A Persistent Identifier Practice For A Research Data Repository

depositar
研究資料寄存所

The 18th International Conference on Open Repositories (OR2023)
June 13, 2023

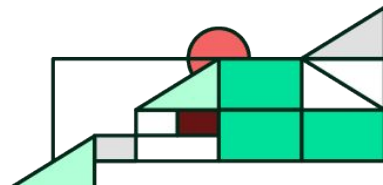
Cheng-Jen Lee, Tyng-Ruey Chuang
Institute of Information Science, Academia Sinica, Taiwan

Outline

- Introduction to ARKs
- ARKs in the *depositar* — a research data repository
- Discussions

Slides: https://n2t.net/ark:37281/k562n4m0z

Introduction to ARKs



Archival Resource Keys (ARKs)

- A multi-purpose URL scheme suited to being a **persistent identifier** for information objects of any type since 2001.
- Similar to **DOIs**, **URNs**, and **Handles**.
- Open, mainstream, non-paywalled, and decentralized
- 8.2 billion ARK numbers from 1,000+ organizations (as of 2023)
  - E.g., BnF, Internet Archive, and Louvre Museum
- Example: [ark:/53355/cl010066723](ark:/53355/cl010066723)

The ARK Alliance: https://arks.org/

Generic ARK Services

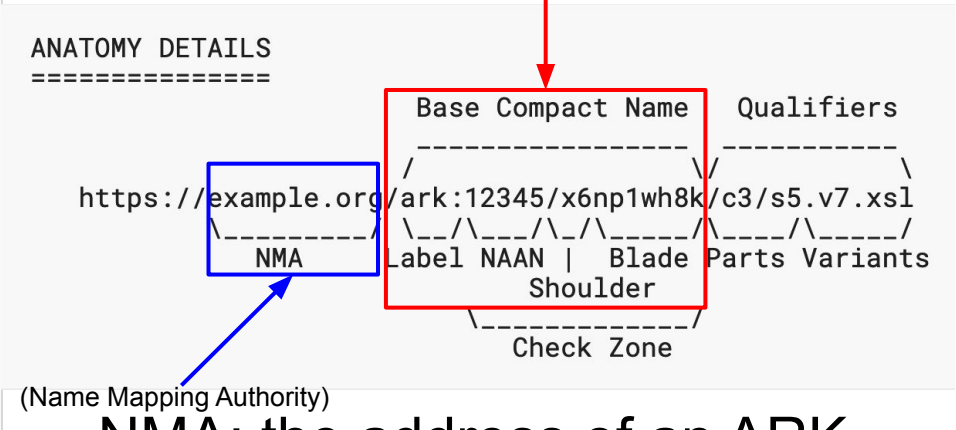
Defined in The ARK Identifier Scheme
(IETF Active Internet-Draft)

- Generic Access Service (*Resolver*)
  - Return the location of the target object via URL redirection
- Generic Policy Service
  - Return declarations of policy and support commitments for given ARKs
- Generic Description Service (*Registry*)
  - Return a description of the target object (ARK URL + ?info)


```
"erc": {
    "what": "Science Europe 研究資料管理指南 | RDM Guides from Science Europe",
    "when": "2020-2021",
    "where": "https://pid.depositar.io/ark:37281/k516v4d6w",
    "who": "Science Europe & 研究資料寄存所 | depositar"
},
```


The description of the object — Electronic Resource Citation (ERC)
https://pid.depositar.io/ark:37281/k516v4d6w?info

ARK Anatomy



- NMA: the address of an ARK service

- Base Compact Name
  - The ark: label*
  - NAAN*(Name Assigning Authority Number)
    - Free registration to get a namespace
  - Shoulder: a string to subdivide a NAAN namespace

Blade*: a unique name for the target object generated by noid (Nice Opaque IDentifier) tool

* mandatory part

Example Blade Template

Quasi-randomly generated

An extended-digit, one of { 0123456789bcdf ghjkmnpqrstvwxz }

A pure digit { 0-9 }

Check character

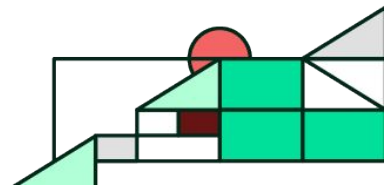
r e d e d e d k

Cardinality: 29 * 10 * 29 * 10 * 29 * 10 = 24,389,000 identifiers

The Decentralized Resolvers

- Local resolver service
  - https://ark.archive.org/ for example
- Global resolver service (longer-lived)
  - https://n2t.net/ (Name-to-Thing) hosted at the California Digital Library

ARKs in the *depositar*

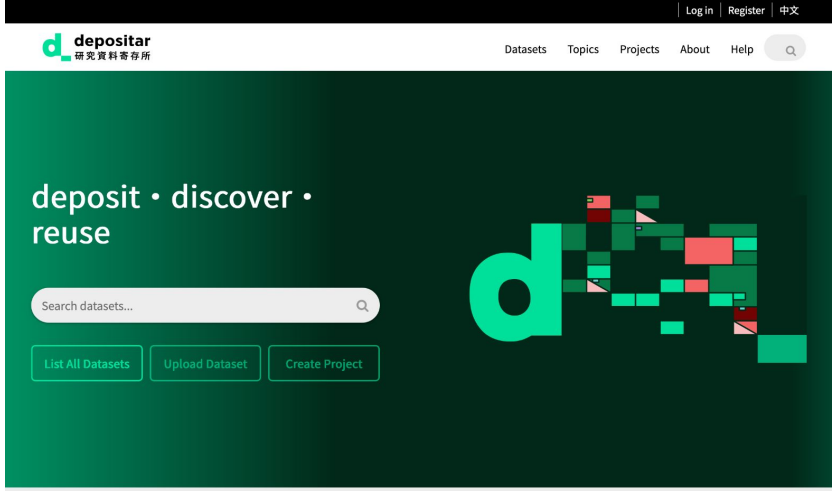


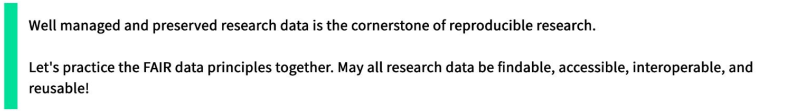
About the *depositar*

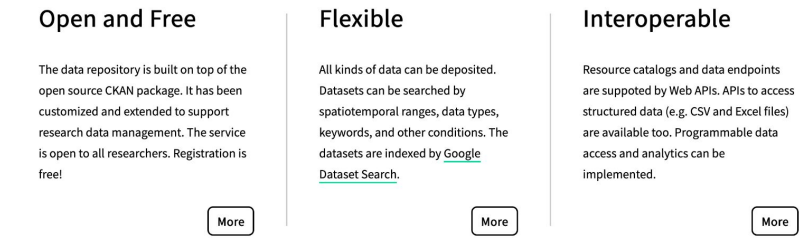
- A data repository open to researchers worldwide for the deposit, discovery, and reuse of datasets since 2018
- Built on top of CKAN , an open source data portal
- 1,716 datasets & 203k views (as of May 2023)

data.depositar.io

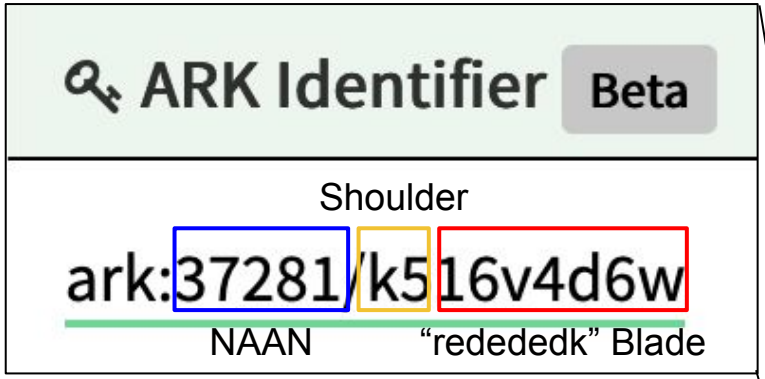
Learn more about the depositar: https://data.depositar.io/en/about







Example



https://pid.depositar.io/ark:37281/k516v4d6w

Science Europe 研究資料管理指南 | RDM Guides from Science Europe

Followers
2

Project

研究資料管理入門
本專案收集研究資料管理相關的入門文件。

read more

Social
Twitter
Facebook

License
CC-BY 4.0

ARK Identifier Beta
ark:37281/k516v4d6w

Dataset Topics Activity Stream Showcases

Science Europe 研究資料管理指南 | RDM Guides from Science Europe

The Practical Guide to Research Data Management (RDM) published by Science Europe and its translation into Traditional Chinese produced by the *depositar* team.

由 Science Europe 所出版的研究資料管理指南。包含原始手冊以及由「研究資料寄存所」(*depositar*) 團隊所完成的中文翻譯版本手冊。

Data and Resources

Practical Guide to the International Alignment... 🔥 Explore
《國際合用的研究資料管理實用指南 — 增訂版》英文手冊。台灣華語翻譯版本亦收錄在本資料集。This is the extended edition of...

國際合用的研究資料管理實用指南—增訂版 🔥 Explore
Science Europe 所出版的 Practical Guide to the International Alignment of...

Tags

RDM | Science Europe | data managment plan | research data manag...

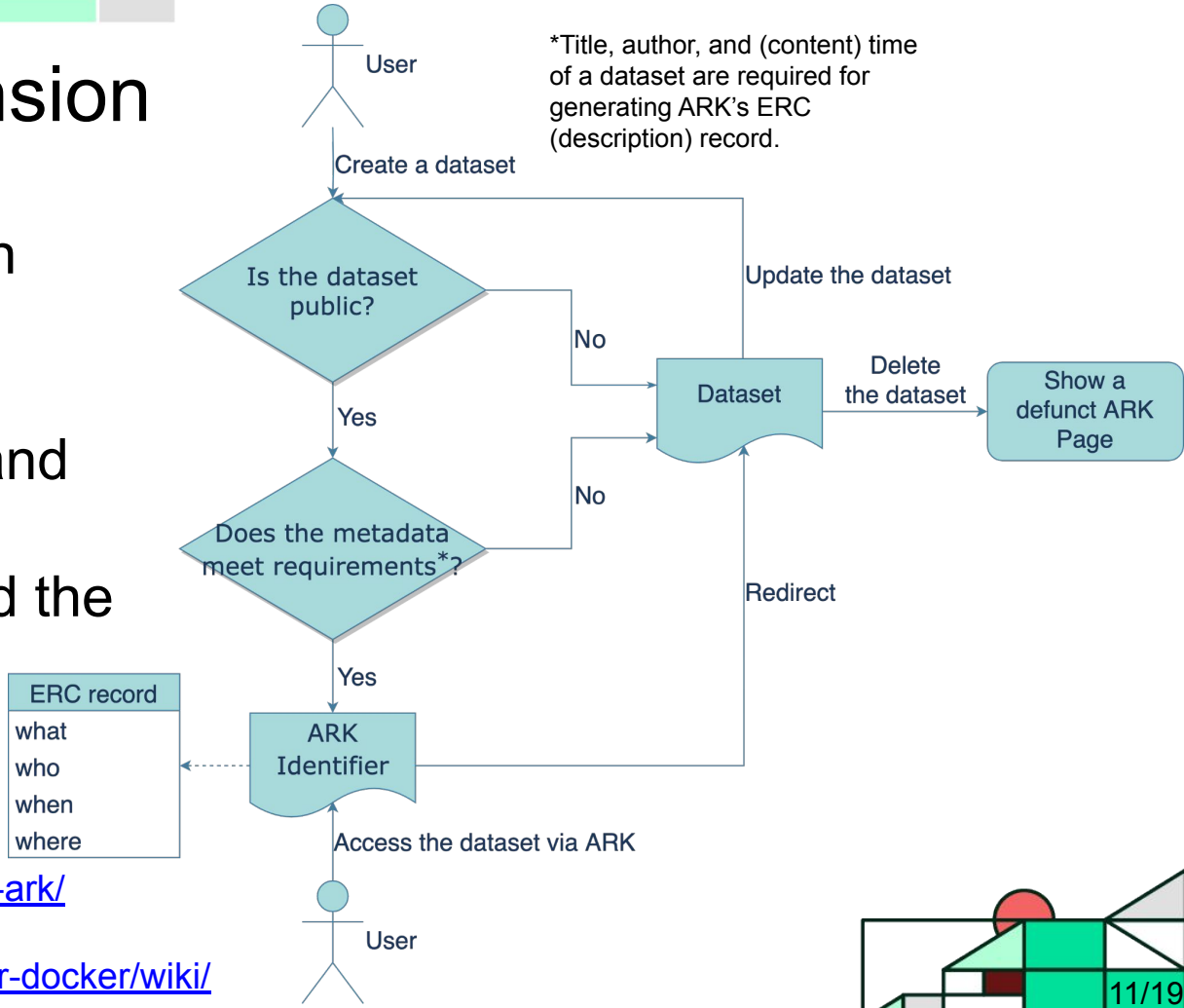
Wikidata Keywords

Science Europe | research data management | depositar | data library | data management plan | Academia Sinica | Institute of Information Science, Academia Sinica | Research Center for Information Technology Innovation, Academia Sinica

ckanext-ark Extension

A CKAN extension which provides:

- A registry for ARKs and their ERC records
- A resolver to respond the ARK URLs

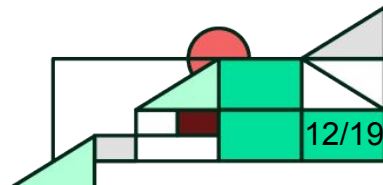


PyPI: https://pypi.org/project/ckanext-ark/
Testing Docker images: https://github.com/depositar/depositar-docker/wiki/Quick-Setup

The Registry: Mint and Bind an ARK

- The noid-mint Python package is used to generate ARKs.

- The binding *granularity*: what is the **target object**?
  - **Dataset:** resources and metadata about the data (current)
  - **Resource:** the data itself
  - **Version:** the snapshots of the dataset (the most ideal)



The Resolver: Make Flask Redirects

- Case #1: show the defunct ARK page if the target object (dataset) has been deleted

ckanext/ark/views.py

```
@blueprints.route('/ark:/<path:path>/')
@blueprints.route('/ark:<path:path>/')
def read(path):
    ...
    else:
        try:
            toolkit.get_action('package_show')({}, {
                'id': ark.package_id
            })
            return toolkit.redirect_to('dataset.read',
                                       id=ark.package_id)
        except (toolkit.ObjectNotFound, toolkit.NotAuthorized):
            # Show defunct page
            return toolkit.render('ark/snippets/defunct.html',
                                  {'ark': ark.identifier})
```

depositar
研究資料寄存所

Defunct ARK

The dataset with ARK identifier: 37281/k562n4m0z is not found.

However, the ERC metadata is still available.

The Resolver: Make Flask Redirects

- Case #2 (/ark:NAAN/): show the policy and support commitments for given ARKs(Generic Policy Service)

```
The depositar, Institute of Information Science,
Academia Sinica, Taiwan assigns identifiers within the
ARK domain under the NAAN 37281 and according to the
following principles:
* No re-assignment.  Once a base identifier-to-object
association has been made public, that association
shall remain unique into the indefinite future.
* Opacity.  Base identifiers shall be assigned with no
widely recognizable semantic information.
* A check character is generated in assigned
identifiers to guard against common transcription
errors.
```

https://pid.depositar.io/ark:37281/

ckanext/ark/views.py

```python
@blueprints.route('/ark:/<path:path>/')
@blueprints.route('/ark:<path:path>/')
def read(path):
    # Show NAA metadata
    if path == toolkit.config.get('ckanext.ark.naan'):
        response = make_response(get_erc_support_commitment())
        response.headers['Content-type'] = 'text/plain; charset=UTF-8'
        return response
    ark = ARKQuery.read_ark(path)
    if not ark:
        return base.abort(404, _('ARK not found'))
    # Show ERC metadata
    if 'info' in request.args or request.environ['REQUEST_URI'][-2:] == '/?':
        response = {
            'erc': {
                'who': ark.who,
                'what': ark.what,
                'when': ark.when,
                'where': get_ark_url(ark.identifier)
            },
            'erc-support': get_erc_support()
        }
        response = make_response(response)
        response.headers['Content-type'] = 'application/json; charset=UTF-8'
        return response
```

The Resolver: Make Flask Redirects (Cont'd)

- Case #3 (ARK + ?info): return the ERC record of the target object(Generic Description Service)

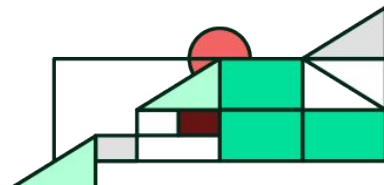

```
{
  "erc": {
    "what": "Science Europe 研究資料管理指南 | RDM Guides from Science Europe",
    "when": "2020-2021",
    "where": "https://pid.depositar.io/ark:37281/k516v4d6w",
    "who": "Science Europe & 研究資料寄存所 | depositar"
  },
  "erc-support": {
    "what": "Permanent: Dynamic Content:",
    "when": "20220708",
    "where": "https://pid.depositar.io/ark:37281",
    "who": "The depositar | Institute of Information Science, Academia Sinica, Taiwan"
  }
}
```


https://pid.depositar.io/ark:37281/k516v4d6w?info

ckanext/ark/views.py

```python
@blueprints.route('/ark:/<path:path>/')
@blueprints.route('/ark:<path:path>/')
def read(path):
    # Show NAA metadata
    if path == toolkit.config.get('ckanext.ark.naan'):
        response = make_response(get_erc_support_commitment())
        response.headers['Content-type'] = 'text/plain; charset=UTF-8'
        return response
    ark = ARKQuery.read_ark(path)
    if not ark:
        return base.abort(404, _('ARK not found'))
    # Show ERC metadata
    if 'info' in request.args or request.environ['REQUEST_URI'][-2:] == '/?':
        response = {
            'erc': {
                'who': ark.who,
                'what': ark.what,
                'when': ark.when,
                'where': get_ark_url(ark.identifier)
            },
            'erc-support': get_erc_support()
        }
        response = make_response(response)
        response.headers['Content-type'] = 'application/json; charset=UTF-8'
        return response
```

Discussions



Long-term Sustainability of ARKs

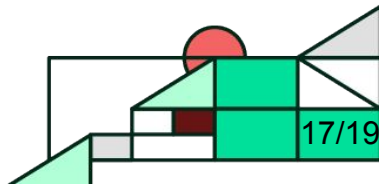
The promise of support from the institute is the key factor to make a PID sustainable.

"...All those identifiers rely critically on thousands of institutional web servers that have adopted ARKs and DOIs, respectively... So in regard to the main PID function of providing long term access, the ARK and DOI infrastructures could be seen as comparable. "

— John Kunze, Author of ARKs
@ ARKs Forum (2023-04-19)

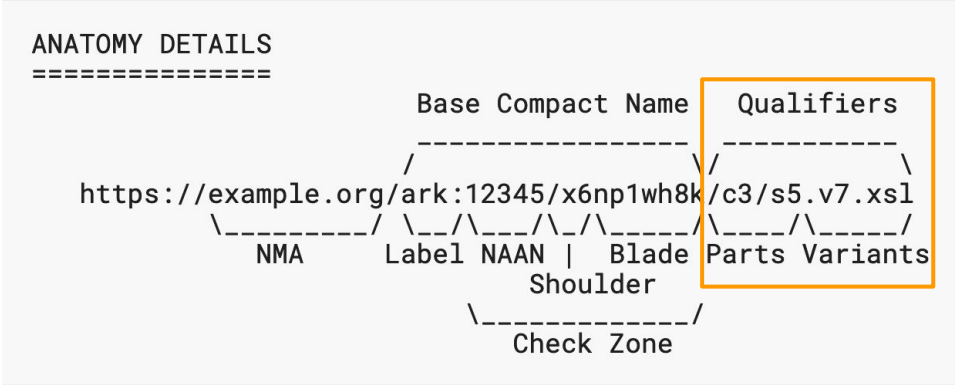
"... ARKs can be implemented directly on a local web server. While some consider this a weakness, citing the "inherent" fragility of DNS names, their argument usually suggests using dx.doi.org, handle.net, or n2t.net instead; the logical flaw is that these are DNS names too, and we note that none of them are as long-lived as bnf.fr."

— The ARK Identifier Scheme: Lessons Learnt at the BnF and Questions Yet Unanswered (2014-10-08)



Future Works

- Assign ARKs to resources (data itself)
- Provide a **resolution report** (statistics) like the way DataCite and Crossref do
- Make the ARK *actionable* via qualifiers
  - Qualifiers: show the hierarchy or variants of the target object
    - E.g., open csv files in plain text or grid view



@_depositar

depositar
研究資料寄存所

謝謝！ Thank You!

https://data.depositar.io/ The *depositar*
https://lab.depositar.io/ The depositar lab

data.contact@depositar.io

The *depositar* is a collaboration at the Institute of Information Science, the Research Center forInformation Technology Innovation, and the Research Center for Humanities and Social Sciences (GIS Center) in Academia Sinica, Taiwan. The project has been supported, in part, by grants from Taiwan's National Science and Technology Council.
The *depositar* project team: T-R Chuang, M-S Ho, C-J Lee & C-H Ally Wang.

「研究資料寄存所」是中央研究院資訊科學研究所、資訊科技創新研究中心、人文社會科學研究中心（地理資訊科學研究專題中心）的協作專案，部份經費來自台灣國科會的專題研究計畫。
研究資料寄存所計畫成員：莊庭瑞、何明諠、李承鑫、王家薰。