
   
     Having organized & documented data simplifies the process of writing papers and PhD thesis

3. For PIs:
   - Do you know where your PhD students leave their data? Do you understand what they did? Can someone else re-use the data and/or carry on the project?

4. Always have a backup in place!
What does it take to actively manage research data?

- Complex process that requires tracking and linking different types of information
1. Code management

1.1 Version control systems

- Software tools specialized on managing and documenting changes to source code over time.
- Used for managing large code bases.
- They are the standard in professional software development.

Tools

- [Subversion](https://subversion.apache.org)
- [GitLab](https://gitlab.ethz.ch)
- [GitHub](https://github.com)

- ID-SIS provides hands-on trainings on git for code management (info @ [https://sis.id.ethz.ch/consulting](https://sis.id.ethz.ch/consulting))
1. Code management

1.2 Interactive notebooks

- Applications that combine documentation, code, input and output generated by the code, e.g. graphs, plots (*Nature* 515, 151–152, 06 Nov 2014)
- Very useful for exploratory data analysis.

**IPython**: Open source
- Python only

**Jupyter**: Open source
- > 40 languages supported (Python, R, Julia, Matlab, IDL, etc)

**RStudio**: Open source + commercial edition
- Integrated development environment for R

**MATLAB**: Commercial
- Used in scientific, engineering, mathematical fields
2. Management of research data files

- Raw data
- Processed data
- Results
- Local HD
- Home dir. (NAS)
- Storage server
- Tapes
- HPC (Euler, CSCS)

Files/folders naming conventions

- Structured organization of data
- Data is annotated with metadata
- Searchable
- One central location
3. Experimental/process description

What to document?

- Goals
- Parameters used in experiment \((T, P, \text{time}, \text{etc})\)
- Experimental/computational procedure \((\text{code version, OS, compiler, etc})\)
- Results
- Links to data

How?

Text files

➢ Not scalable
➢ No sharing
➢ No efficient search
➢ Easy to use

Digital notebooks

➢ Scalable
➢ Sharing
➢ Search functionality
➢ Can link to digital data
➢ Rights management

Title
Date and Goals
Code and scripts
Methods
Analysis
Results
How can we bring everything together?

ETH ID-SIS develops and offers a DM solution to ETH research groups: openBIS.
openBIS for data management

(ETH Scientific IT Services)
openBIS overview

• Platform for managing scientific information. Purpose to support research data workflows from “bench” to publication

• Developed by ID-SIS since 2007

• Used in several ETH labs and facilities, but also at other Swiss and European universities. Also Novartis and Sanofi-Aventis

• Most use-cases in life sciences. However, openBIS underlying structure very generic ➔ amenable to be used in other disciplines
openBIS in a nutshell

Direct upload + Metadata registration

Workflow manager (e.g Snakemake)

Title
Date
Materials
Methods
Analysis
Results
How could openBIS be used at C2SM?

- **BigDataLink**: combines `git`, `git-annex` and **openBIS**.

- **openBIS**: central repository for storing metadata. Has links to data, which can be in different locations.

- Users access openBIS to get information about different (numerical) experiments and see where the data related to those experiments is.
Example: CH2018 project (Swiss Climate Change Scenarios 2018)

- EU-CORDEX (ESGF)
  - continuous updates
  - netcdf

- CORDEX raw
  - continuous updates
  - netcdf

- CORDEX freezes
  - v1.0
  - v2.0
  - v3.0

- Mean changes
  - .csv
  - .png

- Extremes/Indices
  - .csv
  - .png

- Localized data/downscaling
  - netcdf

≈50TB

≈10TB
openBIS workflow example: determine mean changes (CH2018)

- All info is in one central location accessible to all group members and PI (+ external collaborators)
openBIS as a service from ID-SIS at ETH

- From 2018, SIS has mandate to provide active data management services to all ETH

- Basic service for research groups:
  - Install and maintain openBIS on ID infrastructure.
  - Initial training.
  - Continuous support.
  - DMP support.

- Additional services (on demand)
  - Database customization.
  - Migration of existing databases.
  - Instrument integration for direct data upload.
  - Upload of existing historic raw data.
Acknowledgements

• openBIS development team:
  • Juan Fuentes, Franz-Josef Elmer, Antti Luomi, Yves Noirjean, Swen Vermeul

• openBIS support team:
  • Caterina Barillari, Cristian Scurtescu, Manuel Kohler

• SIS Section head:
  • Bernd Rinn

• SSDM Group Manager:
  • Matthew Baker

• Research Informatics Group Manager:
  • Thomas Wüst