Experience in Moving Toward An Open Repository For All

2020-01-22
2021-02-08

[Note: This is a lightly edited proposal that was originally submitted to and accepted by Open Repository 2020 (but was not presented because of COVID-19).]

Authors

Tyng-Ruey Chuang, Institute of Information Science, Academia Sinica, Taiwan (trc@iis.sinica.edu.tw)

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

Chia-Hsun Wang, Institute of Information Science, Academia Sinica, Taiwan (allywang@iis.sinica.edu.tw)

Yu-Huang Wang, Independent Scholar (yuhuangwang@gmail.com)

Abstract

We report on our experience in building a domain-specific research data repository and in moving it toward an open repository for all to deposit datasets.

Keywords

CKAN, Open Source, Linked Data, Wikidata, Sustainability.

A Brief History of *data.depositar.io*

The research data repository data.depositar.io <https://data.depositar.io/en/about> is a repository open for all users to deposit datasets. Initially, it was built for the purpose of facilitating data sharing for two research projects about regional studies. The two consecutive projects were supported by fixed-term research grants for a total of 5 years. Registrations on the repository site were restricted to project members. The site was also used in class room teaching and was open to some students. When the second fixed-term research grant was ending in 2018, it was decided to retool the data repository for general use and to open it up for user registration: Anyone with a valid e-mail address can set up data sharing projects on the repository. Starting in the August of 2019, and supported by a new three-year research grant on sustainable research data management and collaboration, we are now formulating and conducting research with the data repository as an experimental base.

The software underneath data.depositar.io is based on CKAN <https://ckan.org/about/> which is an open source software package for publishing open data. However, many extensions had been added. New in data.depositar.io, to name a fews, there are metadata authoring aid for spatial extent and time period, shapefile preview function <https://github.com/ckan/ckanext-geoview/pull/54>, as well as e-mail verification for online account registration. User manual and installation guide for the repository is available <http://docs.depositar.io/>, and the software has been used to set up other data repositories with the same extended functionalities (for example, ecaidata.org).

The project team for data.depositar.io has been a small operation (at most 3 full-time people at any given time) in a national research institute on information science. Sustaining a general service of data repository is a new challenge to the team who in the past worked mainly on research and development but not on online services. Nevertheless, in this presentation we wish to offer a few observations and insights from our experience.

Build On Open/Libre Infrastructures

Use CKAN as the code base to build upon the software for a research data repository is a very important first step for us. As our team is small and supported mostly by fixed-term grants, using open source software as a base allows us to quickly experiment and add new functionalities. As we are required to release the code of our extensions and modifications (CKAN uses AGPL 3.0 license), this requirement actually also protects the users of our data repository: The users are free to install instances of their own and re-host their datasets.

We wish to emphasize there are open/libre infrastructures involved other than open source software. Open licenses, such as the Creative Commons Licenses, and free digital resources, such as Wikidata items, are components to rely and build upon for interoperable services. Open licenses allow digital resources (code/content/data) to be clearly marked for distribution and reuse freedom. Wikidata and other persistent ID services allow digital resources to be connected in ways that are precise and semantically rich. We have relied on the use of standard open licenses (for reuse conditions) and Wikidata items (for keywords) in the metadata properties of datasets. These small steps add reuse value to the deposited datasets.

Accept Diverse User Communities

When we decided to open the data repository for user registration, we did not have a clear idea of who would use the service the most. We anticipated users from research and education institutions. Gradually we realize the service has been used by quite a few small project teams working on short-term government contracts. The interface of our repository is entirely bilingual (Traditional Chinese and English) so is friendly to Taiwanese users (many other repositories support only English). As the Taiwanese government is pushing for open data as a policy goal, there is a need to deposit, openly somewhere, datasets from project outcome with proper metadata support. Ours meets the needs and is free to use.

We have also been contacted by project grants and government agencies to set up separate instances. Although we will be funded to do so, we have insisted on using data.depositar.io instead for their purposes. Maintaining just one instance is simpler than maintaining several ones. As the user base grows, so is the number of suggestions for new interface features and additional metadata fields. We take their suggestions but will not customize for them (no new instances!). Instead, we work on the next edition of the software in which some of the suggestions are accommodated.

Adopt Common Vocabularies and Formats

Related to the issues from diverse user communities and additional metadata fields, actually we are moving in the opposite direction in harmonizing and simplifying metadata fields. Domain-specific metadata requirements from the two previous research projects, which are about regional studies, are being removed (e.g. bibliographic metadata). Common metadata vocabularies will always be preferred especially when we now publish the data catalogue of the repository in linked data. We are also re-examining and retrofitting all metadata fields so that their values will have well-defined and precise ranges compatible with vocabularies defined in RDF (Resource Description Framework). That is, metadata all have types, they are not just texts. The data catalogue exposed in RDF format and published by data.depositar.io is now indexed by Google Dataset Search. As such the datasets deposited on it shall more likely be discovered and reused.

The change of metadata schema in a data repository is a complicated business; datasets with old metadata need to be updated with new metadata when the repository is upgraded. Fortunately datasets are migrated in batch, and this occurs only when the software upon which the repository is running is upgraded.

Curate Active Datasets, Privately

Datasets are put into a repository often when they are ready to be used by others, and the repository functions as an archive for people to search and discover useful datasets. What are deposited usually are not active datasets (which have not undergone a proper process of selection and hand-over). However, in data.depositar.io it inherits from CKAN the feature of forming a (data sharing) project by associating a group of users as the project members. Project members can share datasets **only** among themselves; these are called private datasets.

We observe data.depositar.io is being used as a platform for projects to curate and manage active datasets that are private to project members. Sometimes the datasets are stored elsewhere; only the links to the datasets are deposited in the repository. The novel value provided by data.depositar.io, in such a scenario, is in its management of metadata for, and the curation of, these active datasets.

Find Paths to Self-Sustainability

Why build and set up your own data repository, especially when your team is small? That is a good question. There can be several reasons: better metadata design and dataset control, meeting local needs, and establishing a community of good practices for data sharing, etc.

How to sustain the data repository in the long run, however, is a more difficult question to answer. It is a good problem to have, however, as the question itself implies there is a demand of the continuous provision of the repository's service. We are thinking of starting a process in which feasible arrangements can be formulated to support the service in the long term. A consortium supported data repository looks like a good idea. The support can be in kinds (e.g. storage and bandwidth) or in grants (e.g. funding operation).