Toward a Reproducible Research Data Repository

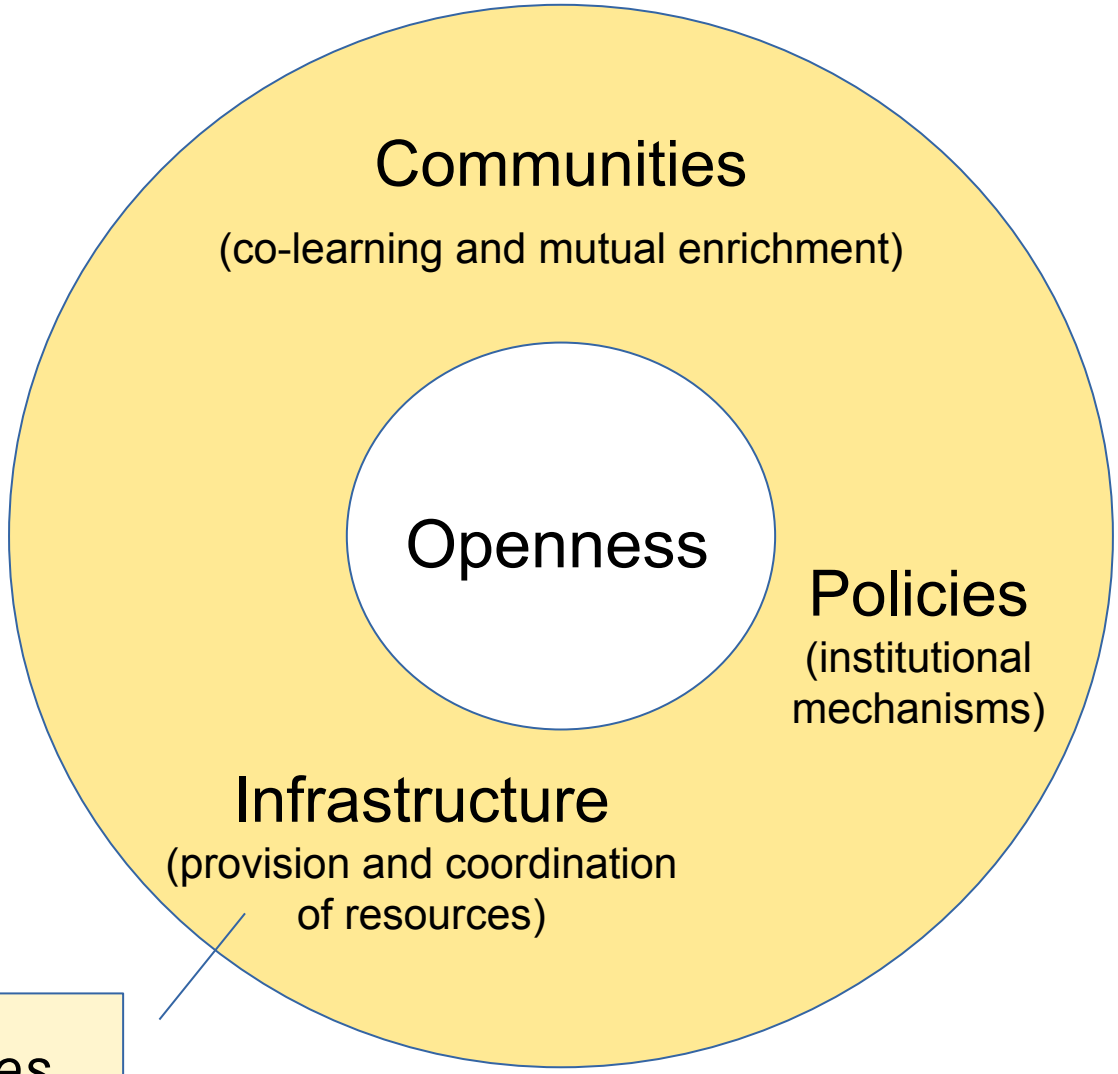
International Symposium on Data Science 2023 (DSWS-2023)
Asia-Oceania Data Forum
December 15, 2023

Tyng-Ruey Chuang
Institute of Information Science, Research Center for Information Technology Innovation,
and Research Center for Humanities and Social Sciences (GIS Center)
Academia Sinica, Taiwan

Cheng-Jen Lee
Institute of Information Science, Academia Sinica, Taiwan

depositar
研究資料寄存所

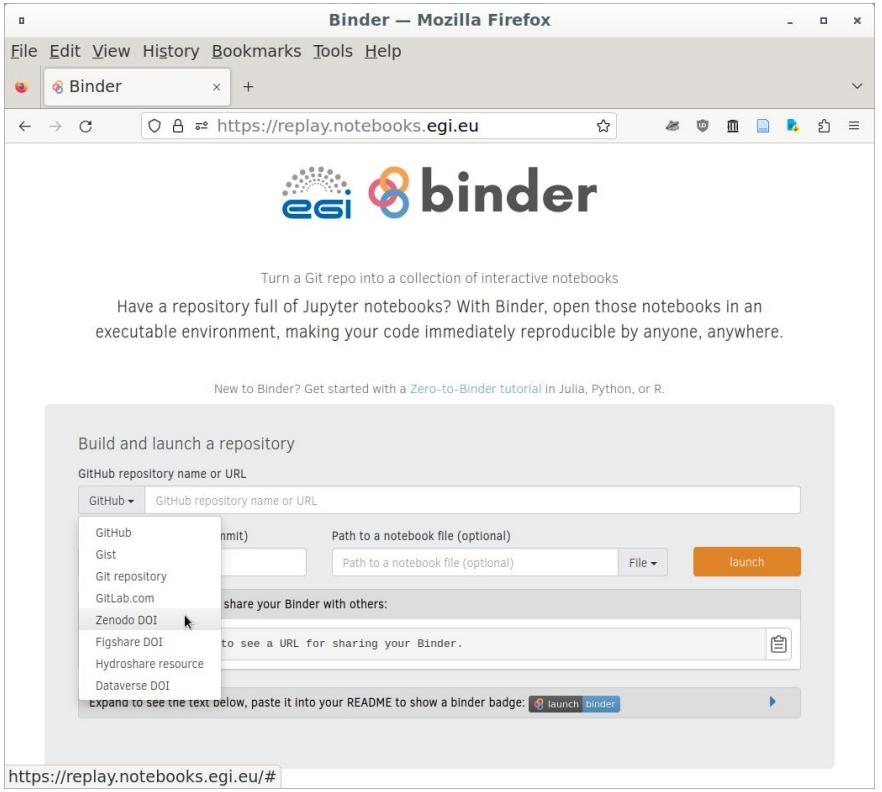
The Roles of Repositories in Open Source/Data/Science…



Repositories

Bringing Computation to Data Repositories

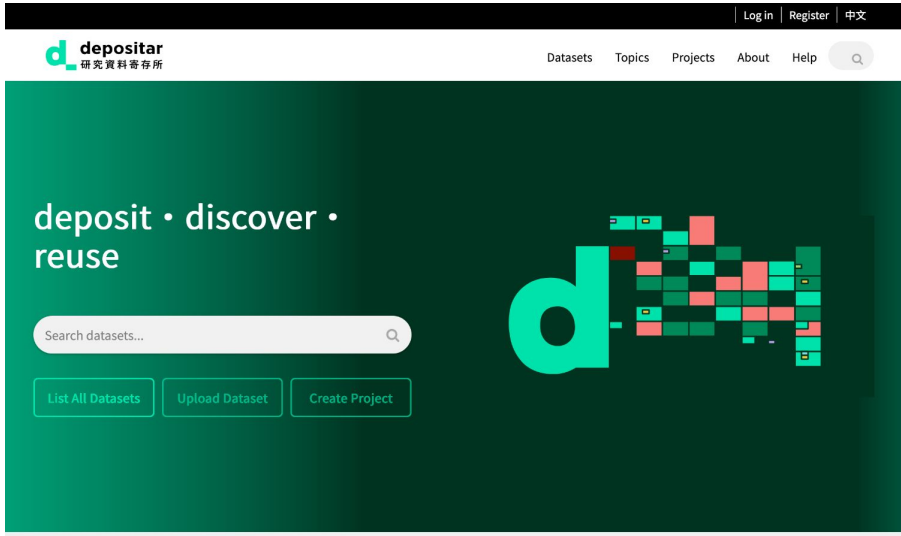
- Our inspirations
  - EGI Notebook and Replay
  - GESIS Notebooks
  - Code Package Function from NII (Japan)

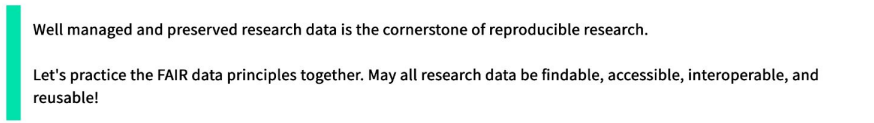


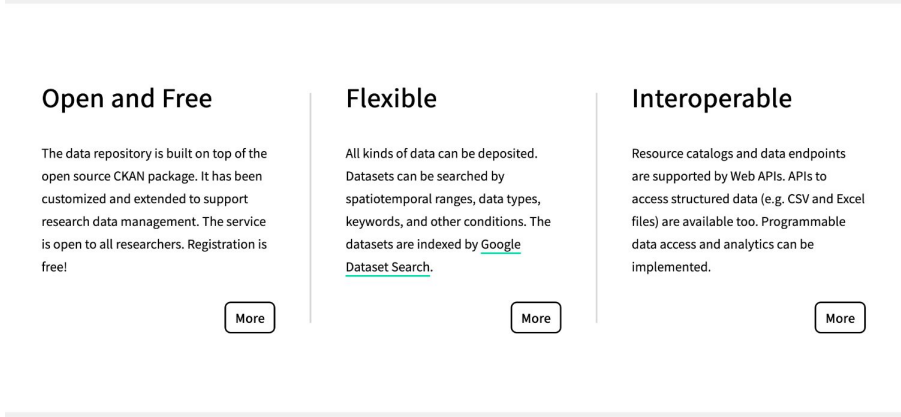
- Repositories for reproducible research
  - Article, code, and data are all open access and at the same place in a repository
  - Analytical results in the article can be automatically generated with the code and the data (by clicking a single button)
  - The repository coordinates computational and storage resources that are needed for the reproducibility of the results

About the *depositar* (研究資料寄存所)

- A data repository open to researchers worldwide for the deposit, discovery, and reuse of datasets since 2018
- Built on top of ckan, an open source data portal
- 1,900+ datasets & 340k+ views (as of Dec 2023)







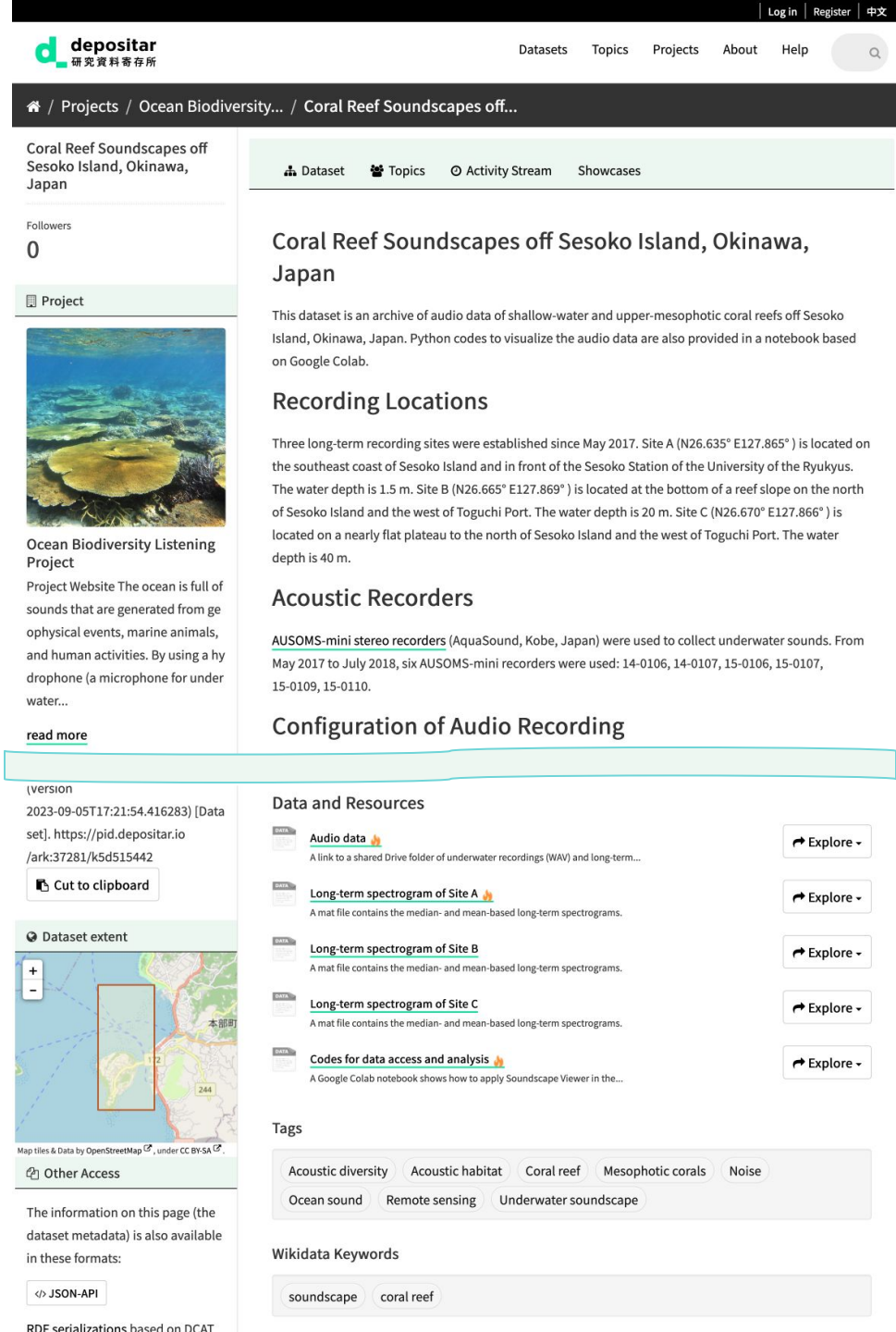
data.depositar.io

Learn more about the depositar: https://data.depositar.io/en/about

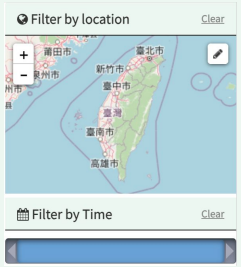
A Sample Dataset at Depostar

- Long description of dataset and project
- (deposited) data and (external) resources; descriptions
- W3C DCAT-based metatada
- Tags and Wikidata keywords

https://pid.depositar.io/ark:37281/k5d515442



Feature Highlights

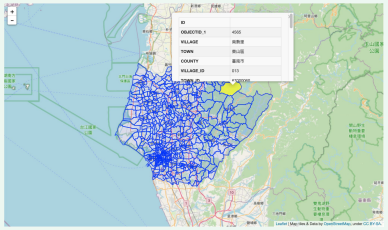


Spatio-temporal Search

Google site:data.depositar.io

Dataset Search

ARK Persistent Identifier



Data Previewers

W3C DCAT-based Metatada

WIKIDATA

Wikidata Keywords

depositar 研究資料寄存所

FAIR data since 2018

Project Management

/api/3/action/

JSON Data API

JSON-LD | XML | Turtle

RDF Serializations

jupyter BinderHub Integration

Dataset Citation CSL

ckan

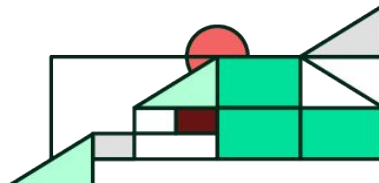
Open Data License Widget

Challenges to Online Interactive Computing

- Jupyter (Jupyter Notebook, JupyterLab)
  - A web-based interactive computing platform
  - Support many kernels (programming languages)

- JupyterHub
  - A multi-tenancy JupyterLab service for group of users

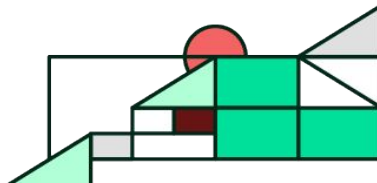
- The current JupyterHub service for specific data partners
  - Limited and inelastic processing power as a single server
  - The need for downloading the data from the repository and then uploading it to Jupyter



BinderHub Integration for Depositar

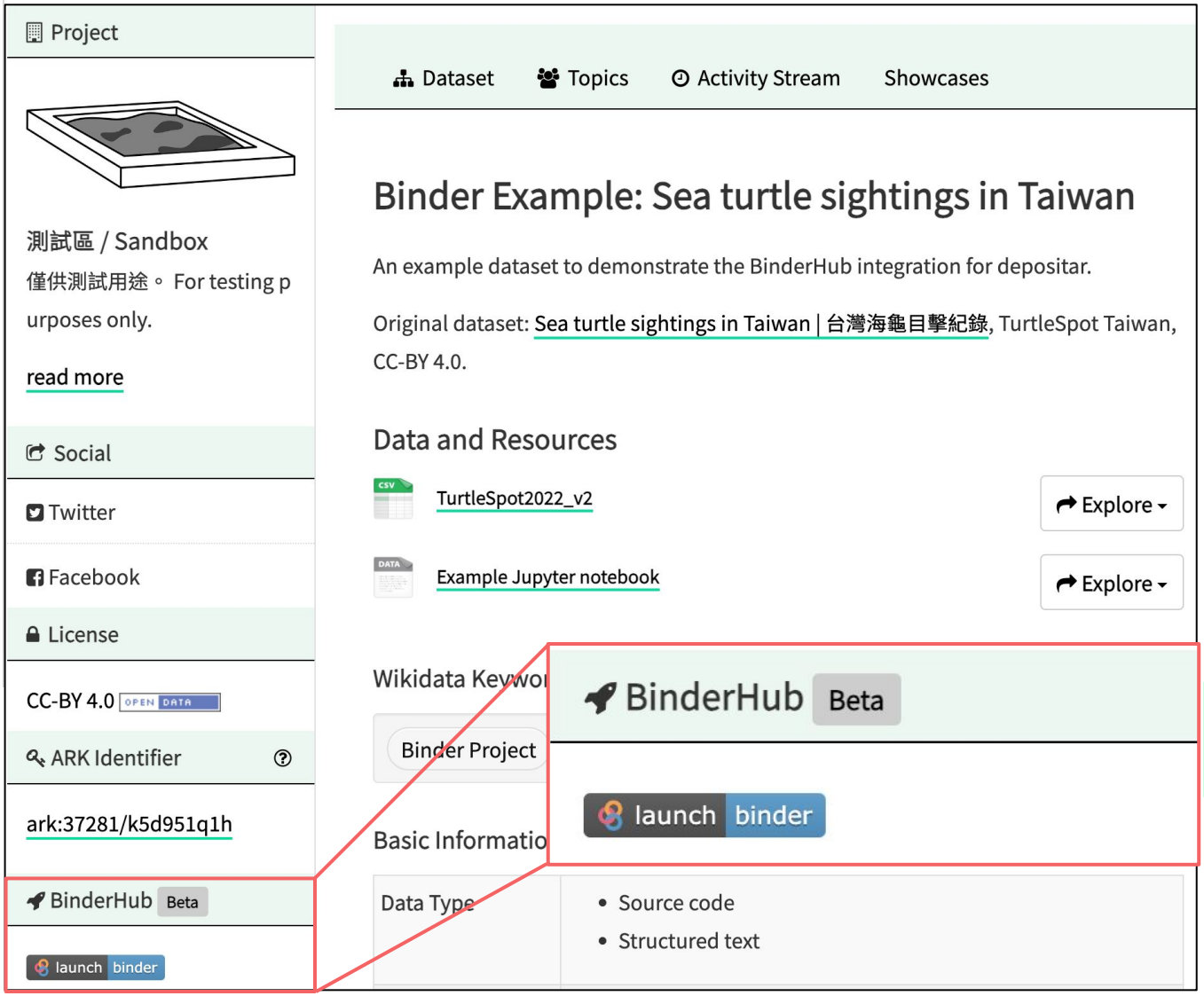
- BinderHub: establish a JupyterHub in the Kubernetes (k8s) environment to create Jupyter notebooks from data repositories such as GitHub, Zenodo, or Dataverse
  - MyBinder.org: a public Binder service

- BinderHub integration for *depositar* (available as of October 2023)
  - binder.depositar.io: a customized Binder service with a "CKAN provider" to support datasets on the *depositar*
  - Launch a Jupyter environment containing resources from a dataset by **clicking a button** on the dataset page without downloading the resources
  - Highly scalable thanks to the k8s technology



Demo

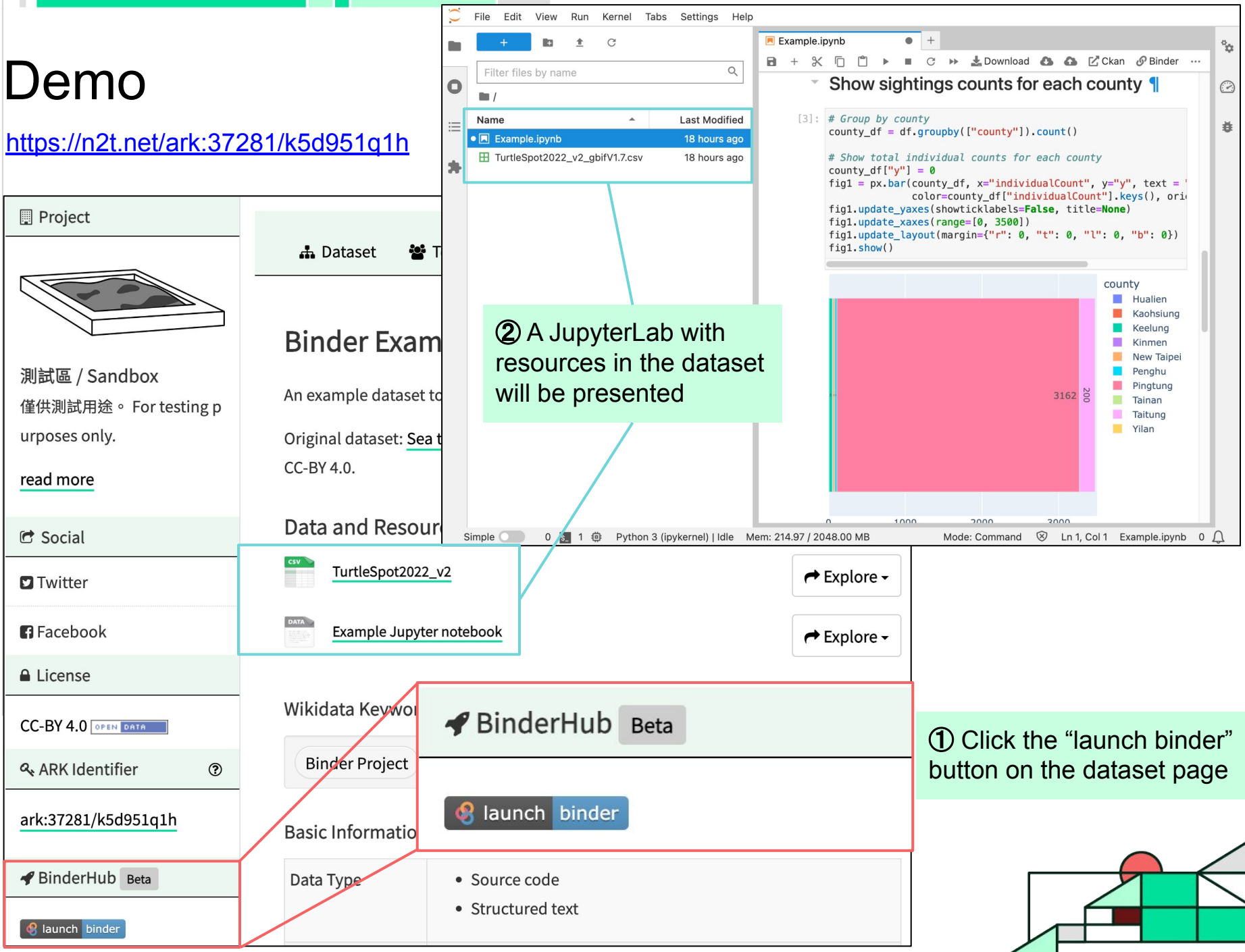
https://n2t.net/ark:37281/k5d951q1h



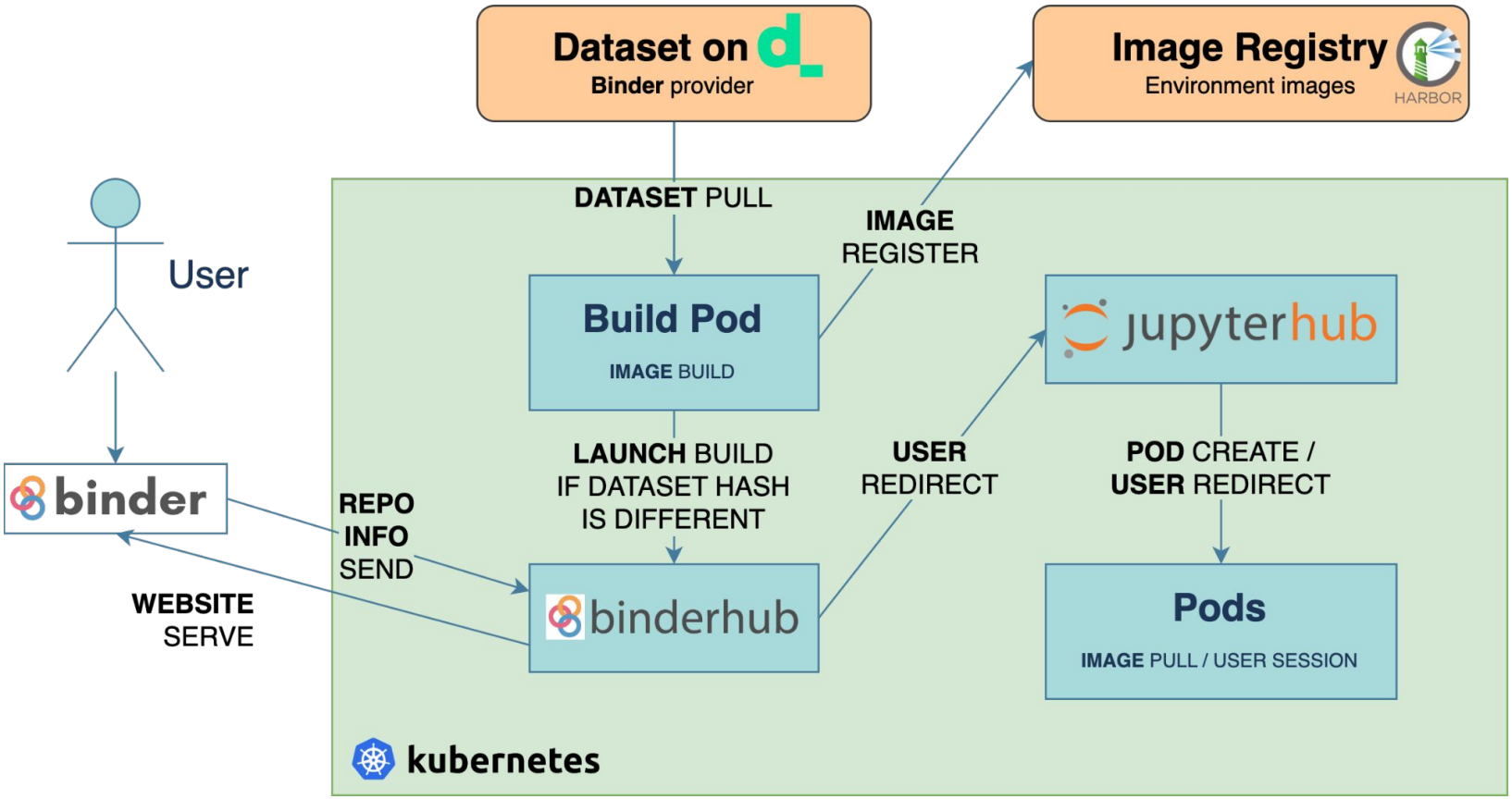
① Click the “launch binder” button on the dataset page

Demo

https://n2t.net/ark:37281/k5d951q1h



Architecture of the BinderHub Integration



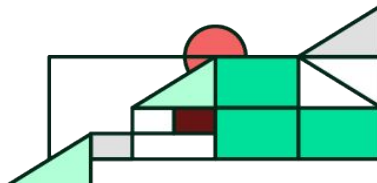
* Adapted from the diagram of the BinderHub architecture.
* Computing resource provided by Academia Sinica Grid Computing Centre, Grant No. AS-CFII-112-103.

Research Workflow Automation

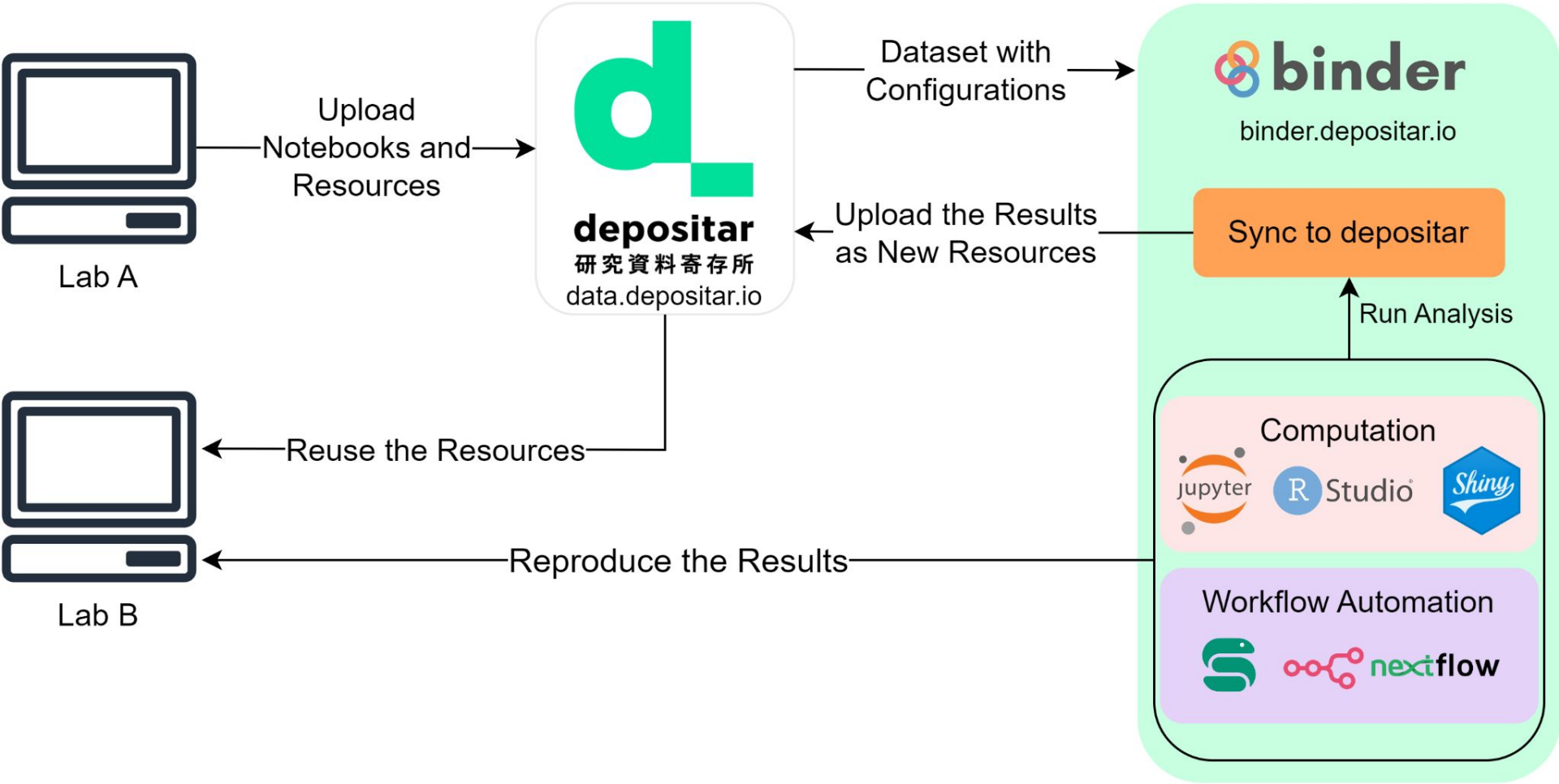
- Vanilla Jupyter is...
  - **IDEAL** for teaching or demonstration purpose
  - **NOT ENOUGH** for reproducing complex data analysis

- Workflow Management System
  - Run human-readable workflows (script) to conduct scalable, automatic, and portable data analysis
  - Candidates for consideration
    - Snakemake (Make + Python)
    - n8n (JavaScript)
    - Nextflow (Groovy)
    - CWL* implementations

* Common Workflow Language



A Preview of Depositar Binder Service (under development)



@_depositar

depositar
研究資料寄存所

ありがとう！ 謝謝！ Thank You!

https://data.depositar.io/ depositar
https://rdm.depositar.io/ RDM Hub

data.contact@depositar.io
https://lab.depositar.io/

The depositar is a collaboration at the Institute of Information Science, the Research Center for Information Technology Innovation, and the Research Center for Humanities and Social Sciences (GIS Center) in Academia Sinica, Taiwan. The project has been supported, in part, by grants from Taiwan's National Science and Technology Council.
The *depositar* project team: T-R Chuang, M-S Ho, C-J Lee & C-H Ally Wang.

「研究資料寄存所」是中央研究院資訊科學研究所、資訊科技創新研究中心、人文社會科學研究中心（地理資訊科學研究專題中心）的協作專案，部份經費來自台灣國科會的專題研究計畫。
研究資料寄存所計畫成員：莊庭瑞、何明諠、李承 鑫、王家薰。