Toward a Reproducible Research Data Repository

Cheng-Jen Lee [1 *], ***Chia-Hsun Ally Wang*** [1], ***Ming-Syuan Ho*** [2], ***Tyng-Ruey Chuang*** [1,2,3 *]

[1*] *Institute of Information Science, Academia Sinica, Taiwan*
[2] *Research Center for Information Technology Innovation, Academia Sinica, Taiwan*
[3] *Research Center for Humanities and Social Sciences, Academia Sinica, Taiwan*
* Email: {cjlee, trc}@iis.sinica.edu.tw

Summary. The depositar (https://data.depositar.io/) is a research data repository at Academia Sinica (Taiwan) open to researchers worldwide for the deposit, discovery, and reuse of datasets. The depositar software itself is open source and builds on top of CKAN. CKAN, an open source project initiated by the Open Knowledge Foundation and sustained by an active user community, is a leading data management system for building data hubs and portals. In addition to CKAN's out-of-the-box features such as JSON data API and in-browser preview of uploaded data, we have added several features to the depositar, including sourcing from Wikidata for dataset keywords, a citation snippet for datasets, in-browser Shapefile preview, and a persistent identifier system based on ARK (Archival Resource Keys). At the same time, the depositar team faces an increasing demand for interactive computing (e.g., Jupyter Notebook) which facilitates not just data analysis, but also for the replication and demonstration of scientific studies. Recently, we have provided a JupyterHub service (a multi-tenancy JupyterLab) to some of the depositar's users. However, it still requires users to first download the data files (or copy the URLs of the files) from the depositar, then upload the data files (or paste the URLs) to the Jupyter notebooks for analysis. Furthermore, a JupyterHub deployed on a single server is limited by the machine's processing power which may lower the service level to the users. To address the above issues, we are integrating the BinderHub into the depositar. BinderHub (https://binderhub.readthedocs.io/) is a kubernetes-based service that allows users to create interactive computing environments from code repositories. Once the integration is completed, users will be able to launch Jupyter Notebooks to perform data analysis and visualization without leaving the depositar by clicking the BinderHub buttons on the datasets. In this presentation, we will first make a brief introduction to the depositar and BinderHub along with their relationship, then we will share our experiences in incorporating interactive computation in a data repository. We shall also evaluate the possibility of integrating the depositar with other automation frameworks (e.g., the Snakemake workflow management system) in order to enable users to reproduce data analysis.

Keywords. BinderHub, CKAN, Data Repositories, Interactive Computing, Reproducible Research.