Master student Internship or Master thesis

Impact of research data curation

Start : January to July 2024
Length : 3 to 6 months
Place : LaBRI (Université de Bordeaux, France) or EPFL (Lausanne, Switzerland)
Supervisors : Guillaume Anciaux (guillaume.anciaux@epfl.ch), Aurélie Bugeau (aurelie.bugeau@labri.fr), Gaël Guennebaud (gael.guennebaud@inria.fr)
Salary : ~570 € /month

Context

Academic knowledge is traditionally disseminated by academic journals. However, nowadays the production of scientific data in any given project exceeds by a vast amount what can be contained in a few journal pages. Reproducible scientific data and publications must be associated to boost scientific collaborations and discoveries. There is today an always increasing pressure coming from universities and funding institutions towards publishing open data.

Naturally, computations and storage of scientific data come with an energy cost, which can in principle be matched with equivalent CO2 emissions. If it is envisioned to keep all the relevant numerical production of science in data archives, hence with the largest retention periods, it becomes an ecological and sustainability problem. However, there is no framework addressing the issue of scientific data storage: what amount of data is it acceptable to keep in the long term?

This question is difficult as it depends on the ratio production/storage costs, and on the practices of each scientific community. For instance, machine learning training sets are massively used, transferred and extended, making these datasets very dynamic. Some experimental datasets are particularly costly to produce but will not evolve in time and should therefore be kept in cold storage facilities. Finally, on the other extreme is the production of numerical simulations where the ratio time-to-compute/storage interacts with the frequency of accesses made by the targeted research groups.

Objectives

This project proposes to set the basis of a model describing this system. The important parameters are volume/cost of produced data, time/cost for data production, and frequency of access made by the targeted scientific community. Since all sorts of mediums can be used to store data with varying costs, it is planned to use several levels of storage each with associated CO2 emissions for storage, access/transfer, and renewing of equipment (only estimations of the CO2 costs can be obtained in general). Once the model is assembled, collecting data will be made to draw conclusions.
Access to the central computing clusters of EPFL will permit to collect the information on the calculation duration (CPU.hour) as well as the consumed storage. The data should be ordered per research groups and scientific disciplines to allow feeding the practices of research communities into the model. A detailed report of the collected data as well as the model

outcome should be produced. An implementation of this model should eventually be delivered in the form of a Python package, which should be used in emerging data curation tools such as the Solidipes (https://gitlab.com/dcsm/solidipes) project.

Candidate profile

- Curious, motivated
- Master student in computer science or related discipline