Seismic Waves to Marine Pulses: A Curation Pipeline for Building an Earth Sciences and Biodiversity Data Lake in the Portuguese Carabela Jellyfish and Seismology Studies"

Genoveva Vargas-Solar, CNRS, LIRIS
genoveva.vargas-solar@cnrs.fr
Jérôme Darmont, U. Lumière Lyon 2, ERIC
jerome.darmont@univ-lyon2.fr

Scientific position

The integration and fusion of data and metadata in the fields of life and earth sciences calls for the proposal of data and knowledge representations to structure the diverse information collected and produced for/within an experimental framework. Data lakes appear to be a relevant solution for managing and making available this diversity of data. Metadata models need to be devised to connect the data, and appropriate organisation and exploration mechanisms need to be devised that are relevant in the context of life and earth sciences.
The extraction of value through data-driven experiments in the life and earth sciences is determined by two main elements. (1) First, the maintenance of metadata collecting the conditions under which experiments are performed (quantitative perspective) to preserve the memory of the experimental process of knowledge production and to enable understanding and reproducibility. (2) Secondly, an open science perspective that can go beyond the sharing of data and must consider the sharing of know-how, decision-making, elements of expertise, project management and the people within projects who define the context in which experiments are carried out (qualitative perspective).

Context and Objective

The 4 months – 6 months internship will be associated with the activities of the project LETITIA (Lac de donnéEs, expérimenTation, vIe, Terre, curatIon, explorAtion) funded by the Federation of Informatics in Lyon in the laboratories LIRIS and ERIC. **The project focuses on designing and creating a data lake for gathering and integrating (meta)data on data-driven experiments in life and earth sciences.**

LETITIA aims to develop and implement a data lake that encompasses data, models, and the accumulated knowledge in these scientific fields. LETITIA addresses the curation, maintenance, and exploration of data collections within the data lake. The objective is to propose methods to navigate and enrich data collections, facilitating the generation of new data and analytical results. Data curation also entails recording the types of experiments conducted, their outcomes, and the conditions under which they were performed. Maintaining a catalog of questions and experiences related to the data supports the principles of open science, enabling the sharing of data and insights gained with the scientific community. All such information should be systematically stored within the data lake.

Expected Results

1. Instantiation of the metamodel goldMedal for characterizing seismic and biodiversity data associated with the collaboration of LETITIA with Brazil (*specification - contribution*)
2. Participation in the design and first implementation of the LETITIA data lake on ATLAS (configuration and software - *testbed*)
3. Proposal and implementation of a curation pipeline (protocol) for processing Earth sciences (seisms) and biodiversity (Portuguese Carabela jellyfish) experiments and collecting (meta)-data for maintaining the data lake (specification and code)
4. Demonstration on the detection and classification of seisms and on the classification of the Portuguese Carabela jellyfish.

N.B. An extensive amount of life and earth sciences data, drawn from various sources, is accessible in an integrated manner to enhance maintenance, analysis, and experimentation. This integration aims to concisely represent the discipline's knowledge, including vocabulary, concepts, and relationships.

Working method

- Technical and theoretical background provided by the tutors.
- Collective weekly discussion with tutors and the project consortium for designing solutions.
- On-line (slack/discord) or in-presence continuous interaction for technical and day by day discussion.
- Methodology and couching for the preparation of the final report and defense provided by tutors.

Working conditions

- ***Location***: the intern will work with the LETITIA consortium at laboratories ERIC and LIRIS in Lyon.
- ***Workload***: 35H weekly during 4 – 6 months according to the academic context of the intern who can be a master or an engineering student.
- ***Background knowledge***: databases (data models, querying, distributed databases - preferable), data processing techniques (preferable), good programming skills.
- ***Remuneration***: Internships +3 months in France paid in accordance with statutory rules for public research labs depending on the intern's academic level.