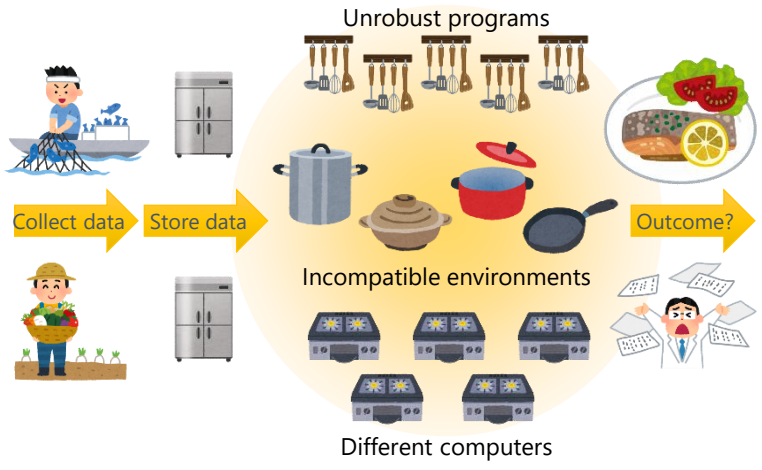
Integration of RDM and data analysis systems

Ikki Fujiwara, Yusuke Komiyama, Shigetoshi Yokoyama, Kazushige Saga, Atsuko Takefusa, Kento Aida, and Kazutsuna Yamaji
National Institute of Informatics

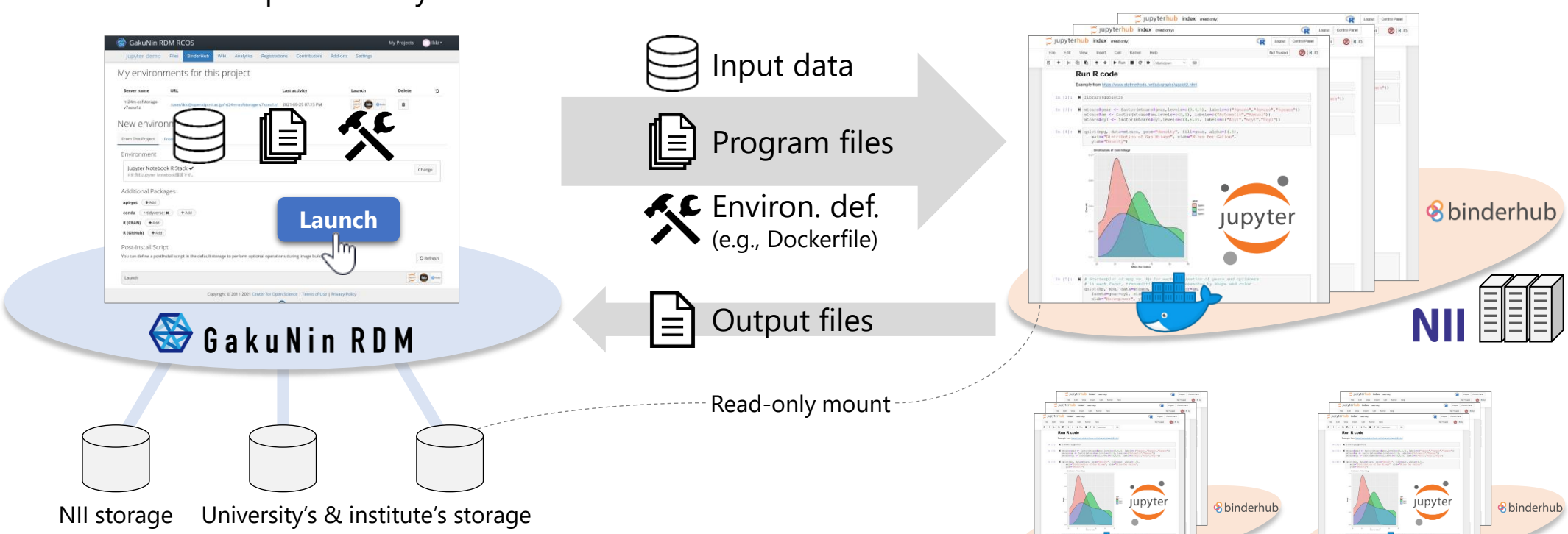
Motivation

Researchers sometimes try to reproduce the results of previous studies from open data. However, it could be hard to run the programs used in the original work in the exact same conditions. When rebuilding the runtime environment, researchers often find themselves in *dependency hell*. In order to ensure reproducibility and to realize the benefits of open science, research data management systems should treat software and its dependencies on par with data and papers.



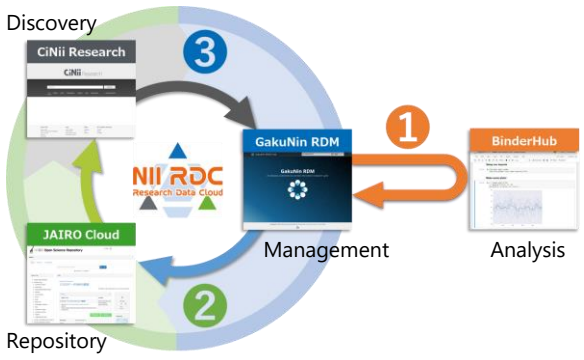
Solution

We integrate **Jupyter**[2] with **GakuNin RDM**[4], NII's research data management service built on top of the Open Science Framework[5]. By integrating the two, we provide researchers with an easy way to create, share, and reuse data analysis programs along with their runtime environments. Both system will be provided to the members of universities and institutes participating in **GakuNin**[3], the authentication and authorization infrastructure Federation in Japan. Unlike Google Colab or Code Ocean, we aim to provide a semi-persistent environment based on the trust provided by GakuNin.



Technology

We deployed **BinderHub**[1] on our on-premise Kubernetes cluster and made it compatible with GakuNin. We then developed a GakuNin RDM add-on that allows users to build and manage their own Jupyter containers. We also developed a Jupyter extension that writes output files back to GakuNin RDM. We will soon allow users to build their Jupyter containers on external computers, e.g., supercomputers at the users' universities or institutes.



Future Plan

Starting with ❶ **integration** of Jupyter with GakuNin RDM, we will develop a series of features to make software FAIR. Our next targets are ❷ **publication** and ❸ **duplication** of *research reproduction packages*, which encapsulate data, programs, and their runtime environments in a portable format. The source code we developed is available on GitHub[6].

Related Work

- Code Ocean. https://codeocean.com/
- GESIS Notebooks. https://notebooks.gesis.org/
- Gigantum. https://gigantum.com/
- Google Colaboratory. https://colab.research.google.com/
- MyBinder.org. https://mybinder.org/
- The Whole Tale. https://wholetale.org/

References

[1] BinderHub. https://binderhub.readthedocs.io
[2] Jupyter. https://jupyter.org/
[3] GakuNin. https://www.gakunin.jp/en
[4] GakuNin RDM. https://rcos.nii.ac.jp/en/service/rdm/
[5] Open Science Framework. https://osf.io/
[6] RCOS repository in GitHub. https://github.com/RCOSDP/

NII National Institute of Informatics
2-1-1 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN
Contact: Ikki Fujiwara, Assoc. Prof. <ikki@nii.ac.jp>
Publicated at eResearch Australasia 2021-10-14