Programmable Data Repositories

2022-03-06

Cheng-Jen Lee (cjlee@iis.sinica.edu.tw)
Institute of Information Science
Academia Sinica
Taiwan

Tyng-Ruey Chuang (trc@iis.sinica.edu.tw)
Institute of Information Science | Research Center for Information Technology Innovation | Research Center for Humanities and Social Sciences
Academia Sinica
Taiwan

Abstract

We present some of the recent development at the *depositar*, a research data repository built on top of CKAN. We will talk about our work in incorporating Archival Resource Keys (ARKs) to the depositar. We will also discuss how existing CKAN APIs can be used to realize an interactive workflow of data analysis and visualization. This leads to interactive and reproducible data repositories and further integrates them into the larger computing infrastructure for collaborative research.

Keywords

ARK, CKAN, Collaborative Research, Data Repositories, Interactive Computing

Audience

Researchers and developers who are working or interesting in data repositories, computing infrastructure, collaborative research, and so on.

Proposal

The *depositar* is a research data repository open to all for the deposit, discovery, and reuse of datasets [1]. It is open source software built on top of CKAN. CKAN is a leading open source data management system for building data hubs and portals (e.g. open government data websites) [2]. While CKAN is often used to set up *data publishers* (who curate datasets for the public), the depositar is *data depository* (who stores datasets for the users). The depositar is both free (anyone can register to use the service) and open source (anyone can copy, modify, and re-purpose its code). In addition, the depositar is an example of what we call programmable data repositories. It is programmable because the repository can be extended and new features added as needed by the repository's developers. The development of depositar has benefited from CKAN's user and developer community. Likewise, any new feature added by us for the depositar is contributed back to the CKAN community. All this is part of a virtuous cycle in which the software for CKAN-based data repository is being continuously enhanced to better serve users and to draw in more users [3] [4]. Previously we reported on several new features added to the depositar (on using Wikidata for keywords [5] and adding Shapefile support [6], for example). These features are now available to CKAN users and developers.

In this presentation, we will highlight some of the recent development at the depositar. We will first talk about incorporating Archival Resource Keys (ARKs) to the depositar. ARKs are persistent identifiers for information objects. They are in the form of URLs and are specified by *The ARK Identifier Scheme*, an IETF Internet Draft [7]. Since 2001, about 8.2 billion ARKs have been created by over 900 organizations [8]. Adding ARK support to the depositar will include two parts: assigning an ARK to each dataset that has been deposited, and resolving an ARK into the URL of the dataset it has been assigned to. Our implementation makes use of *Greens* [9] — an open source ARK minter and resolver. Related issues such as minting ARKs to versioned datasets and mapping the metadata of datasets to ARK's ERC (Electronic Resource Citation) metadata record will also be discussed.

We will also discuss how existing CKAN APIs can be used to realize an interactive workflow of data analysis and visualization. In brief, we can use CKAN APIs to filter, select, and fetch datasets from the depositar as well as to access the constituent resources (like CSV files) and their parts. When programmed in a Jupyter Notebook, for example, a large part of data analysis and visualization can be automated. When configured with automation framework including BinderHub [10] and Pangeo Forge Cloud [11] (which orchestrate the execution of Jupyter Notebooks in a computation farm), this leads to interactive and reproducible data repositories [12] and further integrates data repositories into the larger picture of computing infrastructure for collaborative research [13]. We shall illustrate with a few show cases as well from the depositar.

References

1. *What is depositar*? <https://data.depositar.io/about>; documentation: <https://docs.depositar.io/>; code: <https://github.com/depositar/>. ↩

2. *CKAN – The open source data management system* <https://ckan.org/>; documentation: <http://docs.ckan.org/>; code: <https://github.com/ckan/>. ↩

3. Josh Lerner and Jean Triole. (2000). *The Simple Economics of Open Source*. National Bureau of Economic Research Working Paper, No. 7600. ↩

4. Nataliya Langburd Wright, Frank Nagle, and Shane Greenstein. (2021). *Open Source Software and Global Entrepreneurship: A Virtuous Cycle*. Harvard Business School Working Paper, No. 20-139. ↩

5. Cheng-Jen Lee and Tyng-Ruey Chuang. (2019). *Improving data discovery through Wikidata*. WikidataCon 2019 Poster. ↩

6. Tyng-Ruey Chuang, Cheng-Jen Lee, Chia-Hsun Wang, Yu-Huang Wang. (2021). *Experience in Moving Toward An Open Repository For All*. Open Repositories 2021 Presentation. ↩

7. J. Kunze and E. Bermès. (2022). *The ARK Identifier Scheme*. Internet Draft, Version 34. ↩

8. *ARK Alliance — Home of the Archival Resource Key (ARK)* . <https://arks.org/> ↩

9. *uhlibraries-digital/greens: ARK identifier minter and resolver* <https://github.com/uhlibraries-digital/greens/>. ↩

10. *BinderHub — BinderHub 0.1.0 documentation* <https://binderhub.readthedocs.io/>. ↩

11. *Pangeo Forge Documentation — Pangeo Forge documentation* <https://pangeo-forge.readthedocs.io/>. ↩

12. Chris Holdgraf. (2019). *Binder + Zenodo: A how-to guide* ↩

13. Stern Charles, Abernathey Ryan, Hamman Joseph, Wegener Rachel, Lepore Chiara, Harkins Sean, and Merose Alexander. (2022). *Pangeo Forge: Crowdsourcing Analysis-Ready, Cloud Optimized Data Production*. Frontiers in Climate, Vol. 3. ↩