

Presentaton script in text form: https://hackmd.io/@trc/depositar_at_RDA-P18_script

Presentaton slides in PDF: <https://m.odw.tw/u/trc/m/rda-p18-panel/>

Prepared for “Open Science Initiatives in Asia” panel at the 18th Research Data Alliance Plenary Meeting : <https://www.rd-alliance.org/plenaries/rda-18th-plenary-meeting-virtual/open-science-initiatives-asia>

[Script] Open Repositories for Scholarly Communication and Participatory Research

Tyng-Ruey Chuang
trc@iis.sinica.edu.tw

2021-11-10

1

Hello, I am Tyng-Ruey Chuang.

I am a researcher at the Institute of Information Science, Academia Sinica, Taiwan.

I will talk about “Open Repositories for Scholarly Communication and Participatory Research”.

2

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

I will later give a small tour on how it works.

The *depositor* is an open repository, a data repository open to all for registration. The Registration is free. Any one can use it.

It is also free software. The source code is free to download. It builds upon CKAN, an open source package for setting up open data portals. We extend CKAN, and make the extensions available to the public under AGPL 3.0 which is the license CKAN uses.

One of our main goals is that datasets hosted at the *depositor* will be FAIR: Findable, Accessible, Interoperable, Reusable. I think we achive this goal.

Noted that the *depositor* is a data depository not a data publisher.

By some definition, a publisher engages in “acquisition, copy editing, production, (e-)printing, marketing and distribution”.

While a depository is “a place where something is deposited, as for storage, safekeeping or preservation.”

We will move on to Open Science in Asia later in this presentation.

3

The *depositar* was formally launched at 2018 Pacific Neighborhood Consortium Annual Conference and Joint Meetings in San Francisco, CA, USA, on Oct 27, 2018.

4

This is a photo of the our presentation at that time.

5

Here is a slide I used three years ago. I think this slide pretty much stays the same and speaks of our approach.

Allow me to read out the slide again.

6

Now I give a tour of the *depositar* (研究資料寄存所).

7

Here we look at 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 us to look into:

There are long descriptions about the dataset and the project.

A dataset will include multiple data files as well as links to external resources.

You can use tags and 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.

8

Let's break into three parts.

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

One the left of the page, you see "Ocean Biodiversity Listening Project" which is the project depositing this dataset.

A project can deposit as many datasets as it likes.

9

Here at 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 citation snippets for people to cite this dataset.

10

OK, this is the last part.

You see tags and Wikidata keywords for this dataset.

The dataset’s temporal resolution, time period, and spatial coverage are described.

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

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.

11

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 Chinese labels in stead of the English labels.

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

This is the end of the tour.

12

Since 2018, we have made improvement to the *depositar*.

Because the data catalog published by the *depositar* uses standard vocabularies. The catalog has been indexed by Google Dataset Search.

As the service has been available to the public for a while. We also get to know more about the communities using it.

In addition to researchers, there are also citizen groups users. For example, non-profit organizations working on ecological impact assessment of public works have been using our service to document their works.

We also have a Terms of Use and a Privacy Policy.

We are also tasked with outreach activities about Research Data Management because of a grant we received from the Ministry of Science and Technology.

We also organized RDM workshops and has a RDM website.

We are on twitter too!

13

I wish to share a few thoughts on using data repositories for Scholarly Communication and Participatory Research.

14

For an example, let us look at this paper and, in the data availability section, it says that “the audio dataset used in preparing this paper are available from the authors ... and a dataset ... is available on depositar”.

And the authors provide a link.

We click on the link.

15

And we find the dataset at the *depositar*.

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

This is a link, and it will bring you to the publisher’s website.

So the dataset and the publication links to each other.

This is a wonderful mutual reference!

16

You can also discover datasets at the *depositar* by Google Dataset Search.

You search for “Coral Reef Soundscapes” and you get 29 datasets in return.

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

You click on the “Explore at depositar” button.

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

17

Let’s do another Google Dataset Search using the three Hanzi characters 劉厝溪.

And this brings up a dataset deposited by Dr. Yu-Huang Wang at the *depositar* about some ecological survey of a small river in central Taichung.

18

Here is the dataset at the *depositar*.

I will show you two resources in the dataset.

Actually the two resources are external links, but at the *depositar* detailed descriptions about the external resources have been provided for them.

19

The first is a mosaic of orthophotos, itself is a link to an Open Aerial Map web page.

20

The second is a 360° panorama in the way of a Google Street View link.

21

At this point I wish to share some of my observations and thoughts.

First of all, I think “Open Science” is more about advocacy than policy as now, at least in Taiwan.

We still do not see high-level national policies on open science: how will open science be funded and what to expect of its outcome in terms of research assessment or even evaluation metrics.

So, which approach would researchers and research institutions go for to advocate open science?

Would you go for top-down policy change?

Or would you go for bottom-up culture change?

In my view an open repository is a bottom-up step in cultivating a culture of openness by practicing what we preach, by actually making data available (even there is no high-level data policy in relation to open science).

But why would we need to build our own data repositories?

I think part of the motivation is to serve and to know our communities better.

Often there are culture and language issues and there are different local needs.

Our experience is that users of the *depositor* are happy they can talk to people in Taiwan for supports and about their needs.

By building our own data repository, we also get to learn all the details, and can help others replicate the skills.

I will also suggest that, in building one's own data repositories, one shall reuse as much as possible.

This includes source code, common vocabularies, standards and services, etc.

And now come to the difficult question of how to sustain a Do-It-Yourself data repository.

I don't have good answers, but I think it will depend on stable funding and persistent advocacy.

22

Here are some works in progress.

I will just quickly go over the plan:

We will use Archival Resource Keys (ARKs) for persistent Identifiers.

We are also working on support for large collections of media files (for example, for aerial images generated from drone surveys).

There are tasks related to Research Data Management too.

23

These are some screenshots from the RDM workshops we organized in 2018 and this year.

24

We recently set up a website called Research Data Management Hub, and we hope to outreach to the research communities in Taiwan about good Research Data Management practices.

25

Earlier this year we translated into Taiwanese Mandarin the excellent guide from Science Europe on *Practical Guide to The International Alignment of Research Data Management*.

26

Again, I show you some screen shots of the *depositor* website.

Currently about 150 projects use the service.

2,000 datasets have been deposited to it, of which about 1,000 datasets are open data.

27

Thank you!

The *depositor* website is at [data dot depositor dot io](http://data.dot.depositor.dot.io).

Please check it out.

Please send us e-mail too.

We love to hear from you!

The depositar project team: T-R Chuang, M-S Ho, C-J Lee, Monica Y-C Mu & Ally C-H Wang.

研究資料寄存所計畫成員：莊庭瑞、何明諠、李承鑫、穆昱佳、王家薰。