EGU22-10719, updated on 16 Aug 2022
https://doi.org/10.5194/egusphere-egu22-10719
EGU General Assembly 2022
© Author(s) 2022. This work is distributed under
the Creative Commons Attribution 4.0 License.

Virtual Labs for Collaborative Environmental Data Science

Maria Salama[1,3], Gordon Blair[2,1,3], Mike Brown[4], and Michael Hollaway[4]
[1]School of Computing and Communications, Lancaster University, Lancaster, UK
[2]UK Centre for Ecology and Hydrology (UKCEH), UK
[3]Centre of Excellence for Environmental Data Science, Lancaster University, UK
[4]UK Centre for Ecology & Hydrology (UKCEH), Lancaster, UK

Research in environmental data science is typically transdisciplinary in nature, with scientists, practitioners, and stakeholders creating data-driven solutions to the environment's grand challenges, often using a large amount of highly heterogeneous data along with complex analytical methods. The concept of virtual labs allow collaborating scientists to explore big data, develop and share new methods, as well as communicate their results to stakeholders, practitioners, and decision-makers across different scales (individual, local, regional, or national).

Within the Data Science of the Natural Environment (DSNE) project, a transdisciplinary team of environmental scientists, statisticians, computer scientists and social scientists are collaborating to develop statistical/data science algorithms for environmental grand challenges through the medium of a virtual labs platform, named DataLabs. DataLabs, in continuous development by UKCEH in an agile approach, is a consistent and coherent cloud-based research environment that advocates open and collaborative science by providing the infrastructure and software tools to bring users of different areas of expertise (scientists, stakeholders, policy-makers, and the public) interested in environmental science into one virtual space to tackle environmental problems. DataLabs support end-to-end analysis from the assimilation and analysis of data through to the visualisation, interpretation, and discussion of the results.

DataLabs draw on existing technologies to provide a range of functionality and modern tools to support research collaboration, including: (i) parallel data cluster services, such as DASK and Spark; (ii) executable notebook technologies, such as Jupyter, Zepplin and R; (iii) lightweight applications such as RShiny to allow rapid collaboration among diverse research teams; and (iv) containerisation of application deployment (e.g. using Docker) so that technologies developed can be more easily moved to other cloud platforms as required. Following the principles of service-oriented architectures, the design enables selecting the most appropriate technology for each component and exposing any functions by other systems via HTTP as services. Within each component, a modular-layered architecture is used to ensure separation of concerns and separated presentation. DataLabs are using JASMIN as the host computing platform, giving researchers seamless access to HPC resources, while taking advantage of the cloud scalability. Data storage is available to all systems through shared block storage (NFS cluster) and object storage (QuoBye S3).

Research into and development of virtual labs for environmental data science are taking part within the DSNE project. This requires studying the current experiences, barriers and opportunities associated with virtual labs, as well as the requirements for future developments and extensions. For this purpose, we have conducted an online user engagement survey, targeting DSNE researchers and the wider user community, as well as the international research groups and organisations that contribute to virtual labs design. The survey results are considered are feeding into the continuous development of DataLabs. For instance, some of the researchers' requirements include the ability to submit their own containers to DataLabs and the security issues to access external data storage. Other users have indicated the importance of having libraries of data science and data visualisation methods, which are currently being populated by DSNE researchers to be then explored in different environmental problems.