aarnet

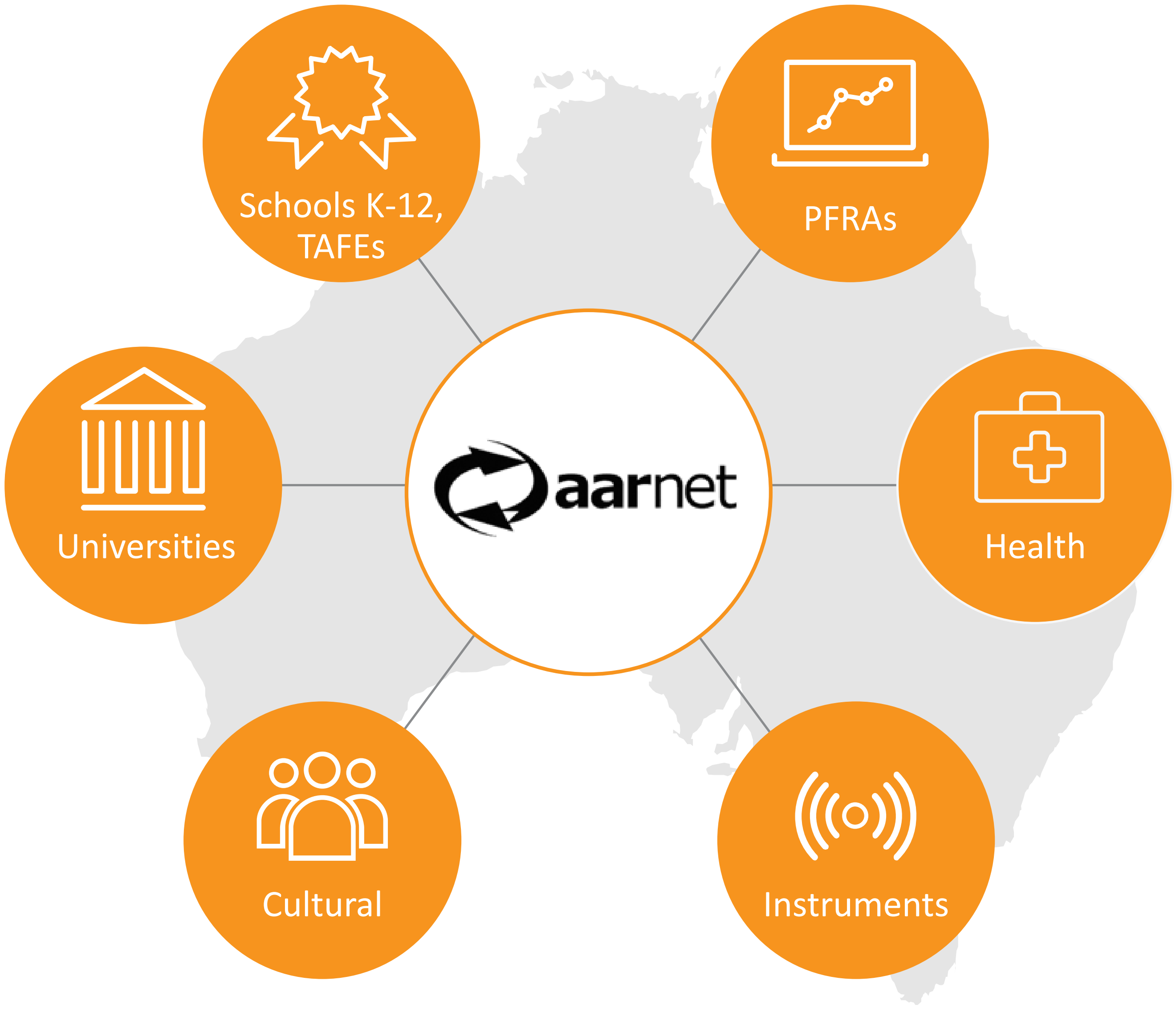
A SIDE SERVE OF METADATA

EVOLVING AARNET'S TOOLS FOR DATA PACKAGING

GAVIN KENNEDY
AARNET CLOUD SERVICES PRODUCT MANAGER

JANUARY 2019

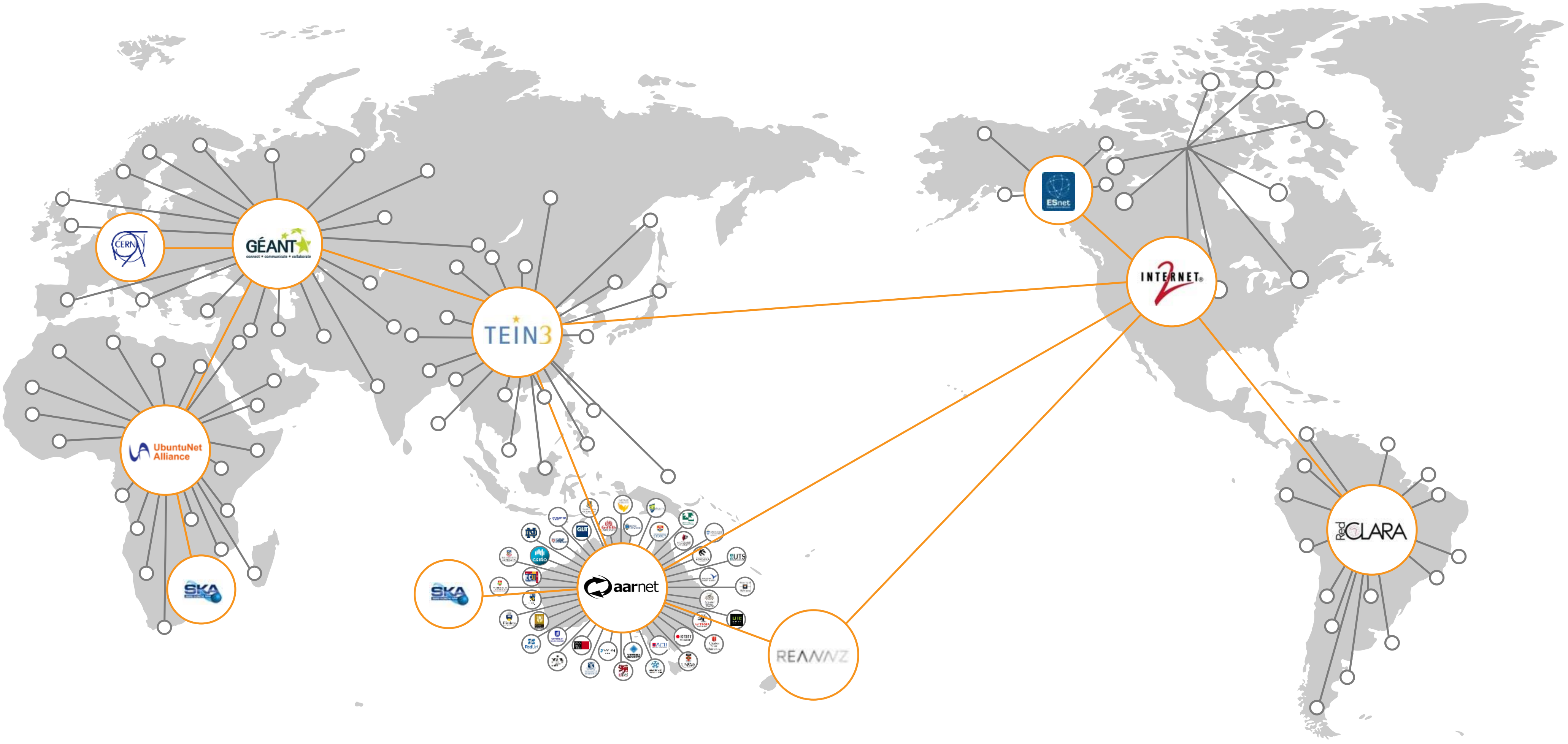
NATIONAL RESEARCH AND EDUCATION COMMUNITY



cloudstor

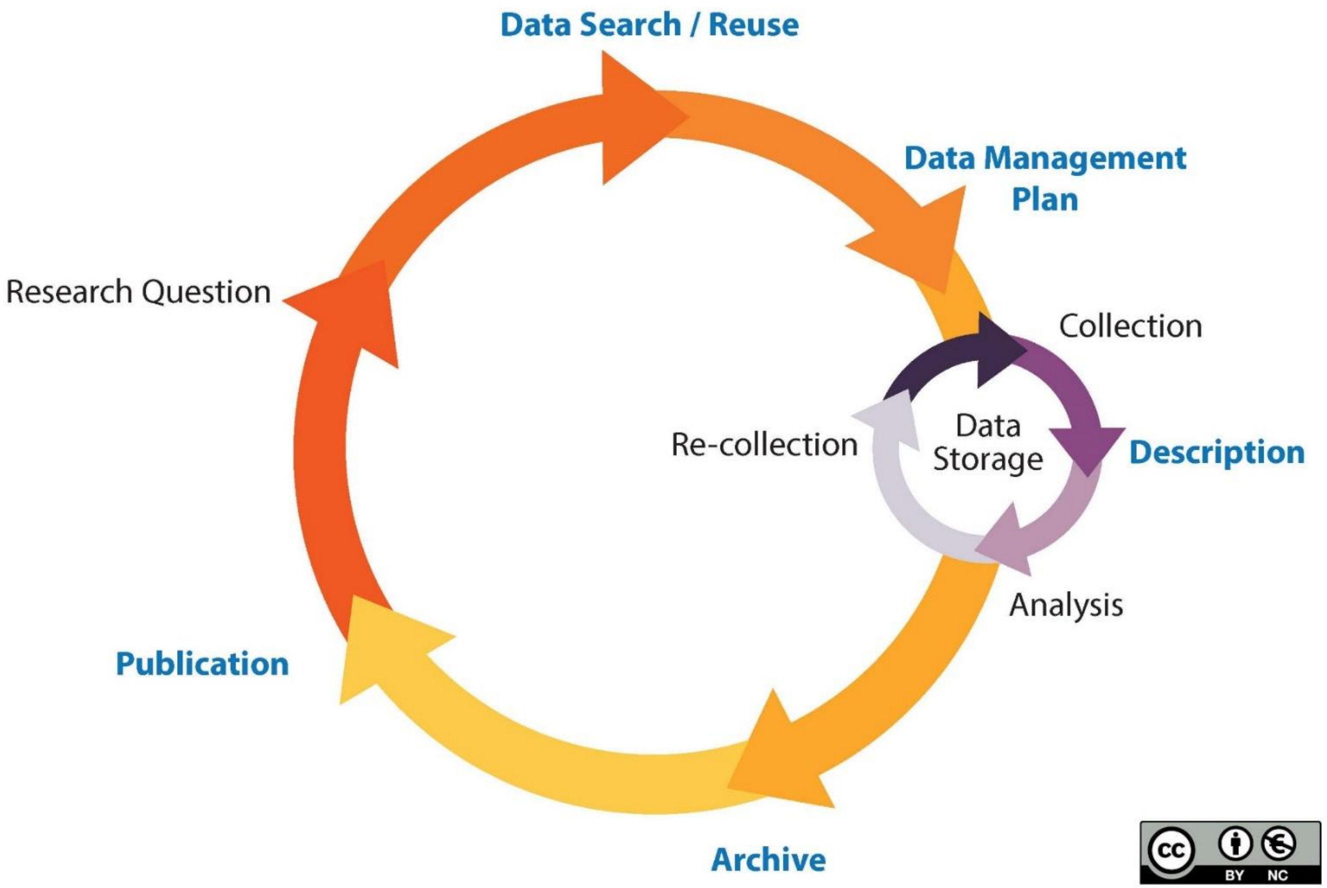
- ✓ Storage
- ✓ Sync & Share
- ✓ FileSender
- ✓ Rocket
- ✓ Swan Jupyter Notebooks
- ✓ Image Viewers
- ✓ Audio Players
- ✓ Data Packaging

INTERNATIONAL NATIONAL RESEARCH NETWORK COMMUNITY



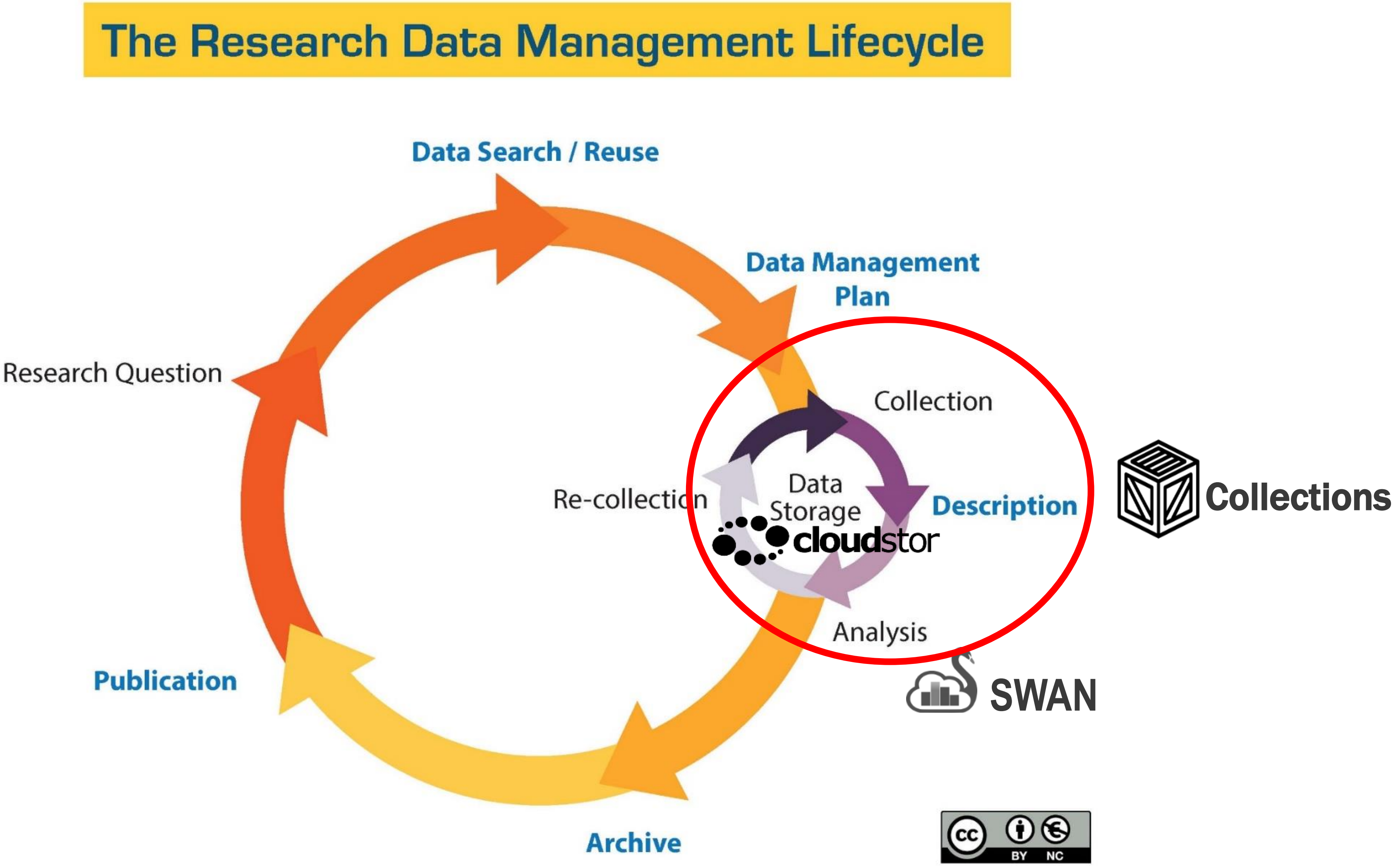
RESEARCH DATA MANAGEMENT LIFECYCLE

The Research Data Management Lifecycle



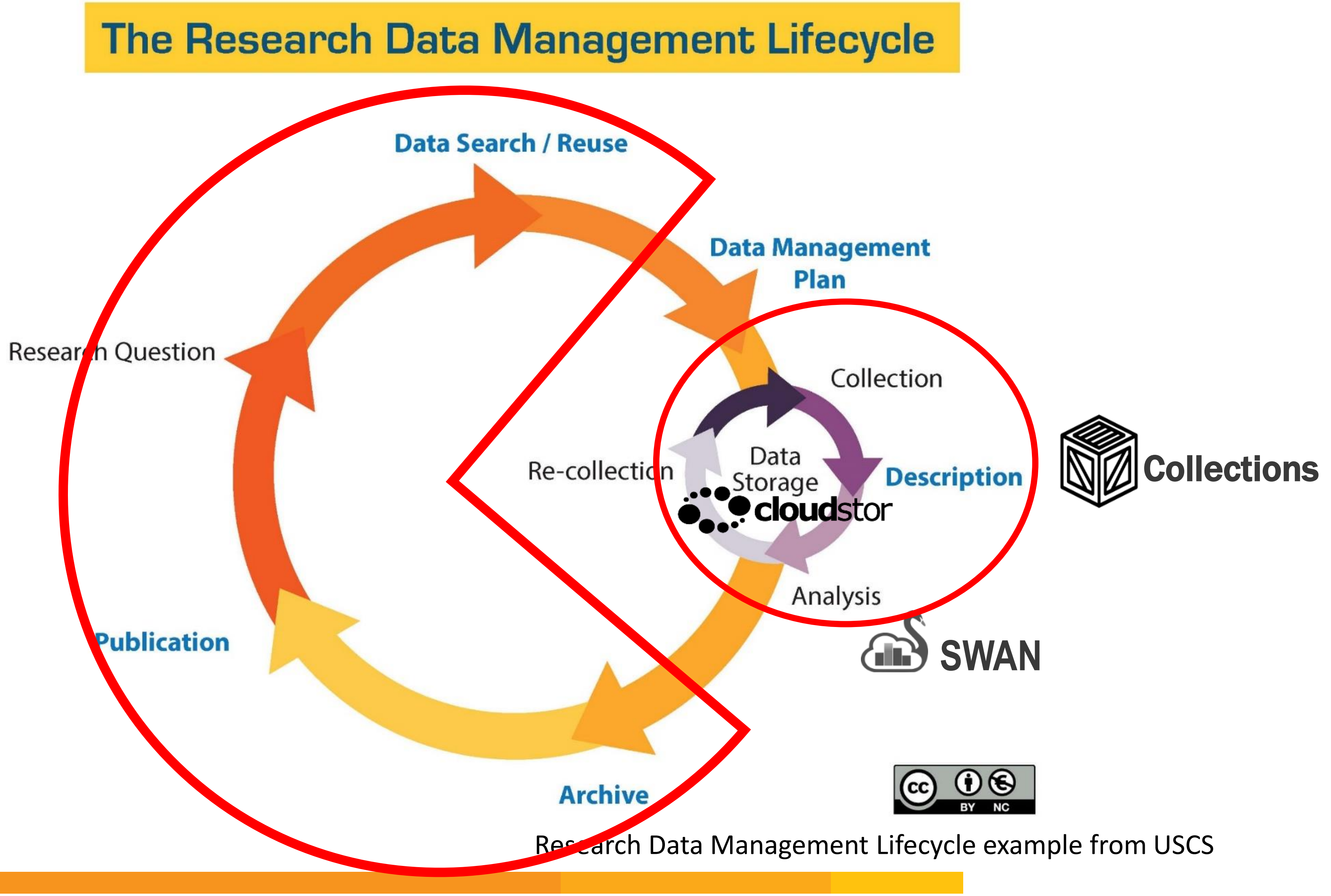
Research Data Management Lifecycle example from USCS

RESEARCH DATA MANAGEMENT LIFECYCLE



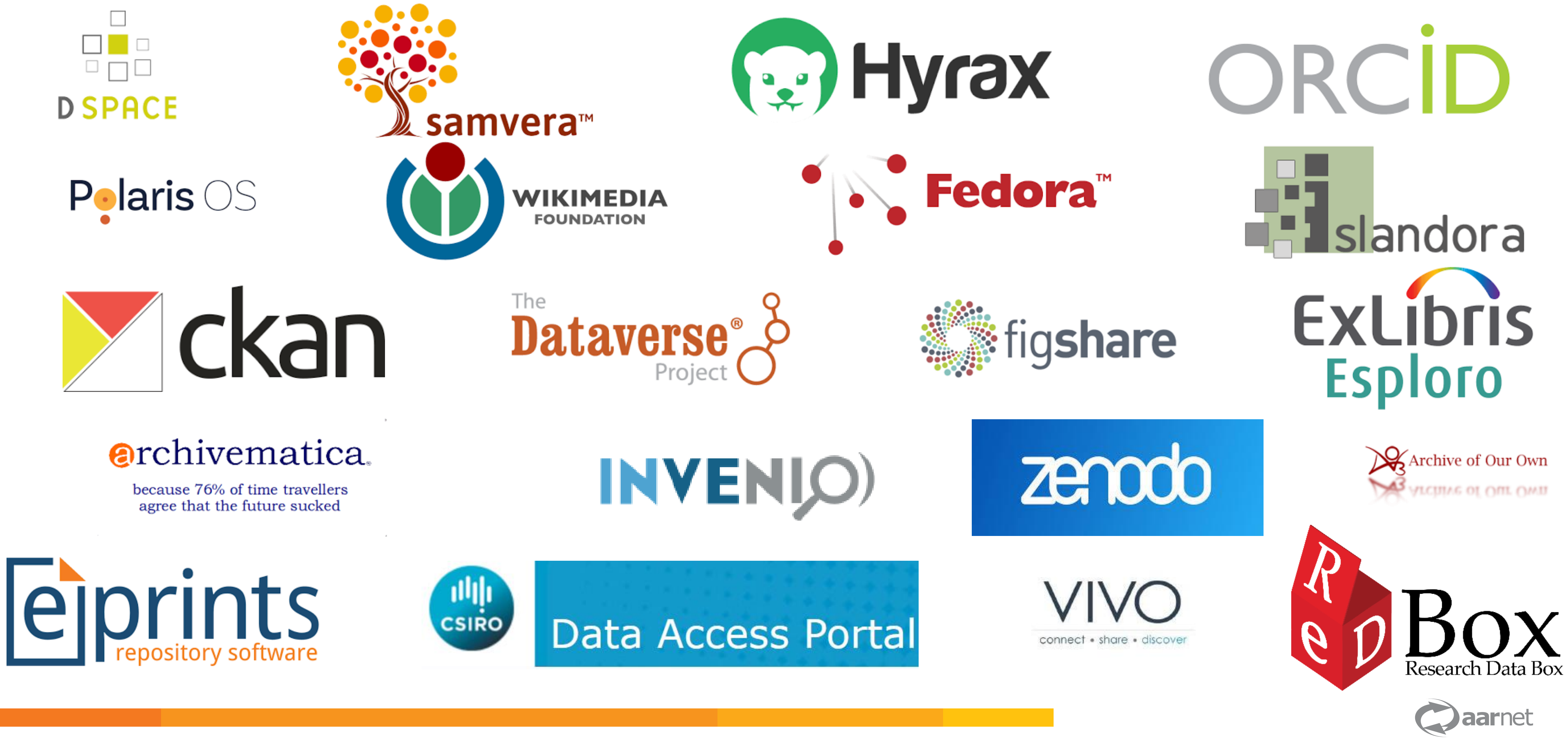
Research Data Management Lifecycle example from USCS

RESEARCH DATA MANAGEMENT LIFECYCLE



Research Data Management Lifecycle example from USCS

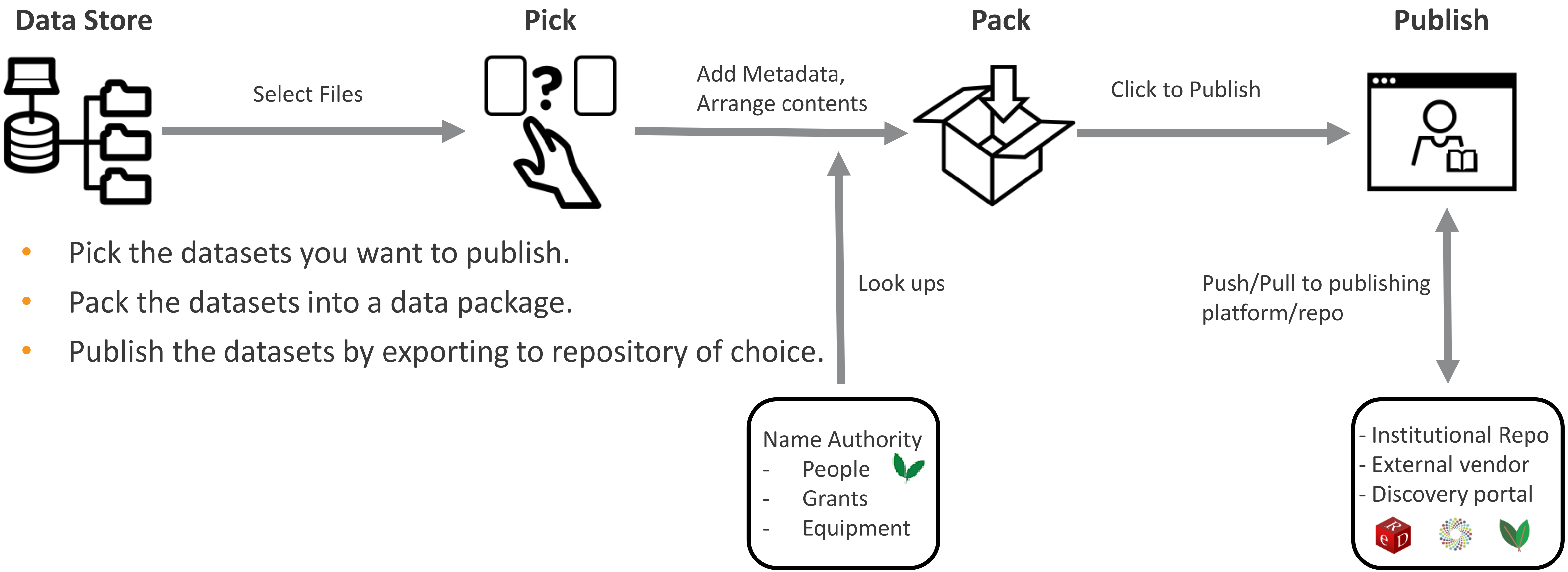
RESEARCH DATA MANAGEMENT – OPEN REPOSITORIES



DATA PACKAGING

- **The combination of a dataset with metadata that describes the dataset.**
- **Provide sufficient contextual data to make the dataset (re-)usable to others.**
- **Basis of “Loosely Coupled” data management: information in the dataset and not the platform.**
- **Facilitates depositing of research data to repositories.**
- **Facilitates sharing of the data (FAIR principles)**
- **Data packages typically contain:**
  - The data
  - The metadata
  - Renditions - metadata outputs in a preferred format
    - Formats incl HTML, XML, JSON-LD
    - Schemas incl DC, MARC, Schema.org

PICK, PACK AND PUBLISH



Data Store

Select Files

Pick

Add Metadata,
Arrange contents

Pack

Click to Publish

Publish

Look ups

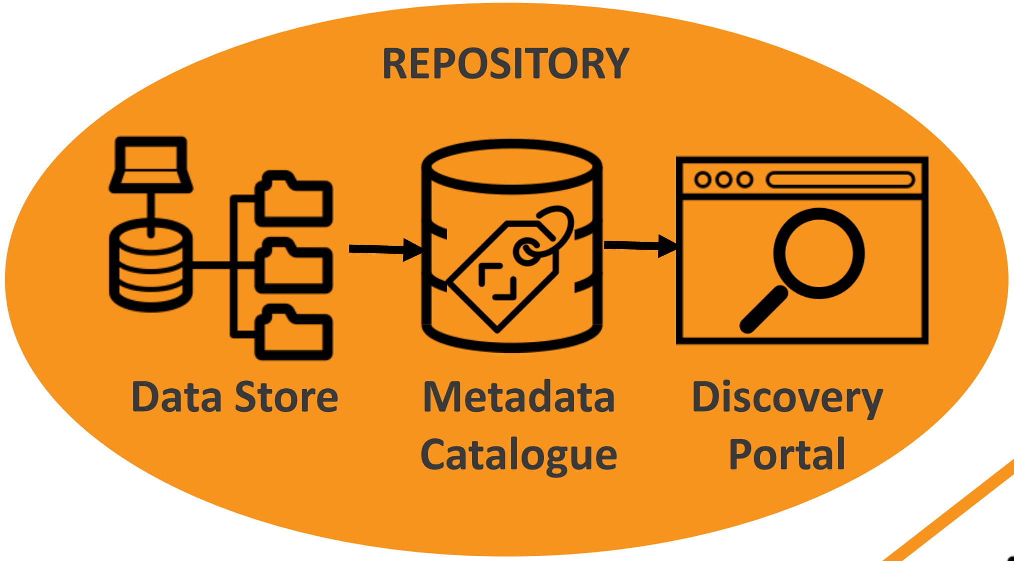
Name Authority
- People
- Grants
- Equipment

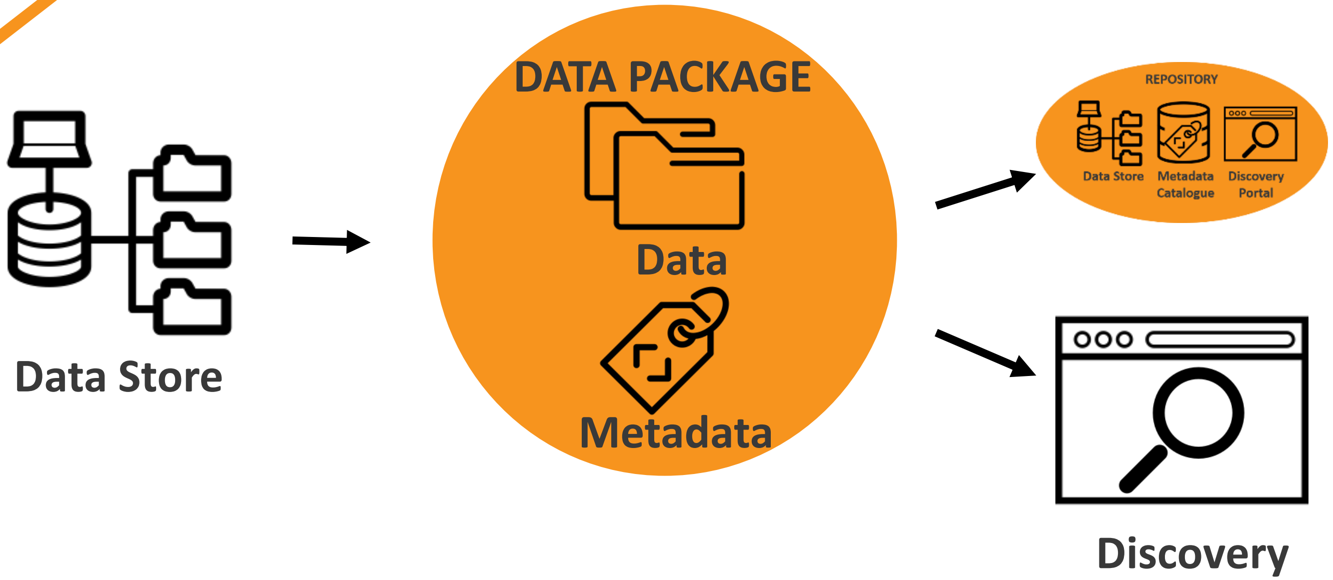
Push/Pull to publishing platform/repo

- Institutional Repo
- External vendor
- Discovery portal

- Pick the datasets you want to publish.
- Pack the datasets into a data package.
- Publish the datasets by exporting to repository of choice.

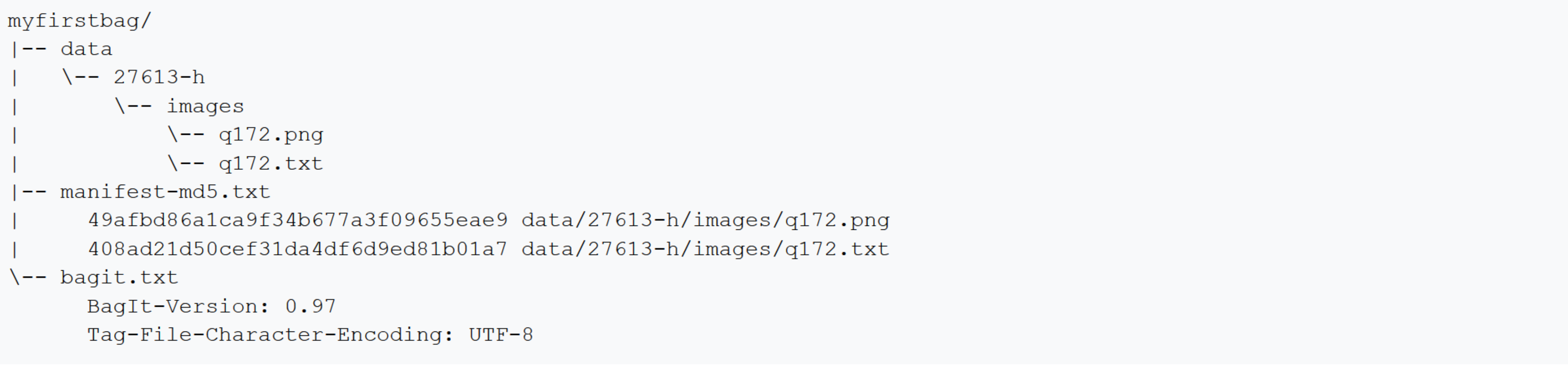
REPO VS DATA PACKAGE





BAGIT

- **Foundational specification for Data Packaging**
- **A ‘Bag’ contains**
  - Payload: File content in a hierarchical file packaging format.
  - Tags: Metadata files documenting the Payload.
  - Checksums: File checksums (e.g. MD5) for validation of shared contents (files and metadata).



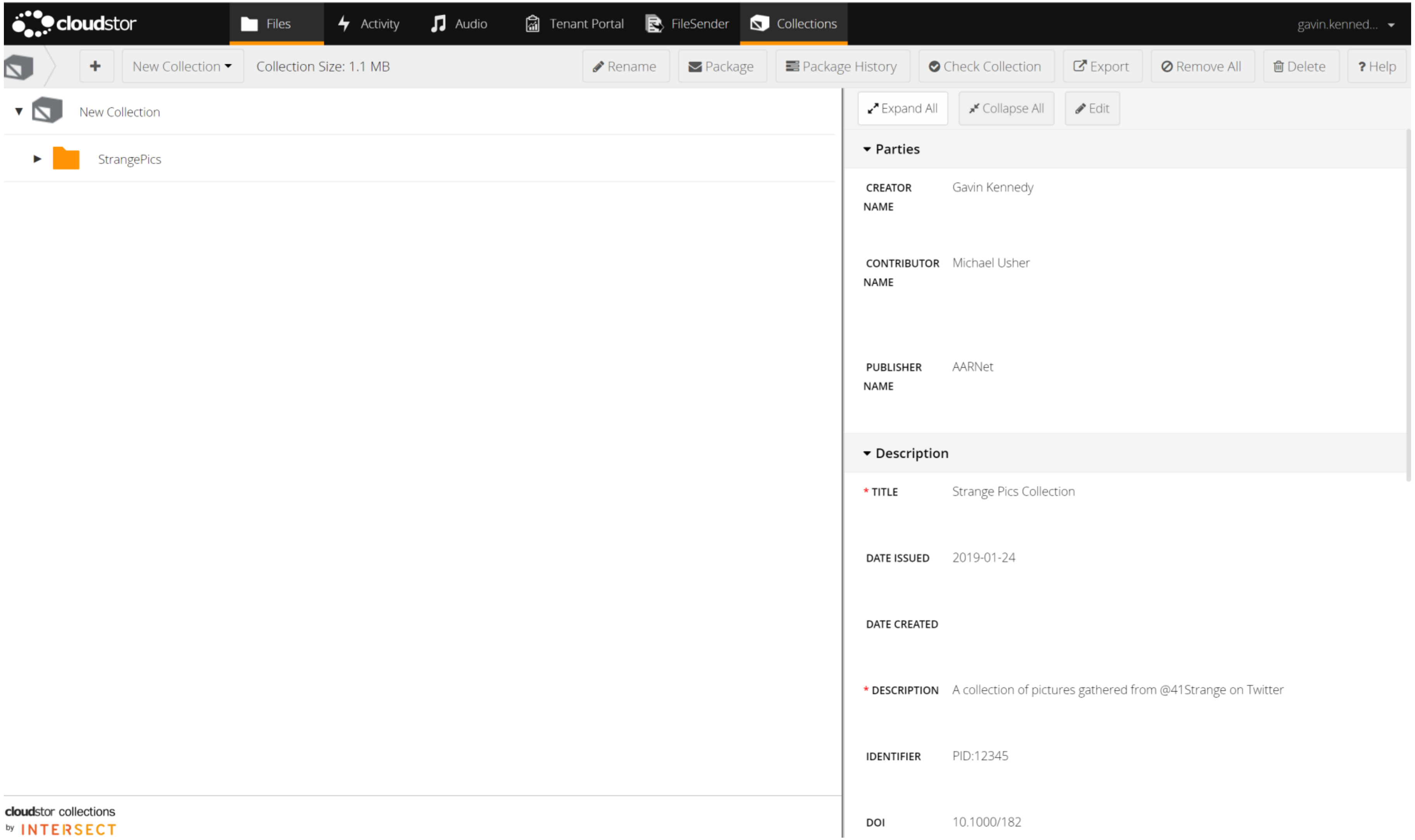
```
myfirstbag/
|-- data
|   \-- 27613-h
|       \-- images
|           \-- q172.png
|           \-- q172.txt
|-- manifest-md5.txt
|     49afbd86a1ca9f34b677a3f09655eae9 data/27613-h/images/q172.png
|     408ad21d50cef31da4df6d9ed81b01a7 data/27613-h/images/q172.txt
\-- bagit.txt
      BagIt-Version: 0.97
      Tag-File-Character-Encoding: UTF-8
```

Getting Started with BagIt in 2018 https://patchbay.tech/2018/03/14/getting-started-with-bagit-in-2018/

FIRST PASS – CR8IT & COLLECTIONS

- **CR8IT: a collaboration of University of Western Sydney, University of Newcastle and Intersect Australia.**
  - Implements BagIT with XML Metadata Tags
  - Developed as an ownCloud plugin.
  - Deployed into ownCloud by UWS, UoN
- **Collections 1.1 & 1.2: an implementation of CR8IT for CloudStor by Intersect**
  - Value added service to CloudStor
  - Users can package selected data for re-distribution, sharing and publishing outside of CloudStor
  - User creates a data package, adds files and annotates in package
  - Annotation using selected DC and Marc elements

COLLECTIONS 1.2



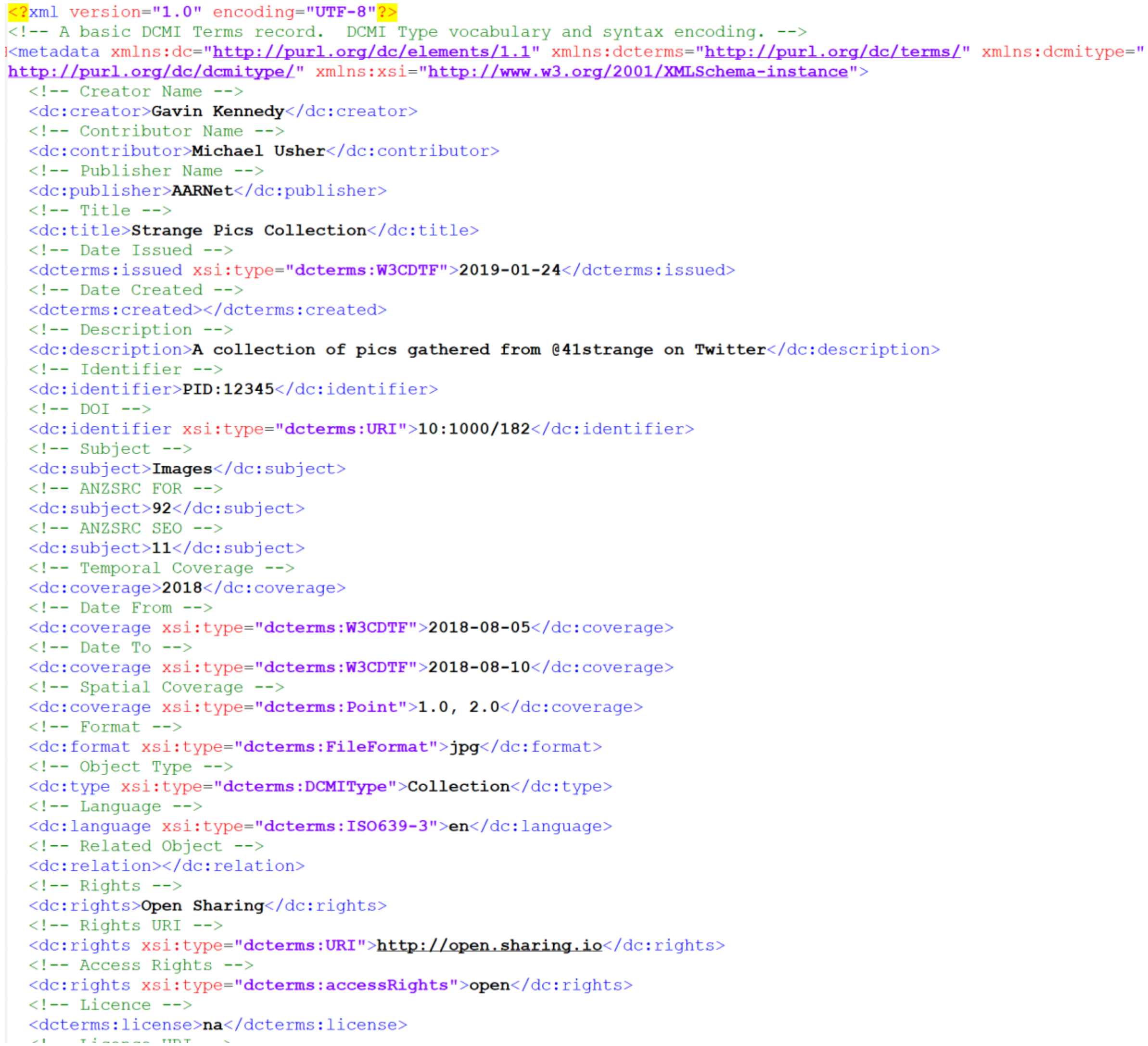
Parties

CREATOR NAME	Gavin Kennedy
CONTRIBUTOR NAME	Michael Usher
PUBLISHER NAME	AARNet

Description

* TITLE	Strange Pics Collection
DATE ISSUED	2019-01-24
DATE CREATED	
* DESCRIPTION	A collection of pictures gathered from @41Strange on Twitter
IDENTIFIER	PID:12345
DOI	10.1000/182

COLLECTIONS 1.2



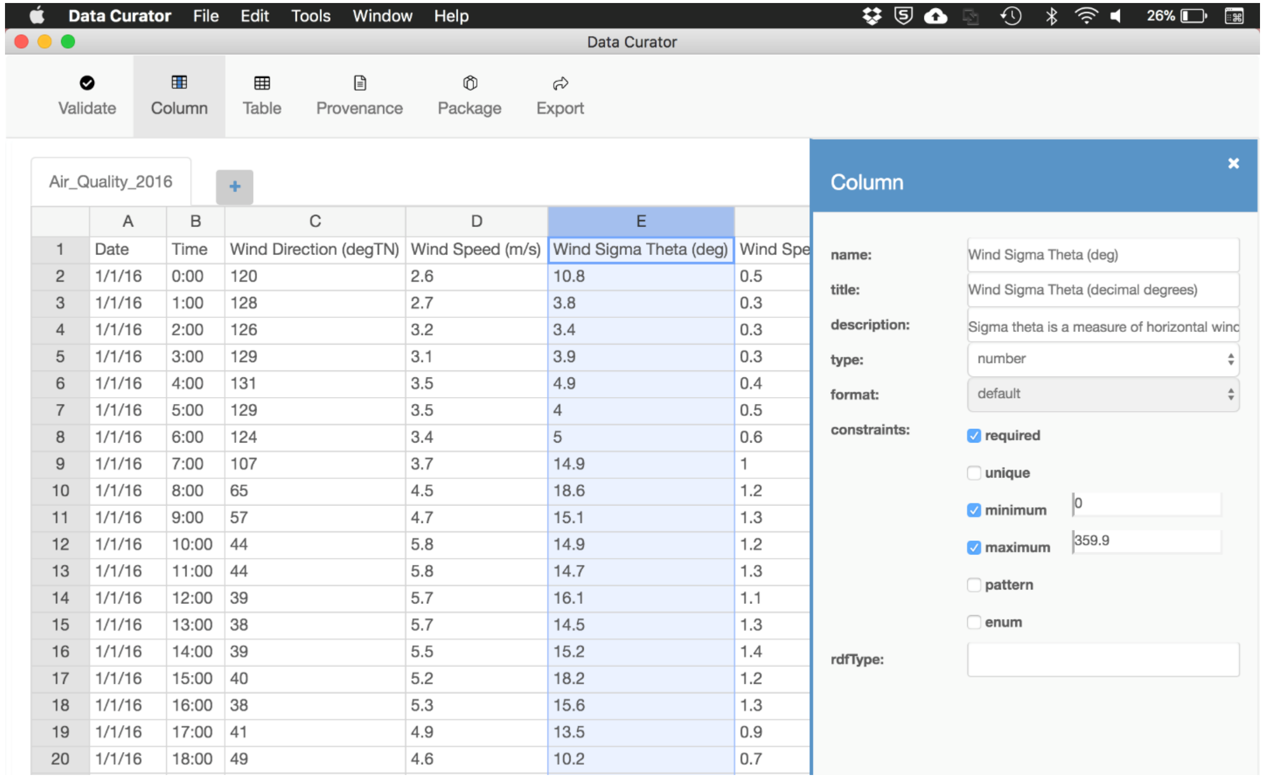
```
<?xml version="1.0" encoding="UTF-8"?>
<!-- A basic DCMI Terms record.  DCMI Type vocabulary and syntax encoding. -->
<metadata xmlns:dc="http://purl.org/dc/elements/1.1" xmlns:dcterms="http://purl.org/dc/terms/" xmlns:dcmitype="
http://purl.org/dc/dcmitype/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  <!-- Creator Name -->
  <dc:creator>Gavin Kennedy</dc:creator>
  <!-- Contributor Name -->
  <dc:contributor>Michael Usher</dc:contributor>
  <!-- Publisher Name -->
  <dc:publisher>AARNet</dc:publisher>
  <!-- Title -->
  <dc:title>Strange Pics Collection</dc:title>
  <!-- Date Issued -->
  <dcterms:issued xsi:type="dcterms:W3CDTF">2019-01-24</dcterms:issued>
  <!-- Date Created -->
  <dcterms:created></dcterms:created>
  <!-- Description -->
  <dc:description>A collection of pics gathered from @41strange on Twitter</dc:description>
  <!-- Identifier -->
  <dc:identifier>PID:12345</dc:identifier>
  <!-- DOI -->
  <dc:identifier xsi:type="dcterms:URI">10:1000/182</dc:identifier>
  <!-- Subject -->
  <dc:subject>Images</dc:subject>
  <!-- ANZSRC FOR -->
  <dc:subject>92</dc:subject>
  <!-- ANZSRC SEO -->
  <dc:subject>11</dc:subject>
  <!-- Temporal Coverage -->
  <dc:coverage>2018</dc:coverage>
  <!-- Date From -->
  <dc:coverage xsi:type="dcterms:W3CDTF">2018-08-05</dc:coverage>
  <!-- Date To -->
  <dc:coverage xsi:type="dcterms:W3CDTF">2018-08-10</dc:coverage>
  <!-- Spatial Coverage -->
  <dc:coverage xsi:type="dcterms:Point">1.0, 2.0</dc:coverage>
  <!-- Format -->
  <dc:format xsi:type="dcterms:FileFormat">jpg</dc:format>
  <!-- Object Type -->
  <dc:type xsi:type="dcterms:DCMIType">Collection</dc:type>
  <!-- Language -->
  <dc:language xsi:type="dcterms:ISO639-3">en</dc:language>
  <!-- Related Object -->
  <dc:relation></dc:relation>
  <!-- Rights -->
  <dc:rights>Open Sharing</dc:rights>
  <!-- Rights URI -->
  <dc:rights xsi:type="dcterms:URI">http://open.sharing.io</dc:rights>
  <!-- Access Rights -->
  <dc:rights xsi:type="dcterms:accessRights">open</dc:rights>
  <!-- Licence -->
  <dcterms:license>na</dcterms:license>
  <!-- Licence URI -->
```

DATA CURATOR – OPEN DATA PACKAGING

Data Curator

share usable open data

- Project between QCIF, ODI and Queensland Government
- Address problem of departments submitting bad data to the Qld Govt Open Data team.
- Electron desktop app for editing, describing and packaging open data CSV files.
- Allows you to annotate the data as well as the dataset (columns, etc).
- Uses the Frictionless table schema
- Validates data against the schema
- Data package can be exported to CKAN
- Opens packages downloaded from CKAN
- https://github.com/ODIQueensland/data-curator



	A	B	C	D	E	
1	Date	Time	Wind Direction (degTN)	Wind Speed (m/s)	Wind Sigma Theta (deg)	Wind Spe
2	1/1/16	0:00	120	2.6	10.8	0.5
3	1/1/16	1:00	128	2.7	3.8	0.3
4	1/1/16	2:00	126	3.2	3.4	0.3
5	1/1/16	3:00	129	3.1	3.9	0.3
6	1/1/16	4:00	131	3.5	4.9	0.4
7	1/1/16	5:00	129	3.5	4	0.5
8	1/1/16	6:00	124	3.4	5	0.6
9	1/1/16	7:00	107	3.7	14.9	1
10	1/1/16	8:00	65	4.5	18.6	1.2
11	1/1/16	9:00	57	4.7	15.1	1.3
12	1/1/16	10:00	44	5.8	14.9	1.2
13	1/1/16	11:00	44	5.8	14.7	1.3
14	1/1/16	12:00	39	5.7	16.1	1.1
15	1/1/16	13:00	38	5.7	14.5	1.3
16	1/1/16	14:00	39	5.5	15.2	1.4
17	1/1/16	15:00	40	5.2	18.2	1.2
18	1/1/16	16:00	38	5.3	15.6	1.3
19	1/1/16	17:00	41	4.9	13.5	0.9
20	1/1/16	18:00	49	4.6	10.2	0.7

Column ×

name: Wind Sigma Theta (deg)

title: Wind Sigma Theta (decimal degrees)

description: Sigma theta is a measure of horizontal winc

type: number

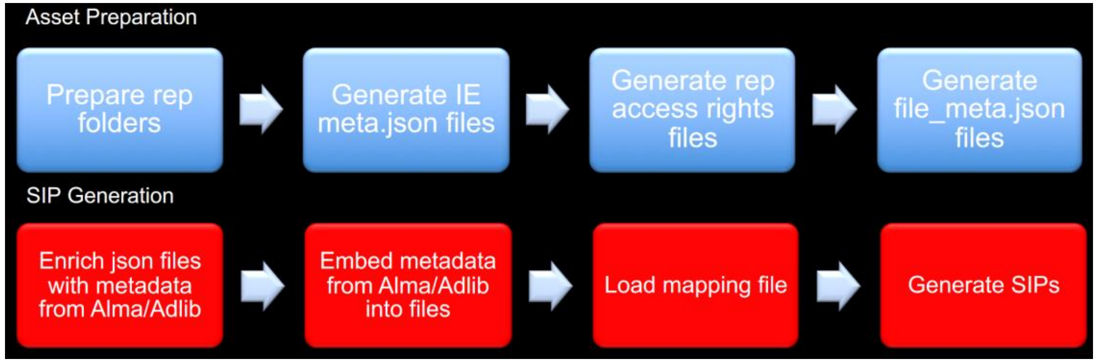
format: default

constraints:
☑ required
☐ unique
☑ minimum 0
☑ maximum 359.9
☐ pattern
☐ enum

rdfType:

PANDA – PRESERVATION AND DIGITAL ACCESS

- Project at State Library of New South Wales (Matt Burgess)
- Addresses problem of packaging archivable data, from multiple sources, prior to ingest.



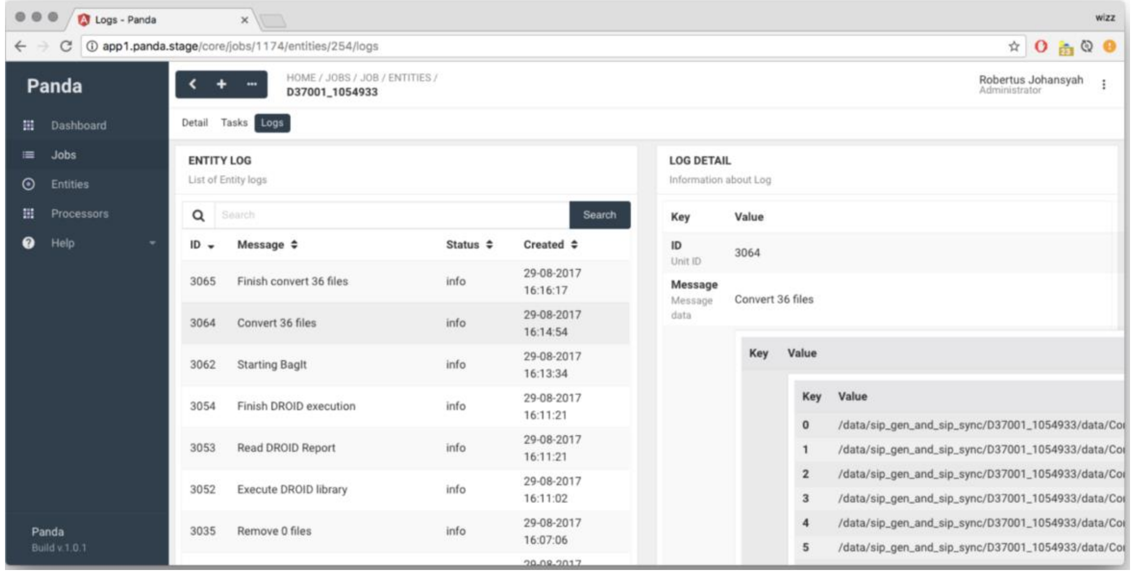
Asset Preparation

Prepare rep folders → Generate IE meta.json files → Generate rep access rights files → Generate file_meta.json files

SIP Generation

Enrich json files with metadata from Alma/Adlib → Embed metadata from Alma/Adlib into files → Load mapping file → Generate SIPs

- Ingest packages (Submission Information Packages) for multiple Digital Archival and Preservation platforms:
  - Rosetta
  - Archivematica
  - Preservica



ENTITY LOG

List of Entity logs

ID	Message	Status	Created
3065	Finish convert 36 files	info	29-08-2017 16:16:17
3064	Convert 36 files	info	29-08-2017 16:14:54
3062	Starting BagIt	info	29-08-2017 16:13:34
3054	Finish DROID execution	info	29-08-2017 16:11:21
3053	Read DROID Report	info	29-08-2017 16:11:21
3052	Execute DROID library	info	29-08-2017 16:11:02
3035	Remove 0 files	info	29-08-2017 16:07:06

LOG DETAIL

Information about Log

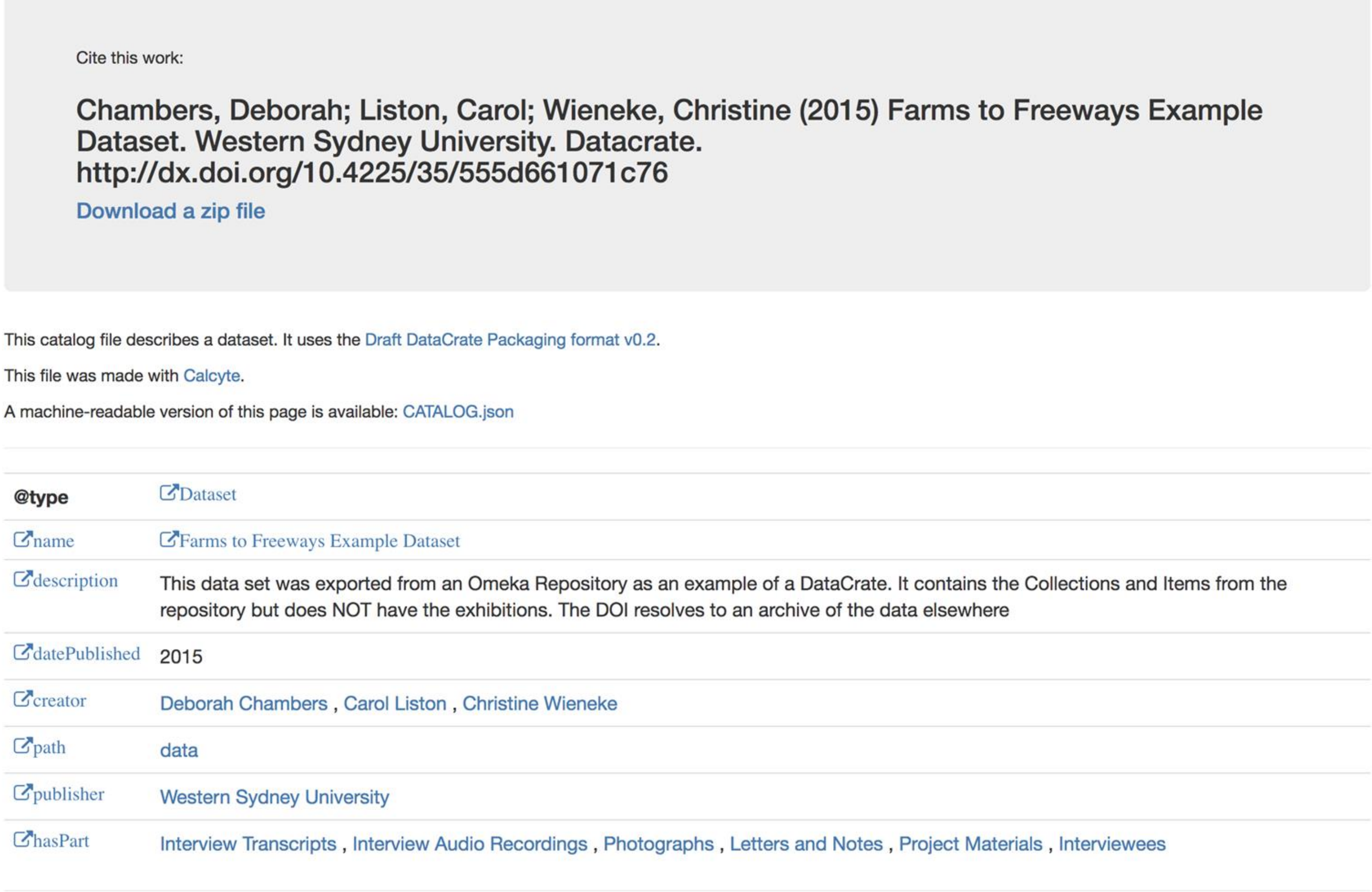
Key	Value
ID Unit ID	3064
Message Message data	Convert 36 files

Key	Value
0	/data/sip_gen_and_sip_sync/D37001_1054933/data/Co
1	/data/sip_gen_and_sip_sync/D37001_1054933/data/Co
2	/data/sip_gen_and_sip_sync/D37001_1054933/data/Co
3	/data/sip_gen_and_sip_sync/D37001_1054933/data/Co
4	/data/sip_gen_and_sip_sync/D37001_1054933/data/Co
5	/data/sip_gen_and_sip_sync/D37001_1054933/data/Co

DATACRATE

- **Developed by University of Technology Sydney**
- **Creates a machine actionable and human readable data catalog**
- **Internal file structure based on OCFL spec: ocfl.io**
- **Schema.org based metadata elements**
  - Standard semantic web metadata as used by search engines
- **Common linked data storage and exchange format (JSON-LD)**
  - Linking elements within the package.
  - Like *FileB is a transalation of FileA.*
- **Ability to link external entitites (parties, grants, publications, DOIs)**
- **Renditions to transform package into a self-contained website.**
- **Still Bagit so contains checksums for internal consistency.**

DATA CRATE



Cite this work:

Chambers, Deborah; Liston, Carol; Wieneke, Christine (2015) Farms to Freeways Example Dataset. Western Sydney University. Datacrate. http://dx.doi.org/10.4225/35/555d661071c76

Download a zip file

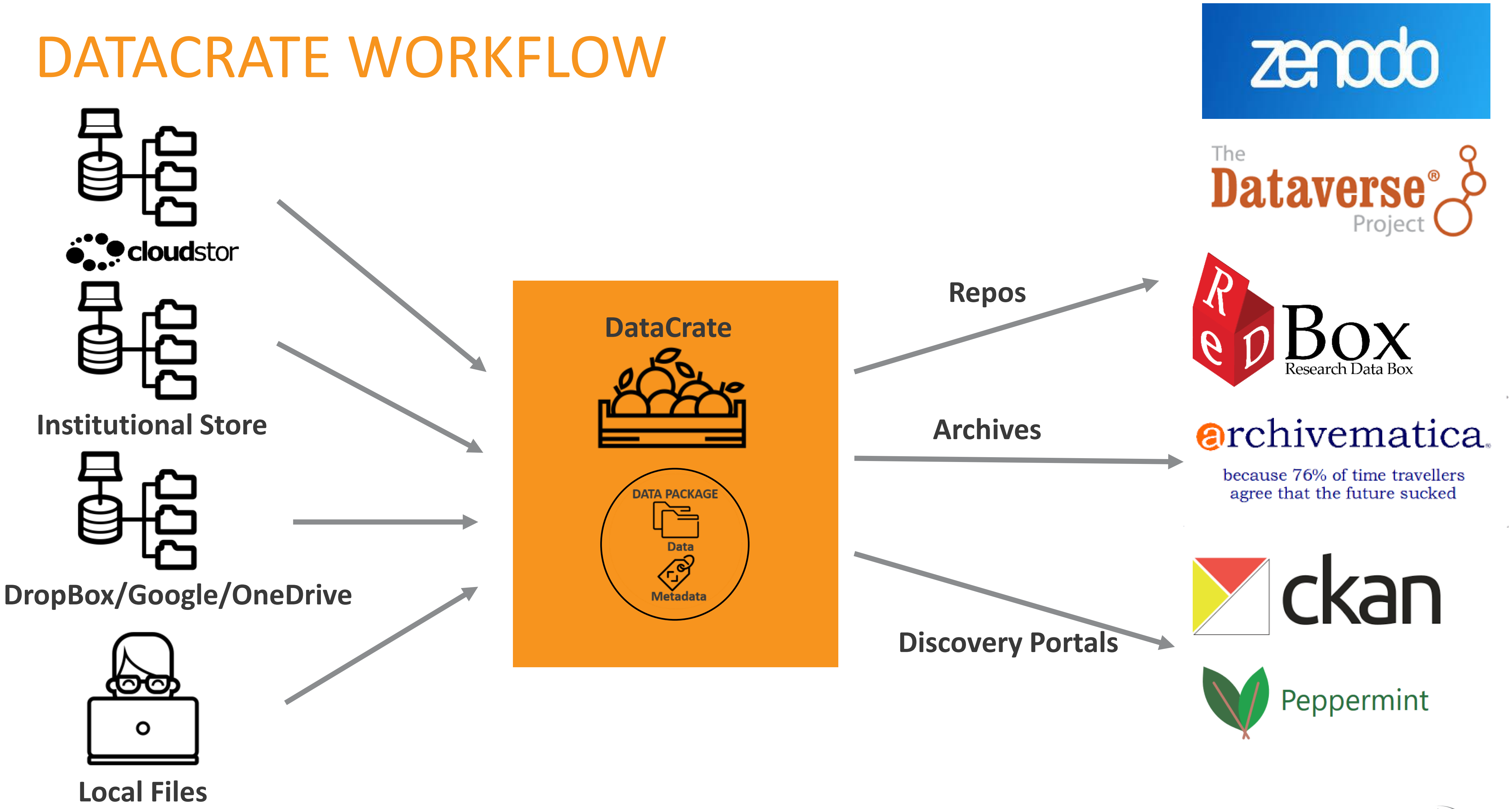
This catalog file describes a dataset. It uses the Draft DataCrate Packaging format v0.2.

This file was made with Calcyte.

A machine-readable version of this page is available: CATALOG.json

@type	Dataset
name	Farms to Freeways Example Dataset
description	This data set was exported from an Omeka Repository as an example of a DataCrate. It contains the Collections and Items from the repository but does NOT have the exhibitions. The DOI resolves to an archive of the data elsewhere
datePublished	2015
creator	Deborah Chambers , Carol Liston , Christine Wieneke
path	data
publisher	Western Sydney University
hasPart	Interview Transcripts , Interview Audio Recordings , Photographs , Letters and Notes , Project Materials , Interviewees

DATACRATE WORKFLOW



cloudstor

Institutional Store

DropBox/Google/OneDrive

Local Files

DataCrate

DATA PACKAGE

Data

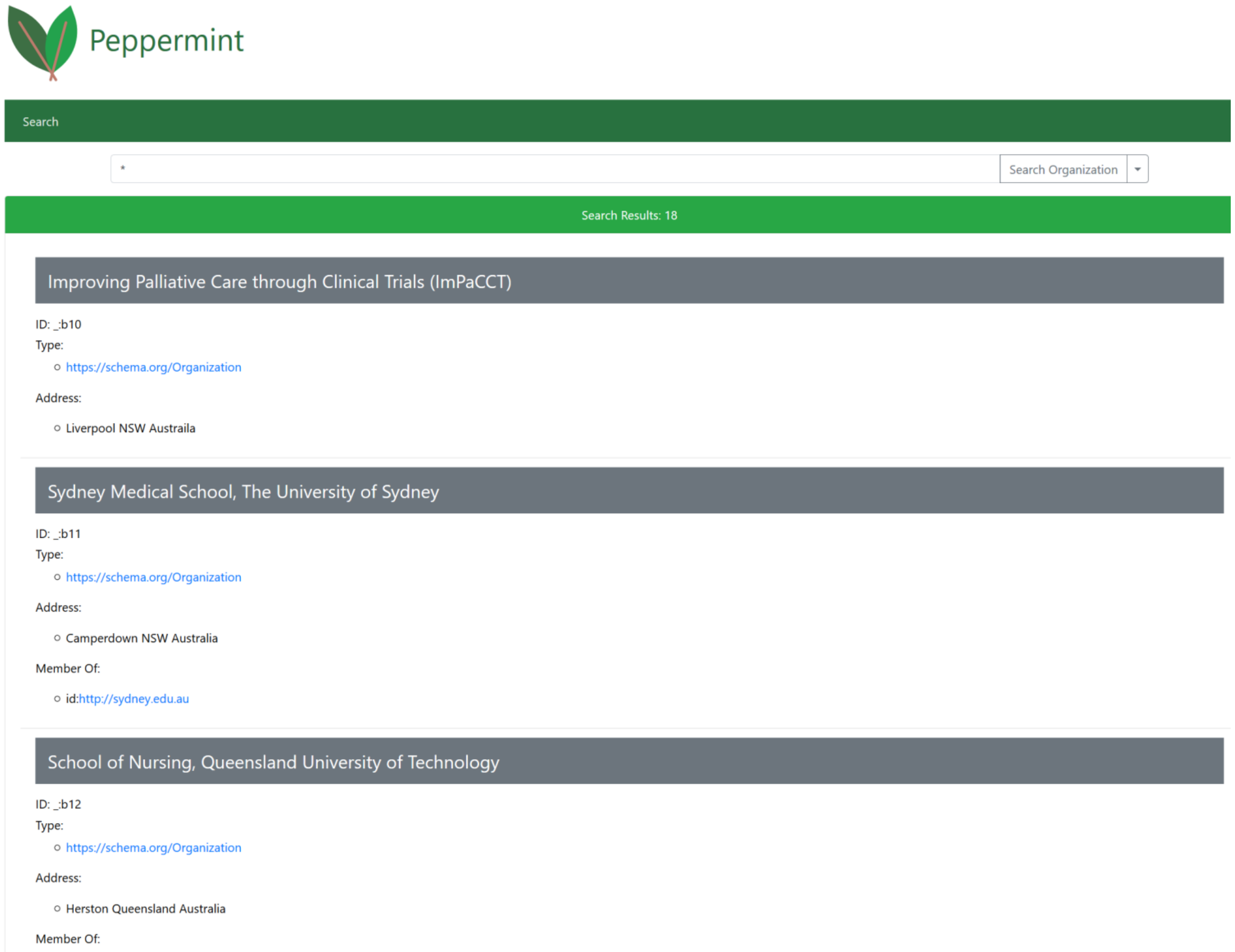
Metadata

Repos

Archives

Discovery Portals

DISCOVERY PORTAL



Search Results: 18

Improving Palliative Care through Clinical Trials (ImPaCCT)

ID: _:b10

Type:

- https://schema.org/Organization

Address:

- Liverpool NSW Austraila

Sydney Medical School, The University of Sydney

ID: _:b11

Type:

- https://schema.org/Organization

Address:

- Camperdown NSW Australia

Member Of:

- id:http://sydney.edu.au

School of Nursing, Queensland University of Technology

ID: _:b12

Type:

- https://schema.org/Organization

Address:

- Herston Queensland Australia

Member Of:

ACKNOWLEDGEMENTS

Co-authors really, and super intelligent thought leaders:

- **Peter Sefton, Director – eResearch Services, University of Technology Sydney**
- **Ingrid Mason, Senior eResearch Analyst, AARNet**

“Make Data Crate Again” – Liz Stokes, UTS Data Librarian, 2017