Data Acquisition, Extraction & Storage Course Project

IASD Masters Program 2023/2024 — PSL Research University

You need to form groups of three persons (or, exceptionnally, of two or four) to work on a project for this course. The project should be about **acquiring**, **extracting/ restructuring/cleaning**, and **storing** data, with the purpose of producing a clean, well-structured dataset. The dataset should be a good basis for further applications (e.g., used to train a classifier, or to build a graphical interface to navigate the data, etc.), but the project should stop at the production of the dataset.

The project should feature the following elements:

- One or several datasets should be acquired (from the Web, from social networks, from open data repositories, or from any other sources). There should be either some difficulty in acquiring the data (e.g., by crawling, by using an API) or if the dataset is readily available, several heterogeneous datasets should be used.
- The data should then be subject to transformations, extractions, restructuring, integration in order to produce a structured dataset.
- A data storage solution (a relational DBMS, a NoSQL DBMS, XML files, text files, etc.) should be proposed and the data should be put within this storage format. You should justify which data storage solution is used.
- The quality of the final dataset (missing or duplicate values, correction of data, etc.) should be assessed (e.g., by automatic or manual sampling methods).

The project will be due **on January 7**.