Coriolis platform: Serving data through OpenDAP

Contribution to JRA3: ‘Facilitating the Re-use and Exchange of Experimental Data’

Authors: C. Bonamy, J. Chauchat, A. Mathieu, G. Moreau, J. Sommeria

Aims

As recalled in the deliverables D10.3 ‘Data Standards Report’ and D10.3’Date Repository Rules’, the Horizon 2020 programme requests that research data are open access, that is providing online access free of charge to the end-user and reusable. Furthermore access must allow the right to copy, distribute, search, link, crawl and mine the data. In addition to these general requests, we aim at achieving the following goals:

1) Allow the end user to scan and visualise the data without downloading.

2) Integrate the process in the data analysis procedure, with minimal additional work.

3) Make use of the data set by the project members themselves, in order to improve the team work organisation, and to verify the data set before final publication.

4) Reward and motivate the additional effort of the research team, by a recognised publication of the data themselves, in addition to the corresponding scientific articles resulting from the data analysis.

5) Rewarding the effort of our own institution by the visibility of relevant data sets.

At LEGI we are developing a software, called **Project-Meta**, to achieve these goals, and applied it to a few examples of data sets.

Beyond Zenodo

The use of Zenodo (https://zenodo.org) is prescribed in the reports mentioned above. Its advantage is the perennity of the storage and its support by European authorities. However, this system has several limitations:

-The data are uploaded once with limited possibility of update. So a preliminary system is useful to build the data set in a more **progressive** way and **verify** them before publication.

-The system is very general and handles various kinds of documents, which must be downloaded. There is no structured access to the data to allow online scanning and visualisation without downloading them.

-A data set is **limited to 50 Go** which may be problematic.

For these reasons we are developing our own data server, which does not preclude a final transfer on the more perennial support of Zenodo: our approach can be used as a way to prepare a final archive for publication on Zenodo. The tools developed are available at . http://servforge.legi.grenoble-inp.fr/projects/soft-trokata/wiki/SoftWare/ProjectMeta

The OPeNDAP protocol

We have chosen to display data with the protocol OPeNDAP (**Ope**n-source Project for a **N**etwork **D**ata **A**ccess **P**rotocol). This includes standards for encapsulating structured data, annotating the data with attributes and adding semantics that describe the data. OPeNDAP is widely used by governmental agencies such as NASA and NOAA to serve satellite, weather and other observed earth science data.

The protocol is based on http, so that data can be scanned with an ordinary web browser. However added functionality of data visualization is provided by graphics programs (like Matlab, GrADS, Ferret or ncBrowse). Compared to ordinary file transfer protocols (e.g. FTP) a major advantage using OPeNDAP is the ability to **retrieve subsets** of files, so it is possible to **work remotely** without downloading whole data files.

Although any file format can be use, data are often in HDF or NetCDF formats. The older NetCDF format is limited to arrays of numbers, while HDF provides wider possibilities of data structures (and it contains NetCDF as a particular case). We choose the NetCDF format which is sufficient for our applications and can be more easily read with a variety of softwares.

Examples of implementation

A DAP server has been implemented at LEGI http://servdap.legi.grenoble-inp.fr/opendap/. This has been used for a few years to publish data from ocean dynamic modeling. Extension to laboratory data, with an improved organization, has been recently undergone in the frame of FREE, using a software developed at LEGI, called **Project-meta**. A few examples are provided.

- A first example (http://servdap.legi.grenoble-inp.fr/opendap/meige/18PROJECT-META-TEST/) is a test containing as data a tutorial movie for the use of Project-meta.
- The example (http://servdap.legi.grenoble-inp.fr/opendap/hyrax/coriolis/14CARR) contains data from an access project performed during Hydralab IV. This project was about 'Internal mixing and near-bed dynamics induced by restricted stratified exchange flows', as it occurs in a river estuary.
- The project (http://servdap.legi.grenoble-inp.fr/opendap/hyrax/meige/15SHEET_FLOW) contains experimental data on sediment transport experiments (sheet flow) carried out in the LEGI tilting flume undertaken in the frame of JRA COMPLEX, as well as numerical data from corresponding numerical simulations.

In addition to the data themselves, each data set contains a file README.txt which gives a short description of the experiments and explains the data structure. Reference to the corresponding publications is given, as well as a link to a wiki page with more information. It contains also a file AUTHORS.txt listing the authors with their institution and e-mail address, a file COPYRIGHT.txt and a file LICENSE.txt. The chosen license could be Licence Ouverte v2.0, in agreement with the French law, but other licenses are possible (open-database-license-v1.0, creative-common-zero-v1.0) and this may evolve in the future… The LICENSE.txt file is just a pure copy of the license taken on the internet. The COPYRIGHT.txt file is a more generic file describing the project, the authors, the licence, the doi publication… It's an important entry point for question about copyright.

The data themselves are stored in disk spaces dedicated to each project, and the OpeNDAP folder contains only symbolic links to the appropriate data folders.

Construction of the data set

The procedure and associated programs are available in http://servforge.legi.grenoble-inp.fr/projects/soft-trokata/wiki/SoftWare/ProjectMeta.

To prepare an archieve on the OpenDAP server, the list of files to display, complemented by some information on the project, is prepared in a text file PROJECT-META.yml. This is a text file structured according to the language YAML (Yet Another Markup Language) more human readable than XML.

AUTHORS.txt and COPYRIGHT.txt could be generated automatically by extracting keys/values from PROJECT-META.yml metadata file and by using template (http://template-toolkit.org/) predefine files. These files and the LICENSE.txt file are mandatory and are part of the open dataset before publishing it to the Internet.

The process relies on a good structuring and documentation of the raw data, prior to the publication on OpeNDAP server. Three steps can be distinguished in the data flow from the experiments, as sketched in figure 1.

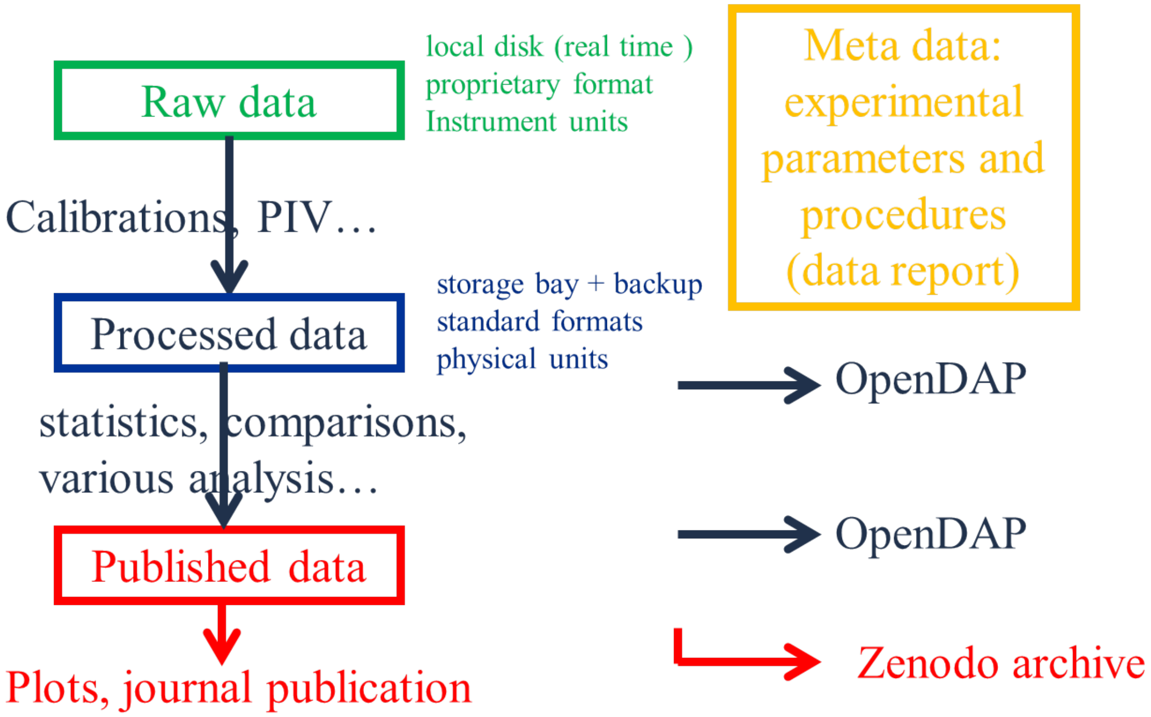

Figure 1: The three stages of data flow, from raw data provided by instruments, to processed data, and analysed data.

The data from instruments are first stored in a local disk, often in proprietary format imposed by instrument providers and by the constraints of real time data transfer. It then leads to processed data after calibration, Particle Image Velocimetry or other processing which are directly linked to the instrumental techniques. In a second stage, various scientific analysis can be performed, leading to the final published data. Those can and should be stored in standard formats, preferably NetCDF.

The first stage can be only performed by the experimental teams, as it requires a deep knowledge of the experimental conditions and instrument features. By contrast the second stage can be performed by external users, which could provide new methods of analysis or seek a close comparison with numerical simulation or theory. Therefore a publication of these data is useful. The minimum requirement is the publication of the data directly underlying the plot published in journal articles. More and more journals impose the publication of the original data in support of a published paper.

Data visualisation on the OpenDAP server

As stated above, an advantage of the OpenDAP protocol is to allow for data visualisation without downloading the data files. The idea is to allow the user to visualise and process data in the same way as if he was on his local disk.

We have developed two complementary approaches for this goal, such that the user can scan and visualised the data in a way which matches his expertise and needs.

1) **Scanning and visualising the file content with Matlab**. Some functions of Matlab are suitable to read data in the OpenDap server. However some adaptations are needed, in particular to get the list of available files. The toolbox with Graphic User Interface UVMAT http://servforge.legi.grenoble-inp.fr/projects/soft-uvmat/ has been adapted to scan and visualise images and netcdf files. For the user, the procedure is the same as for data stored on a local disk.

2) **Reading the files with Jupyter notebooks**. These notebooks provide a documented sequence of Python commands that can be run remotely. This allows the user to reproduce figures published in a related paper, starting from the raw data. The succession of processing operations is documented and can be edited by the user. The approach mainly relies on command lines although elementary Graphic User Interfaces are also provided. The Jupyter notebook runs on an external server, with a suitable installation of Python, so no software installation is needed by the user.