

Toward a Reproducible Research Data Repository

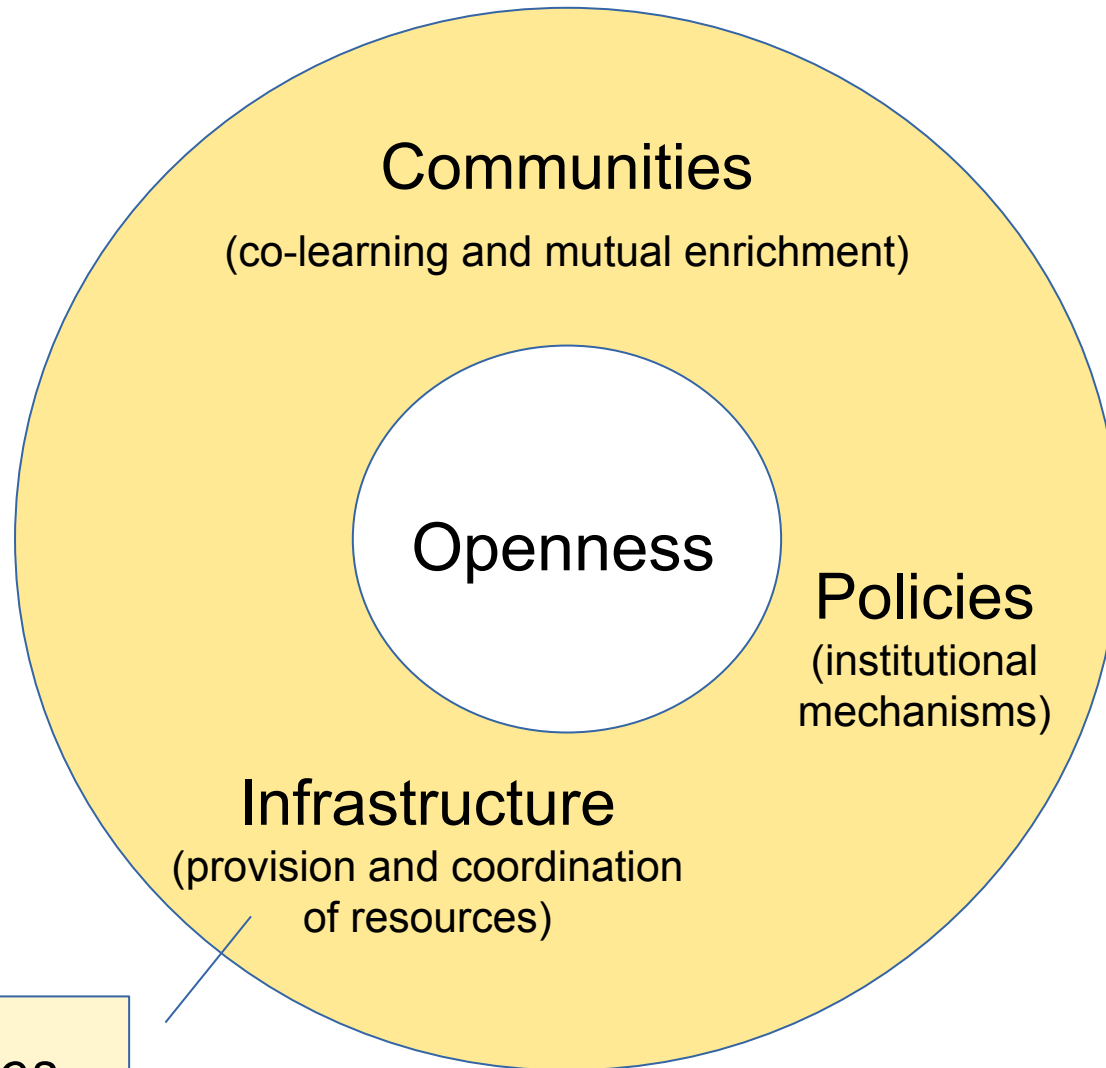
International Symposium on Data Science 2023 (DSWS-2023)
Asia-Oceania Data Forum
December 15, 2023

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

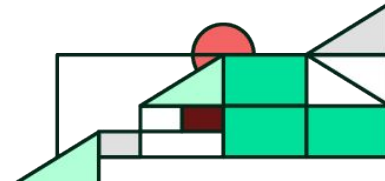
Cheng-Jen Lee
Institute of Information Science, Academia Sinica, Taiwan



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

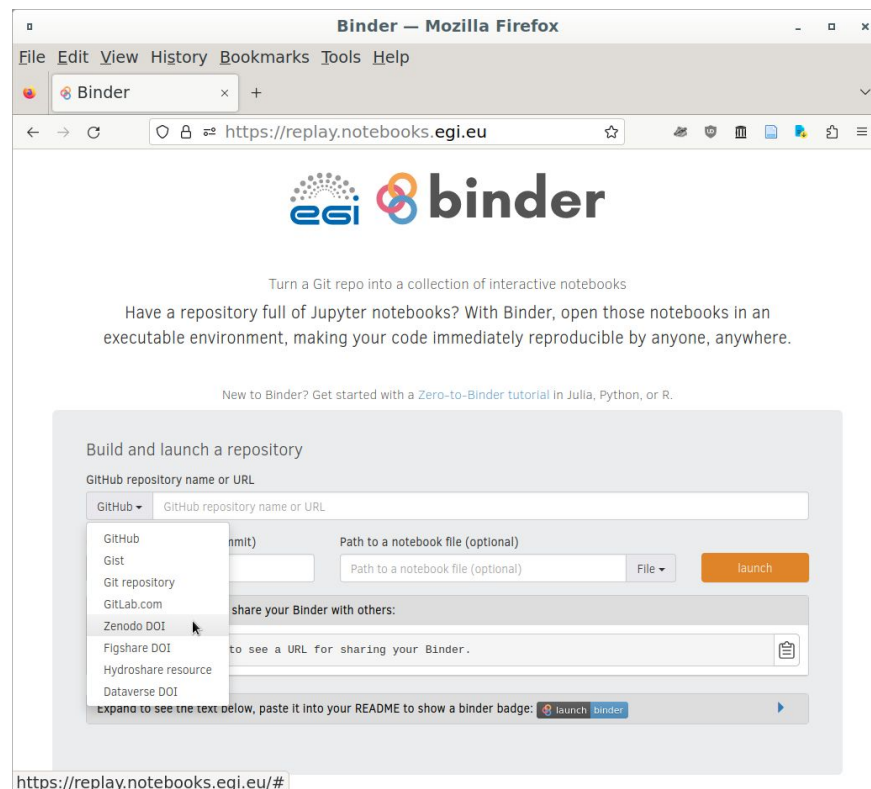


Repositories




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



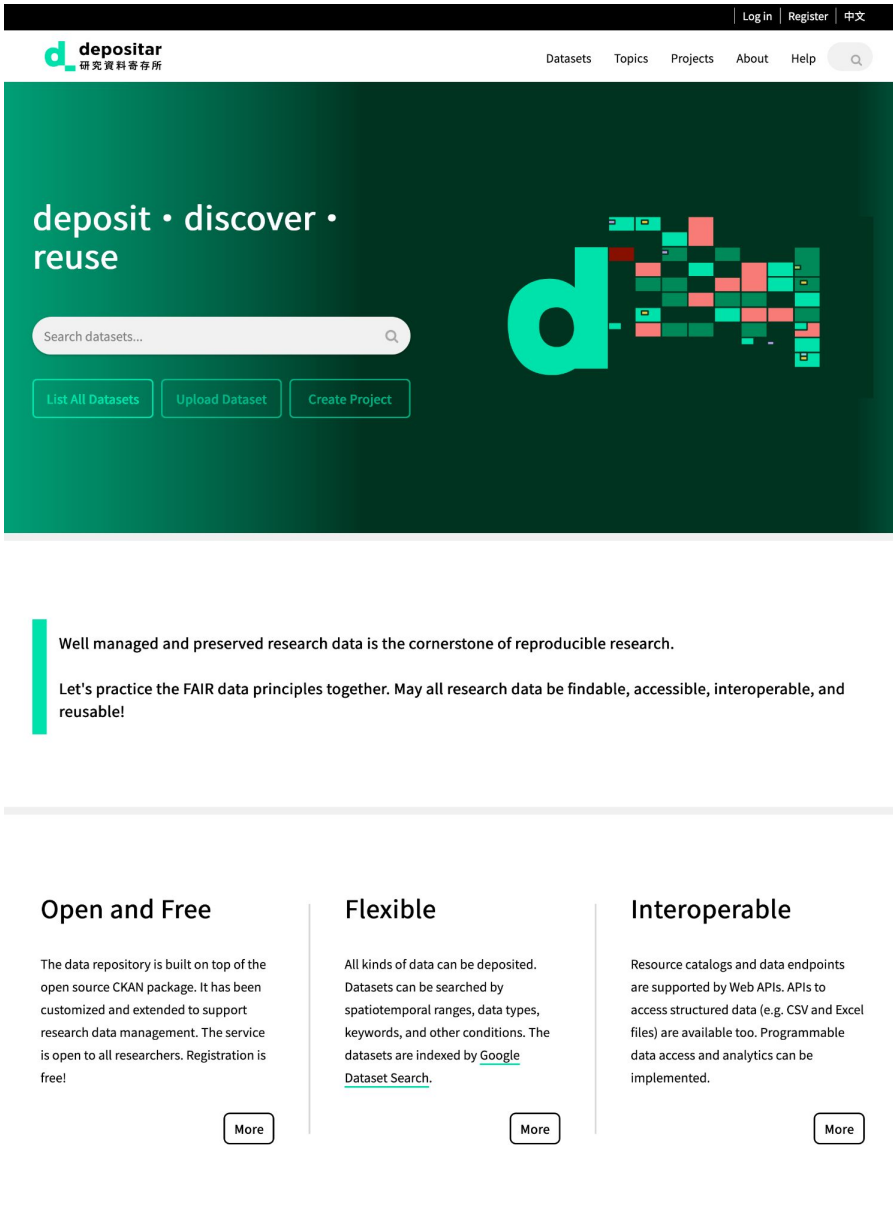
About the *depositor* (研究資料寄存所)

- 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.depositor.io

Learn more about the depositor:
<https://data.depositor.io/en/about>



The screenshot shows the homepage of the depositor website. The header includes the logo 'd depositor 研究資料寄存所' and navigation links for 'Datasets', 'Topics', 'Projects', 'About', and 'Help'. The main content area features the tagline 'deposit • discover • reuse', a search bar, and buttons for 'List All Datasets', 'Upload Dataset', and 'Create Project'. Below this, there is a section with the text: 'Well managed and preserved research data is the cornerstone of reproducible research. Let's practice the FAIR data principles together. May all research data be findable, accessible, interoperable, and reusable!'. At the bottom, there are three columns: 'Open and Free', 'Flexible', and 'Interoperable', each with a 'More' button.

Log in | Register | 中文

d depositor 研究資料寄存所

Datasets Topics Projects About Help

deposit • discover • reuse

Search datasets...

List All Datasets Upload Dataset Create Project

Well managed and preserved research data is the cornerstone of reproducible research.

Let's practice the FAIR data principles together. May all research data be findable, accessible, interoperable, and reusable!

Open and Free

The data repository is built on top of the open source CKAN package. It has been customized and extended to support research data management. The service is open to all researchers. Registration is free!

Flexible

All kinds of data can be deposited. Datasets can be searched by spatiotemporal ranges, data types, keywords, and other conditions. The datasets are indexed by [Google Dataset Search](#).

Interoperable

Resource catalogs and data endpoints are supported by Web APIs. APIs to access structured data (e.g. CSV and Excel files) are available too. Programmable data access and analytics can be implemented.

More More More

A Sample Dataset at Depostar

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

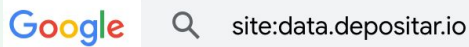
<https://pid.depositar.io/ark:37281/k5d515442>

The screenshot displays the Depostar website interface. At the top, there is a navigation bar with 'Datasets', 'Topics', 'Projects', 'About', and 'Help' links, along with 'Log in', 'Register', and '中文' options. The main content area is titled 'Coral Reef Soundscapes off Sesoko Island, Okinawa, Japan'. It features a 'Project' section with an underwater photograph of coral reefs and a description of the 'Ocean Biodiversity Listening Project'. Below this, there is a 'Dataset extent' section with a map showing the location of Sesoko Island. The 'Data and Resources' section lists several data items, each with an 'Explore' button: 'Audio data', 'Long-term spectrogram of Site A', 'Long-term spectrogram of Site B', 'Long-term spectrogram of Site C', and 'Codes for data access and analysis'. A 'Tags' section contains buttons for 'Acoustic diversity', 'Acoustic habitat', 'Coral reef', 'Mesophotic corals', 'Noise', 'Ocean sound', 'Remote sensing', and 'Underwater soundscape'. A 'Wikidata Keywords' section shows 'soundscape' and 'coral reef'.

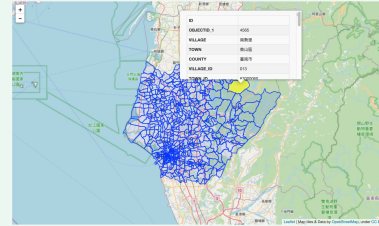
Feature Highlights



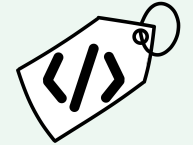
Spatio-temporal Search



Dataset Search



Data Previewers



W3C DCAT-based Metadata



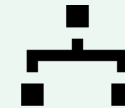
ARK Persistent Identifier



Wikidata Keywords



FAIR data since 2018



Project Management



/api/3/action/

JSON Data API



JSON-LD | XML | Turtle

RDF Serializations



Jupyter

BinderHub Integration

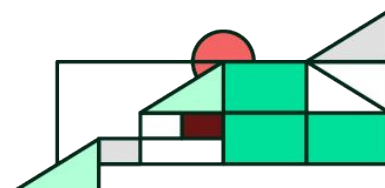
Dataset Citation 



ckan

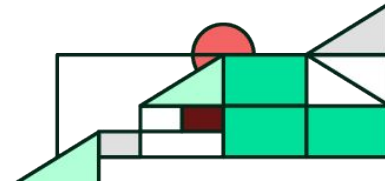


Open Data License Widget



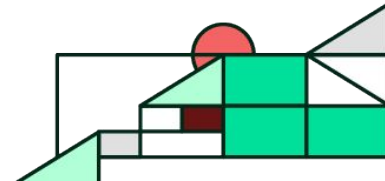
Challenges to Online Interactive Computing

- 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

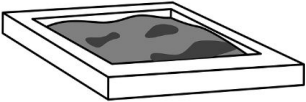
- [BinderHub](#): establish a JupyterHub in the Kubernetes (k8s) environment to create Jupyter notebooks from data repositories such as GitHub, Zenodo, or Dataverse
 - [MyBinder.org](#): a public Binder service
- **BinderHub integration** for *depositar* (available as of October 2023)
 - [binder.depositar.io](#): 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>

Project



測試區 / Sandbox
僅供測試用途。 For testing purposes only.

[read more](#)

Social

Twitter

Facebook

License

CC-BY 4.0 OPEN DATA

ARK Identifier ?

[ark:37281/k5d951q1h](https://n2t.net/ark:37281/k5d951q1h)

BinderHub Beta

[launch binder](#)


Dataset Topics Activity Stream Showcases


Binder Example: Sea turtle sightings in Taiwan

An example dataset to demonstrate the BinderHub integration for depositar.

Original dataset: [Sea turtle sightings in Taiwan | 台灣海龜目擊紀錄](#), TurtleSpot Taiwan, CC-BY 4.0.

Data and Resources

 [TurtleSpot2022_v2](#) [Explore](#)

 [Example Jupyter notebook](#) [Explore](#)

Wikidata Keyword [Binder Project](#)

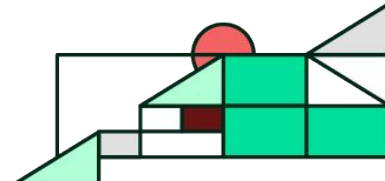
[launch binder](#)

Basic Information

Data Type

- Source code
- Structured text

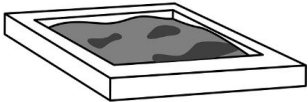
① Click the “launch binder” button on the dataset page



Demo

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

Project



測試區 / Sandbox
僅供測試用途。For testing purposes only.

[read more](#)

Social

- Twitter
- Facebook

License

CC-BY 4.0 [OPEN DATA](#)

ARK Identifier [?](#)

[ark:37281/k5d951q1h](https://n2t.net/ark:37281/k5d951q1h)

BinderHub Beta

[launch binder](#)

Dataset



Binder Exam

An example dataset to

Original dataset: [Sea](#)

CC-BY 4.0.

Data and Resource

-  [TurtleSpot2022_v2](#)
-  [Example Jupyter notebook](#)

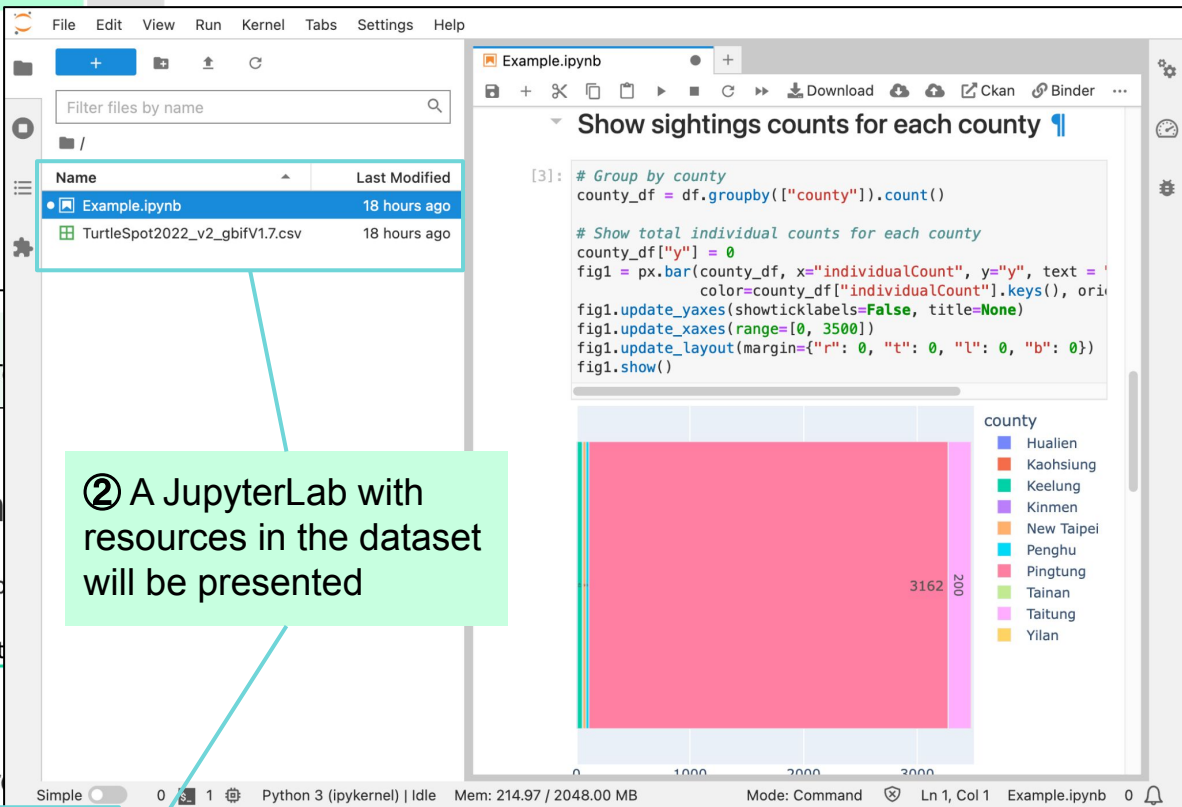
Wikidata Keyword

[Binder Project](#)

Basic Information

Data Type

- Source code
- Structured text



File Edit View Run Kernel Tabs Settings Help

Filter files by name

Name	Last Modified
Example.ipynb	18 hours ago
TurtleSpot2022_v2_gbifV1.7.csv	18 hours ago

```
[3]: # Group by county
      county_df = df.groupby(["county"]).count()

      # Show total individual counts for each county
      county_df["y"] = 0
      fig1 = px.bar(county_df, x="individualCount", y="y", text = '
                    color=county_df["individualCount"].keys(), ori
      fig1.update_yaxes(showticklabels=False, title=None)
      fig1.update_xaxes(range=[0, 3500])
      fig1.update_layout(margin={"r": 0, "t": 0, "l": 0, "b": 0})
      fig1.show()
```

county

- Hualien
- Kaohsiung
- Keelung
- Kinmen
- New Taipei
- Penghu
- Pingtung
- Tainan
- Taitung
- Yilan

3162 200

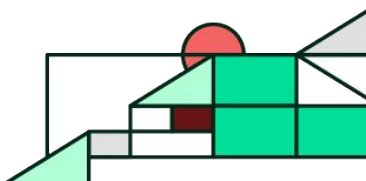
Simple 0 1 Python 3 (ipykernel) | Idle Mem: 214.97 / 2048.00 MB Mode: Command Ln 1, Col 1 Example.ipynb 0

② A JupyterLab with resources in the dataset will be presented

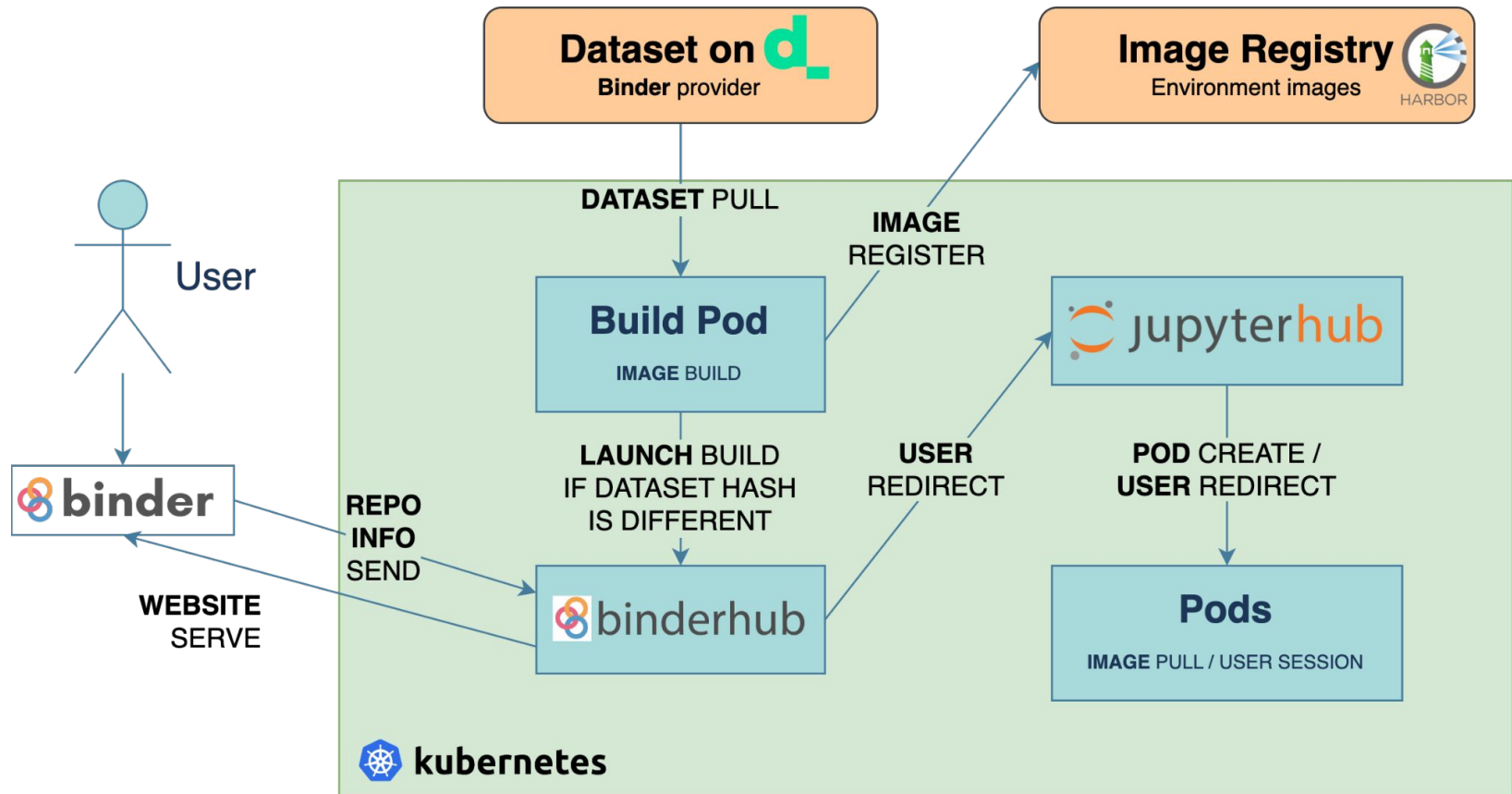
 **BinderHub** Beta

[launch binder](#)

① Click the "launch binder" button on the dataset page



Architecture of the BinderHub Integration



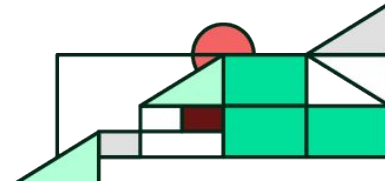
* Adapted from the [diagram](#) of the BinderHub architecture.

* Computing resource provided by Academia Sinica Grid Computing Centre, Grant No. AS-CFII-112-103.

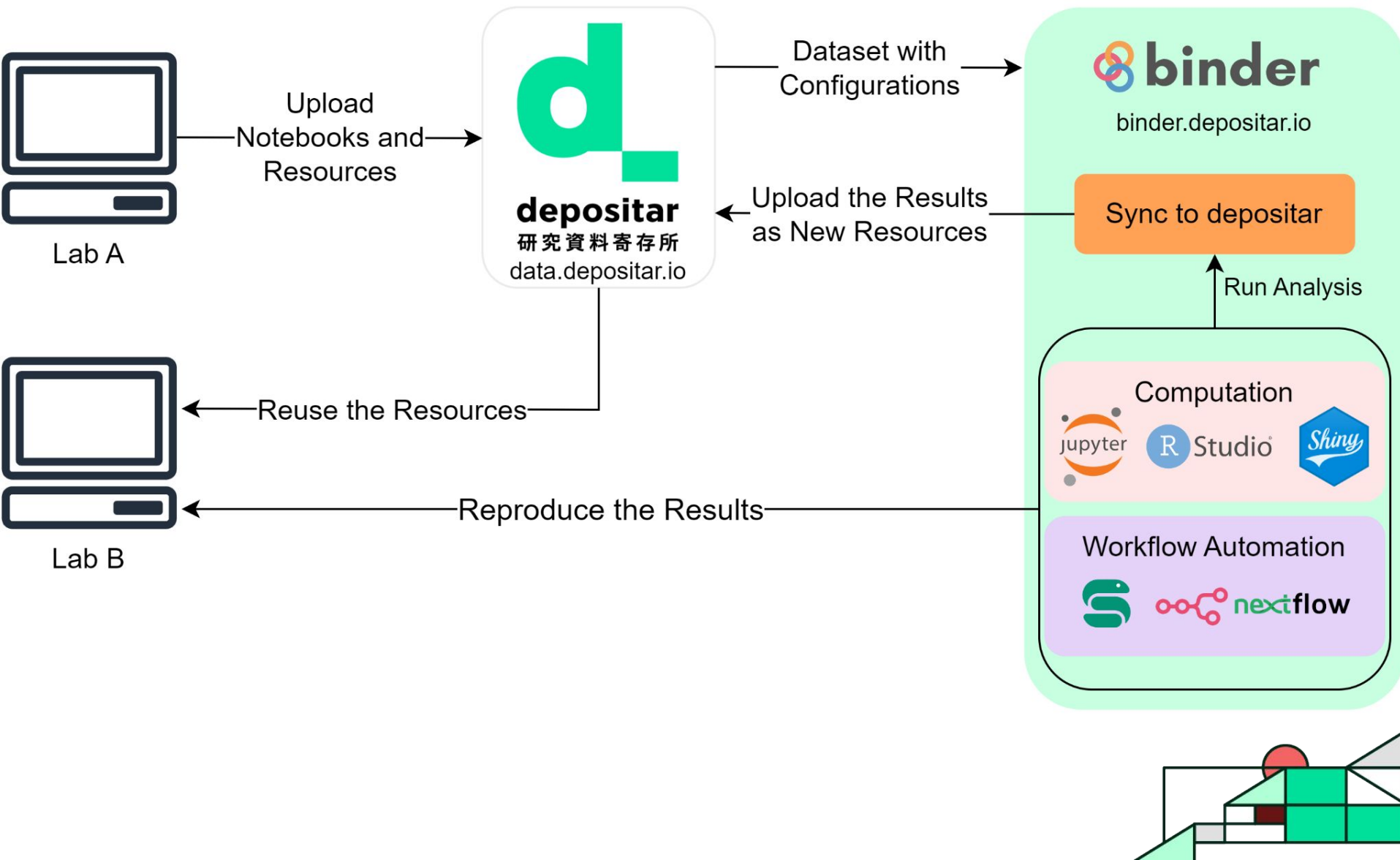
Research Workflow Automation

- 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



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、何明誼、李承 鑫、王家薰。

