This short script is also put online:

https://hackmd.io/@trc/depositar-OR2021-short-script

[Short Script] Experience in Moving Toward An Open Repository For All

1

Hi, I am Tyng-Ruey Chuang.

I will be presenting with my colleagues Cheng-Jen Lee and Chia-Hsun Ally Wang.

We are from the Institute of Information Science, Academia Sinica, Taiwan.

Dr. Yu-Huang Wang, an independent scholar and a collaborator in this research is also joining us at this session.

They will join me at the Q&A session.

2

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

I will give a small tour on how it works.

Then I will talk about the history of the *depositor* and our experience in running this data repository.

We would also raise some issues for discussion.

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.

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.

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 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.

There are citation snippets 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, 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.

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 Chinese labels in stead of the English labels.

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

8

Now I take you back to the English interface again.

Notice the Wikidata keyword is back to Soundscape.

We get these labels in real-time from Wikidata.

9

Now let's go back to the description of the dataset.

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.

10

You look at the 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.

11

And we find the dataset at the *depositar*.

So the dataset and the publication links to each other.

This is a wonderful mutual reference!

12

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.

13

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

14

There are about 130 projects using the *depositar* as now.

In total there are about 800 public datasets.

15

The *depositar* is an open repository for all to use!

The website is at data dot depositar dot io .

16

Now we turn to the part about our experience in running this open repository.

We also raise a few issues for discussion.

17

Why do we build this open repository?

There is some history behind it.

Starting in year 2013, we received two grants from Taiwan's Ministry of Science and Technology on multi-disciplinary area studies.

These were joint projects where each included 4 or 5 teams.

We were the only team with a background in information science.

We worked on tools and systems so as to make it easier for the project to share research data.

Very early on, we decided to use CKAN to set up a data-sharing system.

We use CKAN because it has been used extensively by governments to set up open data portals.

It is open source and written in Python -- we like Python.

We re-purposed CKAN so that it would not be used for publishing open data to the public, but instead would be used to share research data among project members.

We have done a lot of customization however.

For example, we built in domain-specific keywords as a way to do controlled vocabulary.

We extended the file preview function of CKAN to include shapefile so we can preview maps without first downloading them.

We are collaborate with others in Academia Sinica to receive feedback.

We are also rooted in a culture of open content, open data, and open source, so we pretty much set our own goals in exploring this space.

At Open Repositories 2015, actually we had a poster describing our approach.

18

As the second joint project was running to an end, we decided to open up the repository to the public for general use.

Now everyone can deposit datasets .

We re-launched the site at the Pacific Neighborhood Consortium Annual Conference, co-located with Digital Heritage 2018, at San Francisco in October 2018.

There are a lot of preparation to do.

The first is e-mail sign-up and user authentication.

So everyone with an e-mail address can apply for an account.

The site also has a new user interface, and we improve the user manuals.

Later on, we add citation snippets, and we use Wikidata for system-wide keywords.

Since the new launch in October 2018, there are new uptakes of the service.

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

We are also in dialogue with government agencies about using the service to share public sector information.

Very fortunately, we received another grant from Taiwan's Ministry of Science and Technology.

The focus of the new grant is on research data management, about fostering a culture of practicing good research data management.

But the *depositar* will be part of the picture, as researchers need services they trust and like before they deposit their datasets.

Still we need more active users.

19

Our experience in building an open repository is very encouraging.

We are happy to build on the open and libre infrastructures.

In fact, we think that is only way to go.

CKAN is open source.

It is licensed under AGPL 3.0 which means that new code added to a CKAN instance, like the *depositar*, must be released to the public.

We have contributed all our extensions back to the upstream CKAN developers.

These extensions have been incorporated into new releases of the software.

In a way, this protects our users too.

Because they can run an instance of their own, with all the functionality of the *depositar*, if they like.

They can host their own datasets.

We also see CC licenses, Web standards, and Wikidata as important infrastructures to build upon.

Our experience also suggests that we shall prepare to accept diverse user communities ---

communities such as non-profit organizations and government agencies which are not traditional academics.

They have different needs and expectations.

Often time, we would discuss whether to set up a new instance to serve the needs of a particular user community.

We have decided not to do so and we encourage the users just to use the *depositar*.

Because maintaining multiple instances will increase our own responsibility.

Currently the quality of datasets and their metadata at the *depositar* is an issue.

The quality is unbalanced and we are looking for methods to address this.

We encourage the use of common vocabularies, such as DCAT for publishing data catalog, so that the datasets can be properly indexed and discovered.

And we need to formalize the data types of metadata.

This will allow for proper data validation and interoperability therefore will improve data quality and reuse.

20

Now we come to the last slide and I list several issues for discussion.

We come to the community of data repositories from a different path.

We are tools developers and we build systems.

Software tools by themselves do not keep data.

From software tools we can build information systems which can process and keep our own data.

But when we move from systems to services, we are processing and keeping other people's data.

There is an increase of responsibility.

To keep the service running and to preserve people's data is a sustainability issue we are learning to cope with.

On the one hand, we can aim to keep the service running forever.

But we need to look for resources in order to do this.

The other viewpoint is to just preserve users' datasets but not necessarily at the service we run.

We think collaborations with other long-running data communities is very important.

For example, we are looking into use ARKs, Archival Resource Keys, as a way to issue PIDs for datasets.

We are an ARK Name Assigning Authority Numbers (NAAN) and we plan to provide some services later this year.

There are opportunities for two-way enrichment between data repositories and the Wikidata as well.

We currently use Wikidata as the source of keywords.

But we can also contribute metadata from the *depositar* to Wikidata as well.

This concludes my presentation.

21

Thank you!

22

The *depositor* website is at data dot depositar dot io .

Please check it out.

Please send us e-mail too.

We love to hear from you!