https://hackmd.io/@trc/depositar-SciDataCon2022-script

The script for a presentation at the Data Collaborations Across Boundaries session at SciDataCon 2022, part of the International Data Week.

[Script] Bottom-Up Research Data Repositories

2022-06-21

Tyng-Ruey Chuang

1

Hi, I am Tyng-Ruey Chuang.

I am researcher at Academia Sinica.

My background is in Computer Science and has been working over the years with peoples from other disciplines to make better use of research data.

Today we will share our experience in running a bottom-up data repository.

I will be presenting with my colleagues Cheng-Jen Lee, Chia-Hsun Wang, and Ming-Hsuan Ho.

We are from the Institute of Information Science and the Research Center of Information Technology Innovation, Academia Sinica, Taiwan.

2

I will make a brief introduction to the *depositor*, a research data repository we build.

It is an open repository for all to deposit datasets.

I will give a small tour on how it works.

Then I will explain why I call *depositor* a bottom-up data repository.

3

So, here is an actual dataset from the *depositar* on Corel Reef Soundscapes in Okinawa, Japan.

On the left is the page you will get about this dataset at the *depositar* website.

I will highlight some places for you to look at them:

There are long descriptions about the dataset and the project.

A dataset will include multiple data files and the links to external resources.

There are Wikidata keywords to annotate a dataset.

And the metadata comes in three categories:

basic information, spatio-temporal information, and management information.

There are also license and citation information attached to a dataset.

There are machine-readable data endpoints to access a dataset and its metadata.

4

Let’s break into three parts.

On the top of the page, you see the tile and description of the dataset.

One the left, you see “Ocean Biodiversity Listening Project” which is the project depositing this dataset.

A project can deposit as many datasets as it likes.

5

Here at the middle of the page, we see the data files and the links to external resources that together constitute the dataset.

The “explore” bottom will take you to the files.

On the left, we see the dataset is CC BY licensed.

There are also citation snippets, in many citation styles, for people to cite this dataset.

6

OK, this is the last part.

You see tags and Wikidata keywords for this dataset.

The dataset’s temporal resolution, starting and ending time, and spatial coverage are also described.

There is also management information so you know whom to contact for the dataset.

On the left of the page, there is a map showing the spatial coverage, as well as the machine-readable data endpoints of the dataset.

7

The *depositar* has a bilingual interface.

Now I am showing you the Traditional Chinese interface for the same page.

Please take a look at the Wikidata keywords, it now displays the Traditional Chinese labels in stead of the English labels.

So you can see the two Traditional Chinese characters 聲景 for Soundscape as the label of the first Wikidata keyword.

8

Now I take you back to the English interface.

Let’s look at the description of the dataset.

The creators of this dataset have provided an associated publication which is in the journal Biological Conservation.

There is a link, and it will take you to the publisher’s website.

9

You look at the paper and, in the data availability section, it says that “the audio dataset … are available from the authors … and … on depositar”.

And the authors provide a link.

We click on the link.

10

And we find the dataset at the *depositar*.

So the dataset and the publication links to each other.

This kind of mutual reference, I think, is wonderful!

11

Datasets at the *depositar* are found by Google Dataset Search.

You key in the term “Coral Reef Soundscapes” and you get some links to datasets in return.

This dataset from Okinawa, Japan, shows up at the second place.

You click on the “Explore at depositar” button.

12

And this will take you to the dataset at the *depositor*.

13

The *depositar* is a data repository built from the bottom up.

We can describe the bottom up nature in terms of process and tools.

14

The precursor of *depositar* was developed for two research projects in order to help project members better share their data.

We knew in 2018 the funding was coming to an end, and decided to turn the system we built for the projects into a service everyone can use.

We launch the *depositar* as a public service in October 2018 at the Pacific Neighborhood Consortium Annual Conference in San Francisco.

As our background is in information science and open source, to open the service to the public is just natural to us.

This is a big change of our role, however, as we are no longer acting as a data publisher for a small number of researchers but is now providing data deposit service to the public.

15

The *depositar* has an organic growth since 2018.

Surprisingly, the first uptakes are from non-profit organizations.

With the help from Dr. Yu-Huang Wang, a few organizations started to use the service to deposit and disseminate datasets about ecological impact assessments.

We also receive a new grant from MOST with some focus on Research Data Management.

As a result we are interacting with other project teams working on Sustainability Research and on Long-Term Social and Ecological Research.

Some of the teams start to use the *depositar* for data release and sharing.

Of course, there are researchers, in Taiwan and elsewhere, who use the service.

Finally in July last year, the depositar's Terms of Use and Privacy Policy were finally in place.

16

I want to show you a series of workshops co-organized by the non-profits on Research Data Managment related to public construction works and their ecological impact assessments.

That was in 2019.

For the new grant from the MOST, the *depositar* team organized several RDM workshop as well.

We also translated the excellent RDM guide from Science Europe.

These activities are not directly related to the operation of a data repository, but they connect the *depositar* to the wider data communities in Taiwan and elsewhere.

17

The *depositar* is built on top of CKAN.

CKAN is an open source software package from the Open Knowledge Foundation.

It has been used to publish open data, for example in setting up open data portals.

We have re-purposed CKAN to build an open repository for all to use.

We build on top of CKAN.

But we add user sign-up and registration by e-mail, and we support enriched metadata vocabularies for research data.

We also added features to connect datasets to other resources on the Web.

For this, we use Wikidata as a source of keywords.

We add a citation widget so datasets are ready to be cited.

And we publish data catalogs in linked data so datasets are findable.

We are now working on using Archival Resource Key as a generic mechanism for persistent identifiers so that the deposited datasets will all have PIDs.

Finally BinderHub integration!

We hope the *depositar* will soon become a data source in BinderHub for interactive computing with Jupyter Notebook.

18

CKAN is a tool for sharing data but by itself the tool does not keep data.

We use CKAN and build a system to keep and share our own data.

Based on the system we build, we start to provide a service to keep and share data from others.

This is an increase of reponsibility and we shall worry about how the *depositar* can be sustained as a service to the public in the long term.

Our strength will necessarily come from the data communities around the services we provide.

Collaborations among diverse data communities, I think, will show directions and give supports to a service the communities need and like to use.

Currently the *deposiar* is a public good sustained by public funding.

In the long term, I think the *depositar* may need to be sustained by resources put together by the communities.

19

Thank you!

We encourage you to take a look at the *depositor*.

The website is data dot depositar dot io .

Please send us e-mail too.

We love to hear from you!