# Communal Data Projects and Sustainable Digital Preservations

ECAI Workshop
PNC 2019 Annual Conference and Joint Meetings

Oct 15-18, NTU, Singapore

Tyng-Ruey Chuang Academia Sinica, Taipei, Taiwan trc@iis.sinica.edu.tw

## Outline

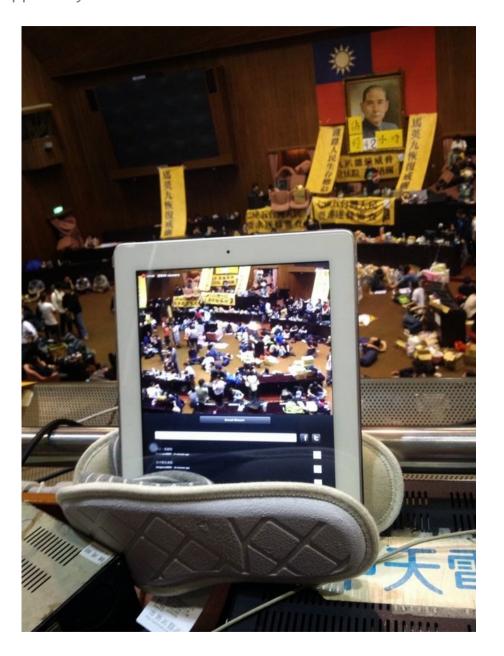
- Two projects
  - The Sunflower Movement Archive
  - Taiwan Roadkill Observation Network (TaiRON)
- Reflections
- Paths of Inspiration
- Experimentations in Progress
  - data.depositar.io

## Sunflower Movement

- On March 18, 2014, students and activists started to storm into and occupy the main chamber of Taiwan's Legislature.
- The chamber was transformed into a theater of protest for weeks; events were streamed live.
- The movement had been instrumental in delaying the passing of the Cross-Strait Service Trade Agreement (CSSTA 海峽兩岸服務貿易協議) by the ruling party at the Legislature, and in increasing public awareness of various issues in trade negotiations with China.
- https://en.wikipedia.org/wiki/Sunflower\_Student\_Movement
- https://zh.wikipedia.org/wiki/太陽花學運

## 專訪 LIVE 阿北熱血轉播 | 即時新聞 | 20140328 | 蘋果日報 http://www.appledaily.com.tw/realtimenews/article/new/20140328/368726/





"Occupy Taiwan Legislature by VOA (8)" by Voice of America https://commons.wikimedia.org/wiki/File:Occupy\_Taiwan\_Legislature\_by\_VOA\_(8).jpg





## The Sunflower Movement Archive



0

000

13063 | public.318.io

+



















## 13063

識別號: 13063 主要形式: 平面

形式分類:海報 印刷

內容分類: 繪畫

產製日期: 產製地點:

測量: 599x427mm

描述:

兩張海報組成。左邊畫面為一個披著中華民國國旗的 人,身邊有多朵太陽花。右邊畫面為兩隻手正在握手, 周圍有許多太陽花,標題為「謝謝在前線的你們台灣加

油」

內容:

材質: 合成紙

姓名標示值:

< 上一筆 温瀏覽

(共7389筆,本頁顯示第1筆)

 $\equiv$ 

<u>File Edit View History Bookmarks Tools Help</u>

( 12958 | public.318... × +

(+) (i) public.318.io/12958

C Q Search

☆自→命▽



關係藏品:

關聯

新年2 日本年2日本 日本年2日本 日本年2日本 日本年2日本

部份







10492





10494





10489

10496

节清和海

C. VARIABLE



化原型 化基础作品

10497

10490



HETENAN THEORY THEORY

Males of the sales

10491



10499



10500

10493













識別號: 12958 主要形式: 平面

形式分類:海報 黏貼 內容分類: 鼓勵語

產製日期: 產製地點:

測量: 1005x655mm

描述:

紙板,上面黏貼大量的便利貼以及「香港中文大學」的字條,便利貼內容為支持反服

貿的鼓勵語。

內容: 香港中文大學 反暴力 撐台灣

材質: 塑膠瓦稜板 姓名標示值:

權利狀態: 權利狀態不明 ①

指認與授權

若您是本藏品著作人,請點選上面連結進行指 認授權作業,想了解更多,請參考相關說明。

引用網址: http://public.318.io/12958



12958





12958

12958

<u>File Edit View History Bookmarks Tools Help</u>

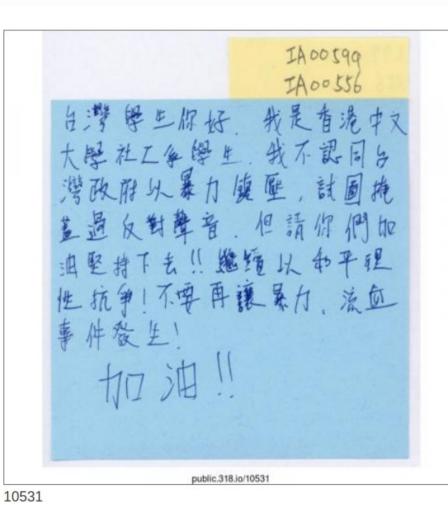
⊕ 10531 | public.318... × +

( public.318.io/10531

C Q Search

☆自♣♠♥■

↑ > ALL > 10531



10531

識別號: 10531 主要形式: 平面

形式分類: 便利貼 書寫

內容分類: 鼓勵語

產製日期: 產製地點:

測量: 75x75mm

描述:

便利貼,上有文字「台灣學生你好,我是香港中文大學社工系學生,我不認為台灣政府以暴力鎮壓.....加油!!」

內容:

台灣學生你好,我是香港中文大學社工系學生, 我不認同台灣政府以暴力鎮壓,試圖掩蓋過反對聲音,

但請你們加油堅持下去!!

繼續以和平理性抗爭!不要再讓暴力、流血事件發生!

加油!!

關係藏品:

整體



12958

File Edit View History Bookmarks Tools Help

(a) news.ltn.com.tw/news/life/paper/1052141

C Q Search

☆自◆

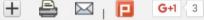
首頁 > 生活

## 318學運文物 台史博接手典藏















2016-11-15

[記者劉婉君、楊淳卉/綜合報導] 兩年半前受全民關注的三一八公民運動(太陽花學運), 相 關文物中研院歷經兩年整理, 昨移交國立台灣歷史博物館, 七千兩百餘件實體物件中, 包括當 時內部運作關鍵的「值班日誌」, 還有過去事件中少有的網路數位媒材資料及數位電子檔案, 將由台史博接手保存、典藏與後續的深入研究,為未來子孫留下珍貴文物。



中研院移交318公民運動文物給台史博。(記者劉婉 君摄)



#### 中研院整理出7200餘件實物、網路資料

移交的文物中除了關鍵的值班日誌,還有書 信、民眾為參與者加油打氣的便利貼、抗議 標語、海報、布條、文宣等,由中研院台灣史 研究所所長謝國興代表移交台史博館長王長 華,文化部文化資源司長陳登欽也到場見 證。由於數量龐大, 台史博預估, 所有文物僅 清潔、入庫、上架,就約需兩年時間才能完 成. 將借助大數據分析協助整理。

### 台史博估花2年時間才能入庫、上架

過去兩年,中研院已清點、記錄,並完成「三 一八公民運動文物紀錄典藏庫簡述」 (http://public.318.io/about)及「三一八公 民運動」(http://expo.318.io/),於網路平 台公開。台史博昨也以「為明日收藏:三一八

#### 近期網友關注

#### 瀏覽記錄

- 「平台哥」大爆紅! 王大陸親自解 惑:平台是...
- 真要打破一中了? 彭斯上電視又講 「台灣總統」
- 3 歐盟官員19歲女兒遭難民姦殺 頭髮 還被拔...
- 4 推特發文狂嗆中國 川普:南海擴軍 也沒問美國
- 5 獨家》陳錕山投共中大遭依採購法 停權3年

### 市場動態

- · Acer有三寶! 尋寶樂園等你來
- 聰明喝好茶 綠茶益處更多
- 看準舌尖商機 百貨大打美食戰

更多市場動態 »



## Preserving Online Collaborative Documents



Hackpad
318.hackpad.com



Hackfoldr hackfoldr.org/congressoccupied

## Hackpad Documents



- Online colaborative editing platform.
- 318.hackpad.com

   documenting
   activities; verbatim
   transcripts of
   speech.
- hackpad.com
   was acquired by
   dropbox. Services
   were closed in
   2017.



## Hackfoldr Documents



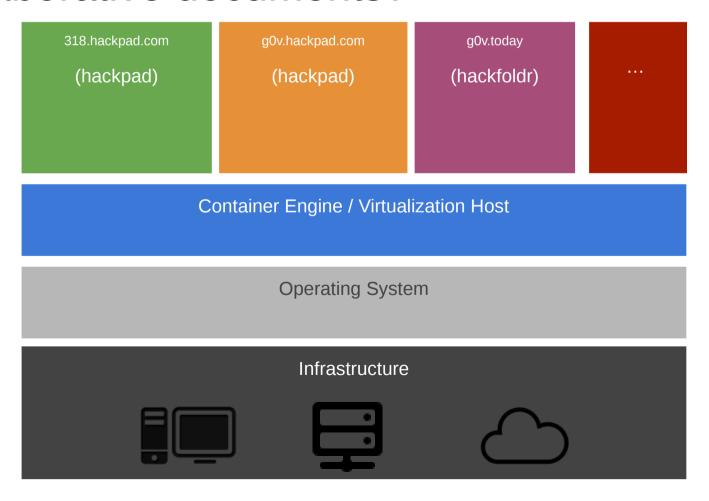
- Folder-like online document aggregator.
- g0v.today –
   integrates
   important and
   current
   information about
   the movement.
- Actual preservation target:

http://hackfoldr.or g/congressoccupi ed/



## The Web and Long-Term Preservation

 What is the idealized preservation model for collaborative documents?





https://www.nmth.gov.tw/exhibition\_236\_405.html







# TaiRON https://roadkill.tw



## Taiwan Roadkill Observation Network

HOME

**ABOUT** 

**PARTICIPATE** 

CAMPAIGN

REPORT

台灣中文版

GOAL

1

### Citizen Science

Popularize civil participations in science researches, detect environmental problems around our life, and then make theories, design experiments, collect information, analyze, discuss and solve together.

2

### Mitigate Roadkill

Find out the reasons, spots, happening seasons, and categories of wild animals' deaths, trying our best to mitigate it, live in peace with all species and share the Earth. 3

### Environmental Education

By taking part in the simple activity of recording "Roadkill" as the first step, participants gradually carry out the ideas of conservation and sustainability of ecology. 4

#### Cherish Life

Life is equal and should be respected and cherished.
Please help collect the carcass of animals in mishap, making into specimen for academic study and preservation, or for the monitor of epidemic and environmental agents.

## App





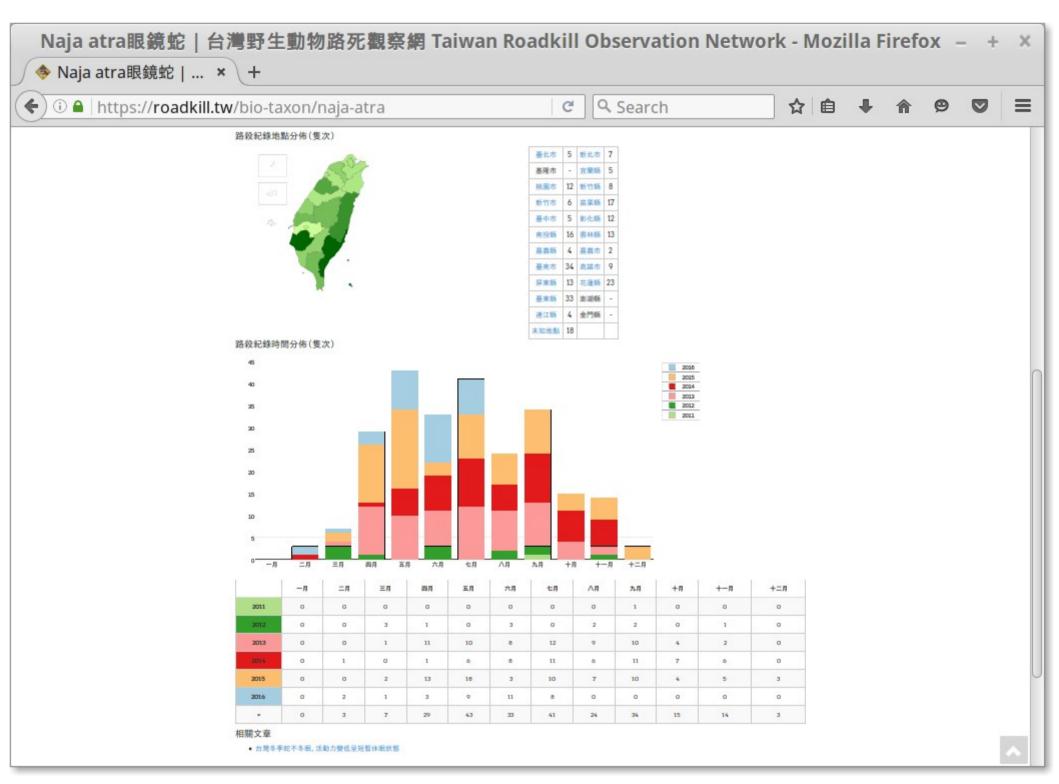
## TaiRON website

## **Facebook Group**



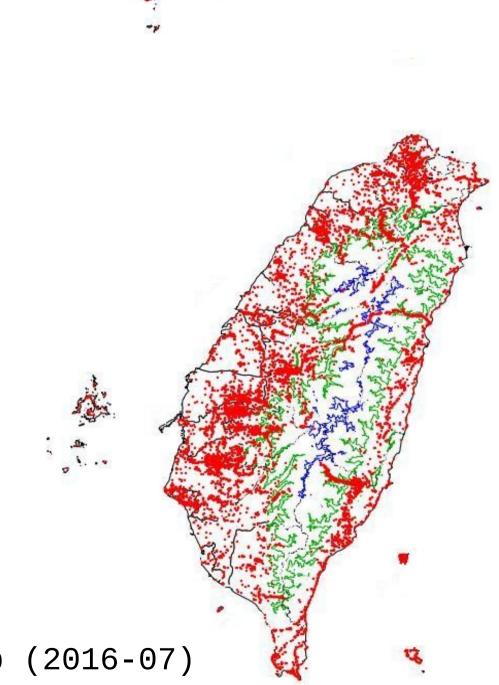




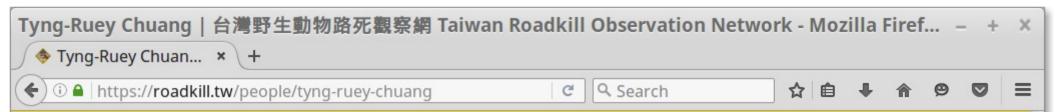


Elevation 1,000m
Elevation 3,000m

Location of roadkill observations



Roadkill Observation Map (2016-07)



### \* > Tyng-Ruey Chuang



帳號:: Tyng-Ruey Chuang (870)

中文姓名: 莊庭瑞

英文姓名: Tyng-Ruey Chuang

暱稱: trc

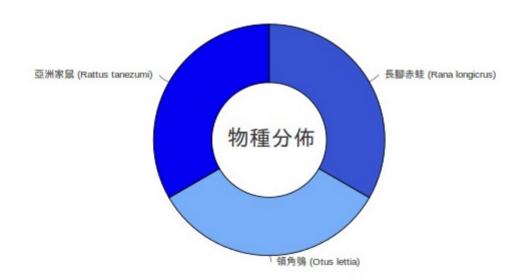
預設授權方式: CCO

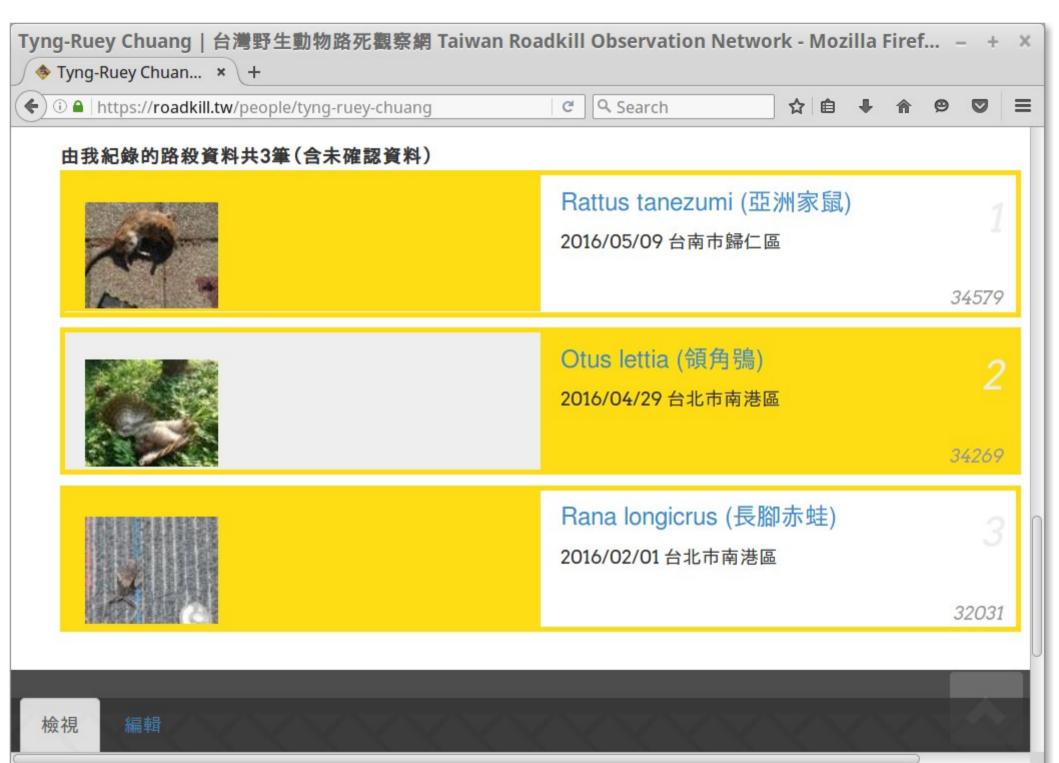
路殺紀錄(已確認): 3筆, 3隻, 3物種

路殺紀錄(含未確認): 3筆

首次回報(上傳)於: 2016/02/01

Member Profile and Collection





## Reflections

## Projects, Institutions, Sustainability

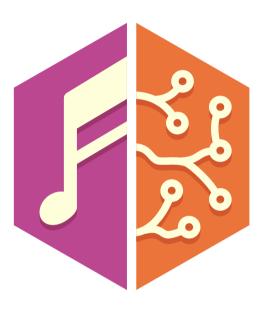
- What we have: Impromptu and experimental research projects curating digital collections of community content/data.
- What we need: Sustainable arrangements for long-term preservation of the digital collections.
  - c.f. roles of research institutions (national labs, universities, grants) & memory institutions (GLAM)
- What we can do: Collaborate across institutional boundaries in practicing communal content/data production, management, and reuse.

## Digital Preservation

- Preservation of digital objects
  - artefactual aspects (e.g. print; hardware/medium)
  - informational aspects (e.g. script; file/program)
  - functional aspects (e.g. performance; interaction)
- Keys for long-term digital preservation:
  - free software, open license, common format
- How to preserve and represent digital objects as events are unfolding?
- Sustainable Digital Preservation?
  - resources and institutions
  - planning, process, and governance

## Communal Data Projects









## WikipediA

### **English**

The Free Encyclopedia 4 000 000+ articles

## **Español**

La enciclopedia libre 900 000+ artículos

### Русский

Свободная энциклопедия

875 000+ статей

#### Italiano

L'enciclopedia libera 939 000+ voci

## **Português**

A enciclopédia livre 742 000+ artigos

### 日本語

フリー百科事典 815 000+ 記事

#### Deutsch

Die freie Enzyklopädie 1 430 000+ Artikel

### **Français**

L'encyclopédie libre 1 270 000+ articles

### **Polski**

Wolna encyklopedia 908 000+ haseł

### 中文

自由的百科全書 500 000+ 條目

Search · Suchen · Rechercher · Zoeken · Ricerca · Szukaj · Buscar · Поиск · 検索 · Busca · Sök · 搜尋 · Tìm kiếm · Пошук · Cerca · Søk · Haku · Hledání · Keresés · 찾기 · Cari · Ara · بصنب · Căutare · نصن · Hl'adat' · Søg · Serĉu · Претрага · Paieška · Poišči · Cari · יויפוש · Търсене · Іздеу · Bilatu · Suk · Bilnga · Traži · खोजें



[ MediaWiki

content

tools

users

services





## Paths of Inspiration

# Civil Infrastructure Platform (CIP) – for supporting long-running systems

- e.g. for railway systems: 25-50 years product life-cycle; How to update and upgrade the hardware and software platforms?
- "The Civil Infrastructure Platform (CIP) project aims to provide an open source base layer for embedded systems that require Super Long-Term Support (SLTS). The base layer is composed of the CIP kernel source code, and the CIP Core source packages. CIP Core source packages are a list of Debian Long Term Support (LTS) source packages that members of the CIP project consider important and want to sponsor." https://wiki.linuxfoundation.org/civilinfrastructureplatform/cip-core
- A collaborative effort: Linux + Debian + Companies + Developers
- Details: https://www.cip-project.org/about

## The CIP Picture

Open
Source
Base Layer

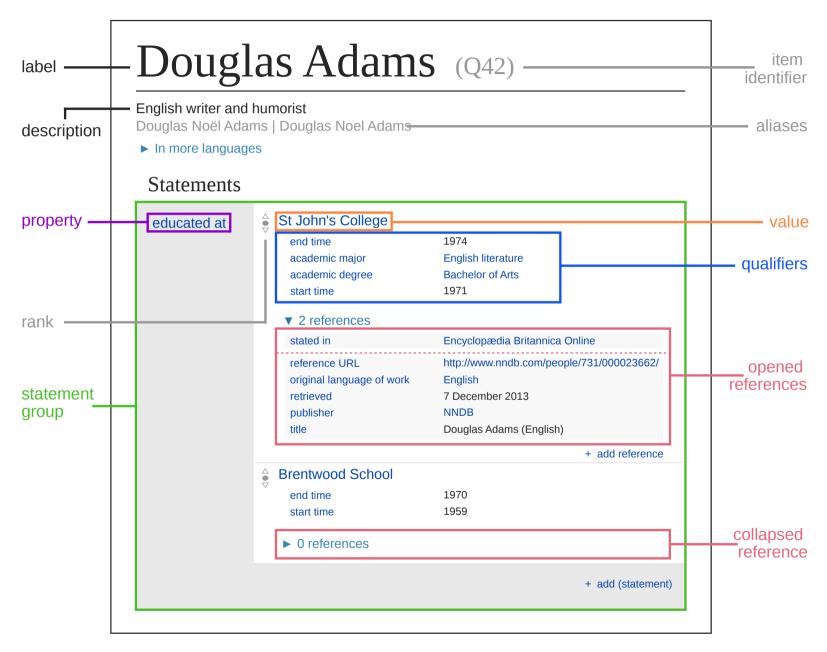
- Open source based reference implementation
- Start from a minimal set for controllers in industrial grade systems

## Non-CIP packages Linux distribution (e.g. Debian) may extend/include CIP packages. User space CIP Reference Filesystem image with SDK (CIP Core packages) Kernel **CIP SLTS Kernel** Hardware CIP Reference Hardware

# Wikidata — "the database to link them all"

- "Wikidata is a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world."
- Free: Data is published under CC0.
- Collaborative: Data is maintained by editors who set the rules.
- Multilingual: Editing in any language is possible and encouraged.
- A secondary database: Wikidata records are also about their sources, and connections to other databases.
- Collecting structured data: Imposing a high degree of structured organization allows for easy reuse of data by third parties
- https://www.wikidata.org/wiki/Wikidata:Introduction

## A Wikidata Item



https://www.wikidata.org/wiki/Wikidata:Introduction

# **Experimentations in Progress**

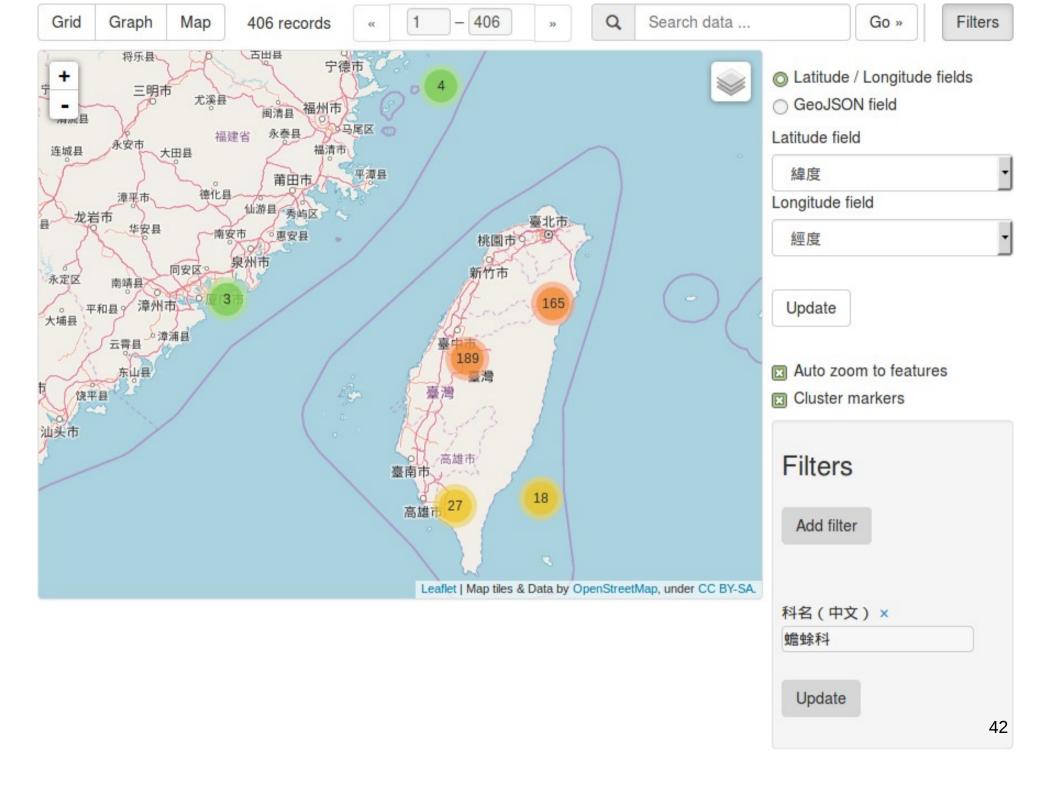
# data.depositar.io

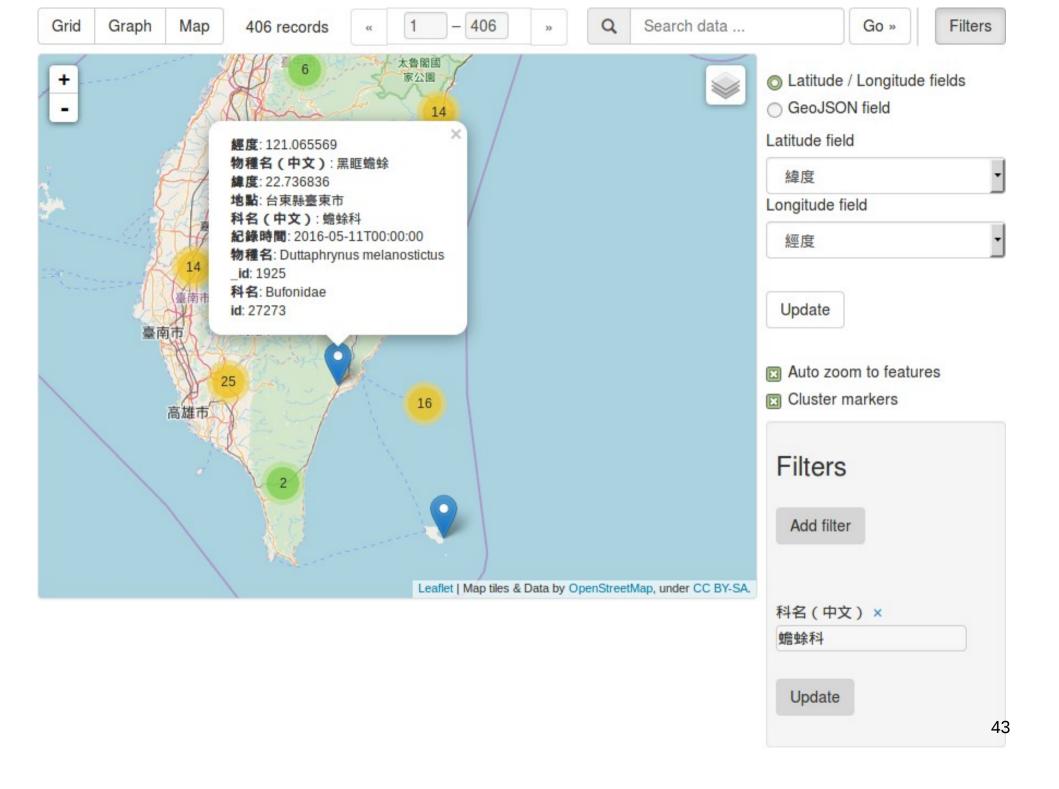
- depositar a Portuguese/Spanish verb for to deposit is an open repository for research data.
- Built on top of CKAN, an open source software package from Open Knowledge Foundation for publishing open data.
- Added extensions in depositar
  - spatiotemporal description and query
  - more data preview and base map
  - simplified metadata authoring
  - keywords sourced from Wikidata
- An online platform for preserving, documenting, and sharing (community generated) datasets.

# Sample Dataset from TaiRON (Taiwan Roadkill Observation Network)

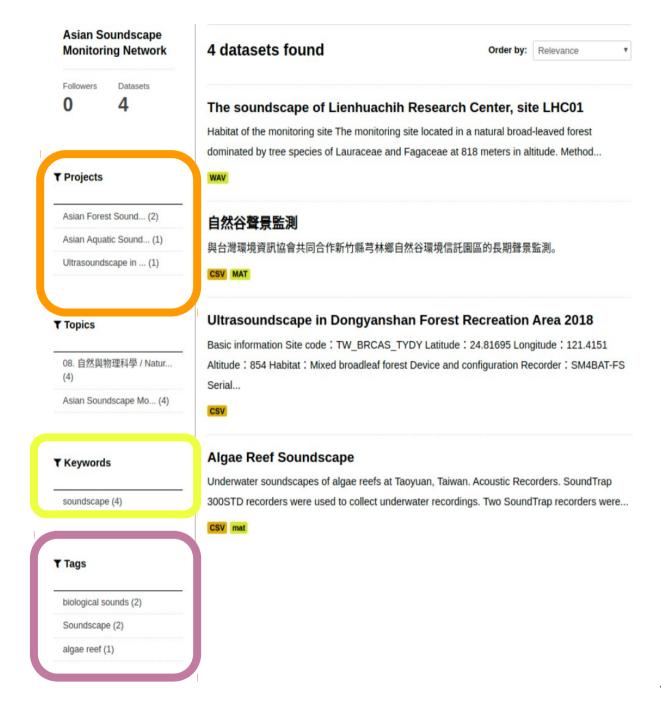
https://data.depositar.io/dataset/2016

Grid	Graph	Мар	3218 re	cords	1 -	20	Q	Search data	•••	Go »	Filters
_id	id		科名	科名(	物種名	物種名	紀錄時間	地點	經度	緯度	
1	2493	3	Bufonidae	蟾蜍科	Rana lon	長腳赤蛙	2016-01	宜蘭縣員	121.636	24.754443	
2	2493	4	Muscica	鶲科	Turdus p	白腹鶇	2016-01	基隆市中	122.08166	25.62861	
3	2493	5	Muridae	鼠科	Bandicot	鬼鼠	2016-01	高雄市茄	120.195	22.894975	
4	2494	4	Talpidae	鼴鼠科	Mogera i	台灣鼴鼠	2016-01	台中市北	120.788	24.173821	
5	2497	1	Colubridae	黃頷蛇科	Lycodon	白梅花蛇	2016-01	新竹縣峨	120.98417	24.698021	
6	2497	2	Viperidae	蝮蛇科	Trimeres	赤尾青竹絲	2016-01	苗栗縣頭	120.974	24.702631	
7	2500	4	Bufonidae	蟾蜍科	Bufo ban	盤古蟾蜍	2016-01	台北市士	121.538	25.127258	
8	2500	6	Potamidae	溪蟹科	Geothel	黃綠澤蟹	2016-01	高雄市大	120.408	22.725732	
9	2500	8	Laniidae	伯勞科	Lanius c	紅尾伯勞	2016-01	台南市柳	120.349	23.261861	
10	2500	9	Muridae	鼠科	Rattus n	溝鼠	2016-01	台中市潭	120.724	24.231545	
11	2502	3	Rallidae	秧雞科	Amauror	白胸苦惡鳥	2016-01	金門縣金	118.412	24.508996	
12	2502	5	Columbi	鳩鴿科	Streptop	珠頸斑鳩	2016-01	金門縣金	118.319	24.471015	
13	2502	6	Bufonidae	蟾蜍科	Duttaphr	黑眶蟾蜍	2016-01	台北市文	121.538	25.008142	
14	2502	7	Columbi	鳩鴿科	Columba	岩鴿	2016-01	新北市貢	121.984	24.99522	
15	2502	8	Rhacoph	樹蛙科	Rhacoph	台北樹蛙	2016-01	基隆市暖	121.734	25.091321	
16	2502	9	Rhacoph	樹蛙科	Rhacoph	台北樹蛙	2016-01	基隆市暖	121.734	25.091321	
17	2503	0	Rhacoph	樹蛙科	Rhacoph	台北樹蛙	2016-01	基隆市暖	121.734	25.091321	
18	2503	2	Colubridae	黃頷蛇科	Ptyas dh	過山刀	2016-01	台中市太	120.7971	24.10225	
19	2503	3	Strigidae	鴟鴞科	Strigidae	鴟鴞科	2016-01	台北市文	121.586	24.982227	
20	2503	4	Sturnidae	椋鳥科	Acridoth	家八哥	2016-01	金門縣金	118.413	24.507908	





## Asian Soundscape Monitoring Network





**追蹤者** 

n

〇跟隨

田 專案



#### Asian Aquatic Sounds cape

此專案沒有任何說明

#### 世 社交

Google+

☑ Twitter

**■ Facebook**

#### △授權(

CC-BY 4.0 OPEN DATA

#### O Dataset extent



#### 資料集 警主題 O動態牆 P歷史紀錄

#### Algal Reef Soundscape

Underwater soundscapes of algal reefs at Taoyuan, Taiwan.

Acoustic Recorders. SoundTrap 300STD recorders were used to collect underwater recordings. Two SoundTrap recorders were used in three surveys: SoundTrap 300STD #1 SN1208004643; SoundTrap 300STD #2 SN805347354.

Configuration of Audio Recording. (1) Duration: 5 min. (2) Sampling Rate: 96kHz. (3) Channels: 1. (4) File Format: WAV (converted by SUD). (5) Audio Gain: High. (6) High Pass Filter: Off.

Field Deployment. Each SoundTrap recorder was fixed vertically on an angle steel at the bottom. The recorder was exposed in the air during low tide, so the collection of underwater recordings were only possible 3 hours before and after the high water.

Data Processing. We used the LTSA\_gui (https://github.com/schonkopf/long-term-spectrogram) to visualize the spectral-temporal change of long-duration recordings. By the median-based long-term spectrogram, we employed the periodicity-coded non-negative matrix factorization (https://github.com/schonkopf/soundscape-viewer) to separate the audio data into two components: (1) high-frequency biological chorus, and (2) low-frequency ambient noise. The relative intensities of the two components were analyzed.

## **Basic site information**

#### 資料與資源

## Spa

#### Spatiotemporal change of soundscape

This csv describes the Location, Latitude, Longitude, Survey Number, Tide...



Field recordings at BY

Audio archive on Asian Soundscape



Field recordings at G1

Audio archive on Asian Soundscape



Field recordings at G2

Audio archive on Asian Soundscape



Field recordings at YX

Audio archive on Asian Soundscape



Field recordings at YA

Audio archive on Asian Soundscape



Long-term spectrogram of the Survey 1 at BY

Please use Soundscape\_Viewer (https://github.com/schonkopt/soundscape-viewer)...



Long-term spectrogram of the Survey 1 at G1

Spatiotemporal change of soundscape (.csv)



→ 探索

**人管理** 

**Audio archive (Pumilio)** 



Visualization of acoustic data (.mat)

## data.depositar.io

## **Topics**

- Asian Soundscape Monitoring Network
- Natural and physical sciences

# Project Asian Aquatic Soundscape

Dataset site1

Dataset site2

.

## **Project**

Asian Forest Soundscape

Dataset site1

Dataset site2

•

## **Project**

Ultrasoundscape in Taiwan

Dataset site1

Dataset site2

•

### **Keywords**

- biodiversity
- geophony
- biophony
- anthrophony
- Vietnam \ Thailand
- Soundscape ...

### **Tags**

- biological sounds
- long-term monitoring
- Chiroptera ...

### **Formats**

• csv \ mat ...

# Wikidata-powered Keywords

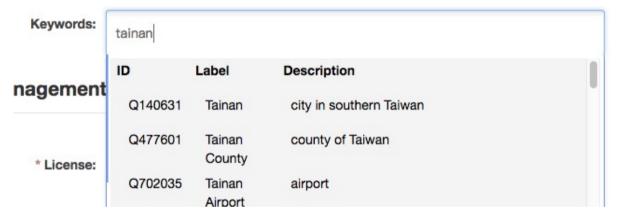
- Keywords
  - controlled vocabularies for tagging datasets
  - DIY vocabulary design: a never-ending story ...
- Use Wikidata as the source for keywords
  - 70M+ entries
  - multilingual



- semantic relations enable data inference
  - · e.g. Taipei is part of Taiwan

# Wikidata Items as Keywords

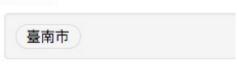
Search and select keywords:



Keywords (as Wikidata IDs) are stored.

Viewed in English:	Keywords			
	Tainan			

Viewed in Chinese: 關鍵字



# Observations & Viewpoints

- Communal data projects organizing themselves to produce datasets of community/public goods.
- Mutual beneficial to researchers to understand and work with them, and to initiate new data projects.
- The need to collaborate across traditional boundaries.
- Communal data production closely related to sustainable digital preservation.
- Key components: free software, open license, common format, and community process.

## Thank You!

http://data.depositar.io/

data.contact@depositar.io



