ETH ORD Programm

Contribute call 3 - deadline 12th December 2022

Number of projects submitted : 18 / Number of projects financed : 8 / 44% success rate

Project title	Abstract	Applicant	Institutions involved
Reprod, a web interface to assess reproducibility in experimental sciences	While recent reports have sparked intense discussion on the existence of a reproducibility crisis in life sciences, the accuracy of most scientific claims is rarely verified. Given this, we have started an ambitious project to assess the reproducibility of 400 articles published in my research field, Drosophila immunity, over a period of three decades. The goal is to provide a database open to the public that allows to better estimate the extent to which discoveries in this field that can be reproduced and trusted from the perspective of an experimental biologist. This project was funded by the SNSF supporting a full-time scientist to annotate the 400 articles and check if their claims have been reproduced. In the framework of an ETH Domain ORD Program, we plan to generate a web interface to present the major, main and minor claims of each article as well as assessments of these claims. This web interface will be interactive allowing members of the scientific community to comment on these articles and our analysis. This resource will provide easy access to summaries of supporting and contradictory data on a broad range of key topics in Drosophila immunity. The present proposal aims to support the realization of this web interface and to make it open access to allow its implementation in other research fields. To my knowledge, our project is unique and could clearly have an impact on today's science by providing a new tool to assess and discuss as a community reproducibility.	Lemaitre, Bruno	EPFL
Reinforcing open data analysis with spam: FAIR packaging and workshop organisation	We aim to make our popular open-source software 'spam' more accessible to the experimental mechanics community to exemplify and encourage good ORD practise in the community and avoid vendor lock-in with closed formats. We also hope to make a community of users with good practice for future joint developments within the ETH domain. We will do this by outsourcing some of the more complex software engineering to a service in EPFL, and also by organising a three-day workshop to promote, explain and illustrate ORD practise with 'spam'	Andò, Edward	EPFL
DaTabases and fRont-end wEb-bAsed Software for performance-based natUral hazaRds Engineering	The growing realisation that data are valuable to extract information for both research and design purposes in natural hazards infrastructure engineering is timely. This project aims first to signify the further development and curation of open access databases in a consistent format. Predictive models and probabilistic distribution functions that express damage of structural metallic materials and members will be formulated in an interactive web-based software that will also provide enhanced data visualization features. The data curation process, which will be in line with the FAIR principles, will strive to harmonize data formats in a way to maximize data re use within the research and engineering communities. Planned dissemination and maintenance strategies along with the development of comprehensive guidelines on how to standardize data storage and visualization will embrace the project's sustainability in the long term.	Lignos, Dimitrios	EPFL

Project title	Abstract	Applicant	Institutions involved
TSDF (Time Series DataFrame) - A data storage architecture for scalable processing of heterogeneous and geospatial time series	Geospatial time series data play a central role in environmental sciences, with example applications ranging from fixed sensors such as meteorological stations to mobile sensors. However, despite their ubiquity, methods to process and store geospatial time series datasets are often disperse and lack standard practices. A key reason for such a shortcoming is the inherent heterogeneity and irregularity of time series data. In this proposal, we aim to address these challenges by developing TSDF, a new data specification that provides a flexible framework to access, process, store and share geospatial time series datasets. A binary format based on Apache Parquet will be designed to provide flexible and effective hierarchical structuring of time series measurements, enabling scalable and distributed processing of irregular and heterogeneous datasets. Additionally, a Python package will be developed to provide an easy interface to load the datasets into pandas data frames, with methods to facilitate operations such as filtering, reducing and combining the data. The proposed work can greatly facilitate the processing of geospatial time series data in a vast body of applications, reducing data wrangling time for researchers while enhancing interoperability, reusability and collaboration.	Bosch, Martí	EPFL
Application Programming Interface for the Modern Ocean Sedimentary Inventory and Archive of Carbon database.	The increasing use of data repositories in marine geosciences has led to large, albeit dispersed and unstandardized datasets that require large amounts of time and effort to compile and harmonize. To overcome this, the Modern Ocean Sediment Archive and Inventory of Carbon (MOSAIC) database was devised to understand the factors that affect the distribution of the quantity, origin and reactivity of organic carbon in marine sediments, a key component of the global carbon cycle of growing interest with respect to carbon stocktaking and ecosystem services. Over the last year, MOSAIC has quadrupled its spatiotemporal coverage, increased by ten-fold the number of variables it contains, and it is currently stored as a PostgreSQL spatial relational database. The complexity of this expanding database requires it to be dynamically queried through a user friendly interface. Hence, this project aims to improve the accessibility to the database by creating an Application Programming Interface (API). This will be achieved by building an interactive and user-friendly web interface that allows users to query MOSAIC. Finally, Python and R packages will also be built that allow researchers to incorporate MOSAIC in their data analysis and processing workflows, ensuring reproducibility of their findings. This project will provide accessibility of the database to users with different backgrounds and interests, building on Open Research Data practices in marine geosciences	Paradis, Sarah	ETHZ
Open JMP - unlocking the potential of global indicator data	Decades of manual data structuring have resulted in the most comprehensive and internationally-comparable information on Water, Sanitation, and Hygiene (WASH) coverage. The WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene (JMP) maintains the database. The data are shared openly but in spreadsheet-based proprietary software, not following FAIR data principles. Data stored in spreadsheets underutilizes the potential those data could have for purposes other than the national, regional and global progress monitoring in WASH. We will approach this unused potential by developing open-source data and software packages that follow FAIR data principles to share the data within the WASH community and beyond. In the process, we engage with the community by hosting free learning events using open-source computational tools, enabling community members to further competencies aligned with FAIR data principles.	Elizabeth Tilley	ETHZ

Project title	Abstract	Applicant	Institutions involved
Building Open-Source Tools for reproducible interaction with biological ORD databases	The life sciences are increasingly reliant on access to and use of centralized online databases for exchanging biological information (e.g., NCBI, EBI, MGnify) (EBI, 2022; NCBI, 2022c; Mitchell et al., 2020). These databases allow researchers to share diverse primary (raw) and secondary (processed) biological datasets, allowing downstream re-use. However, significant rate-limiting steps to scientific discovery are (a) technical challenges for users depositing or withdrawing open research data (ORD) from these datasets, and (b) the poor control, traceability, and reproducibility of some steps involved in interacting with and downstream use of these ORD resources. Together, these reduce the reliability and reproducibility of scientific results, and slow adoption of ORD/FAIR (findable, accessible, interoperable, reusable (Wilkinson et al., 2016)) practices within the life sciences by increasing the activation energy required (Nelson, 2009; Tenopir et al., 2011; Wouters & Haak, 2017). We propose developing ORD tools (software) to facilitate remote, programmatic, and fully FAIR/reproducible interaction with several leading ORD resources that are commonly used in the biological sciences. These will eliminate current barriers to ORD sharing, (re-)use, and practices, and encourage community engagement in ORD practices. In this “contribute” project we will specifically contribute ORD tools for the microbiome research domain, but the mutli-disciplinarity of this field will position our project outcomes to translate to other research domains. This fits the criteria for the Contribute program, as we contribute software tools to facilitate interaction with and re-use of ORD from established ORD databases.	Nicholas Bokulich	ETHZ
Traceable thermodynamic datasets for chemical modelling	At present, thermodynamic datasets do not follow ORD FAIR principles. ThermoHub database provides access to a collection of traceable thermodynamic datasets for various fields of application. These datasets are curated and documented by experts using an open standard JSON format. The aim of the project is to bring ThermoHub to its full potential and demonstrate its ORD capabilities by producing a unified database of several mainstream thermodynamic datasets that are ready to use for chemical modeling. The project also aims to develop and provide a documented semi-automatic workflow for future maintenance and extension with new data. This work can greatly standardize and unify the workflow of chemical thermodynamic modeling, and support iterative improvement of the quality, reliability, and traceability of databases and modeling results. Providing datasets following FAIR principles will be advantageous when used in modeling, as it will remove the burden from modelers of collecting all necessary standard thermodynamic values from vast literature or writing complex scripts for importing these from different formats. ThermoHub can greatly streamline collaboration within Swiss, European, and other international projects by providing traceable thermodynamic data for various modeling applications. It will also be advantageous for the recognized work at PSI/LES (as well as EPMA and ETHZ) onthermodynamic database and modeling code development.	Miron, George-Dan	PSI