DATA PACKAGING

Possible to do, but how and what for?
CAIRSS 28 October 2019

Ingrid Mason, Deployment Strategist, eResearch @AARNet

Data packaging

Technology development

Packaging in lifecycles

A DATA PACKAGE IS…

“A data package is simply the combination of a dataset with metadata that describes the dataset.

The purpose of the data package is to provide sufficient contextual data to make the dataset usable to others. It therefore becomes the basis of a loosely coupled data management platform in which the information is in the dataset, not in the platform. A receiving platform just needs to be able to access and interpret the package.”

A side serve of metadata: evolving AARNet’s tools for data packaging

Gavin Kennedy & Ingrid Mason (AARNet); Peter Sefton (UTS).

CS3 2019 – Cloud Storage Synchronization and Sharing Services, 28-30 January 2019

WHAT DOES A DATA PACKAGE ENABLE?

Facilitates depositing of research data to repositories.

Facilitates data sharing (FAIR principles).

Facilitates the repeatability or reproducibility of research.

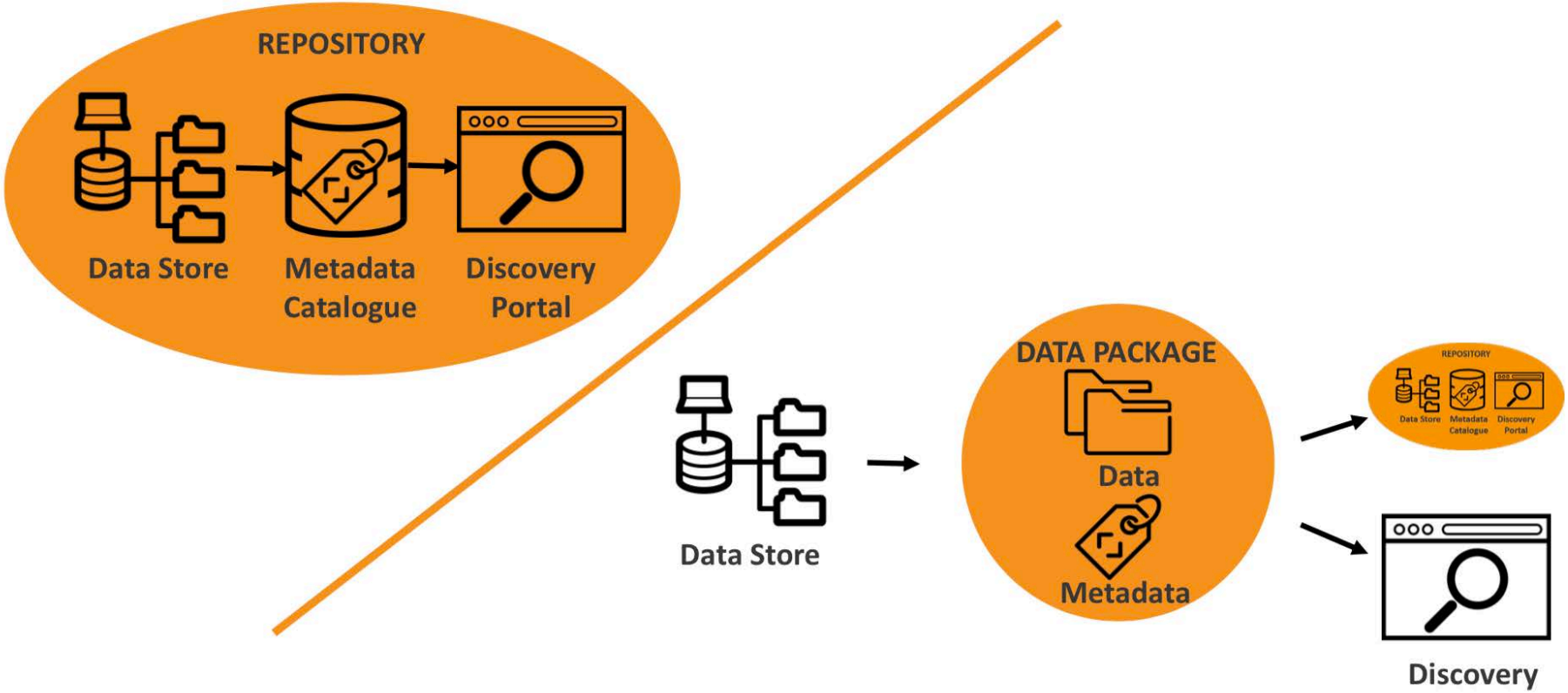
- Data packages typically contain
  - Digital material (files like data, software etc)
  - Metadata (information like description, ownership etc)
  - Renditions – metadata outputs in a preferred format
  - Formats including:
    - HTML, XML, JSON-LD
    - DC, MARC, Schema.org

TECH DEVELOPMENT

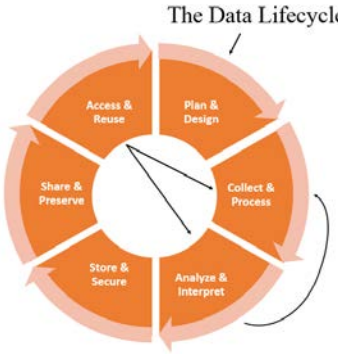
Data packaging technologies and their evolution (here in Australia with global connections)

- First there was BagIt (file manifest spec)
- *cr8it* was developed using BagIt (WSU software) & *PanDA* was developed using BagIt (SLNSW software)
- *cr8it* was redeveloped by Intersect, AARNet and WSU (as an ownCloud plugin) to become *CloudStor Collections*
- *DataCrate* developed (UTS) (based on cr8it)
- *RO-Crate* was built based on *DataCrate* (Research Object and OCFL specs)
- Research Object spec > community effort http://www.researchobject.org/specifications/
- Oxford Common File Layout spec > community effort https://ocfl.io/

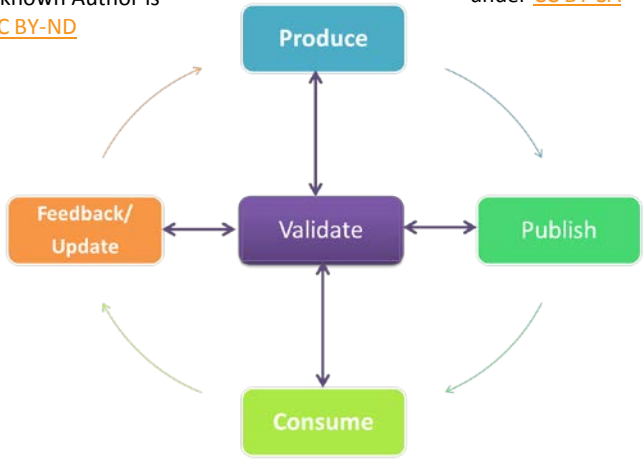
REPO VS DATA PACKAGE



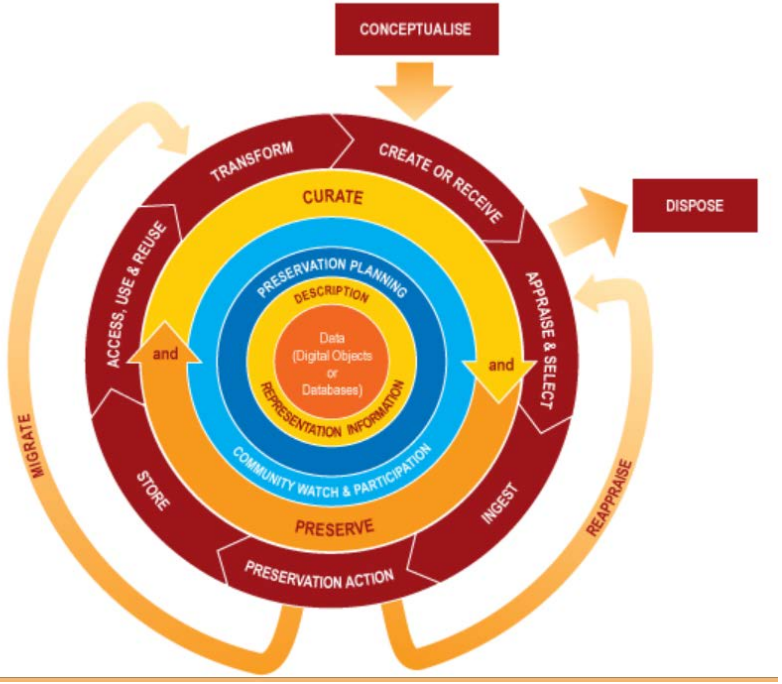
This Photo by Unknown Author is licensed under CC BY-ND



This Photo by Unknown Author is licensed under CC BY-SA



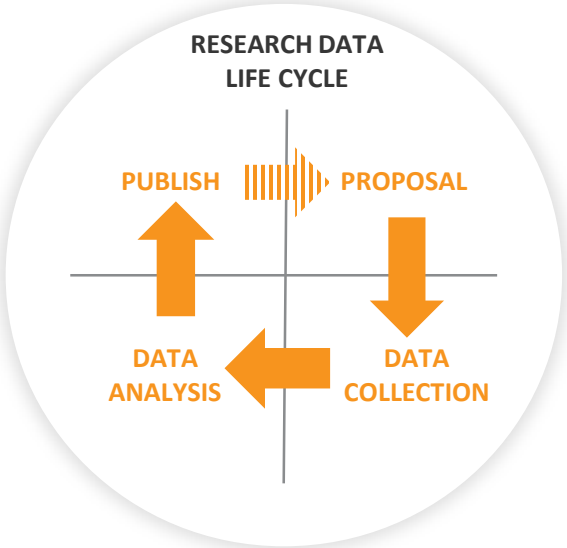
This Photo by Unknown Author is licensed under CC BY-SA



This Photo by Unknown Author is licensed under CC BY

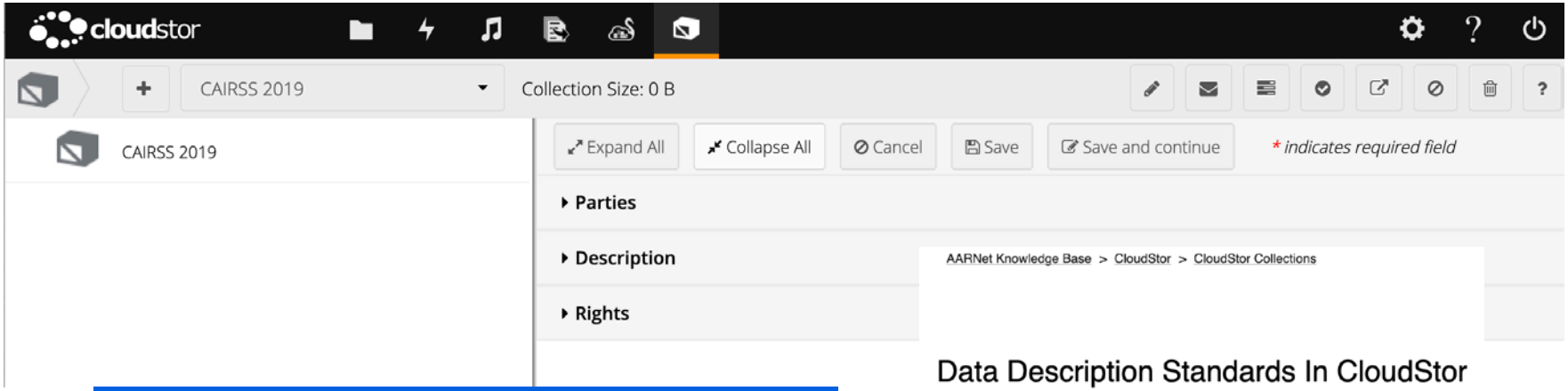
DATA PACKAGING IN RESEARCH DATA LIFECYCLE

This Photo by Unknown Author is licensed under CC BY-SA



This Photo by Unknown Author is licensed under CC BY-SA-NC

CLOUDSTOR COLLECTIONS



CAIRSS 2019	Folder
manifest-md5.txt	Plain Text Document
data	Folder
CollectionData	Folder
CAIRSS 2019.xml	XML
README.html	HTML document
2019-10-27T10-47-0...5db575adda2e6.guid	Document
bagit.txt	Plain Text Document
CAIRSS 2019.zip	ZIP archive

AARNet Knowledge Base > CloudStor > CloudStor Collections

Data Description Standards In CloudStor Collections

Two new metadata models have been created for CloudStor Collections using existing standards heavily used in institutional repositories i.e. unqualified Dublin Core elements and DCMI terms.

The metadata models are visible in the drop down menu as: DCES v1.0 - org.dublincore and DC+DCMI Terms v1.0 - org.dublincore

Why AARNet?

- Packaging data will be a component in research and education workflows
- National and global partnerships in common approach to packaging research data

What's in it for CAIRSS?

Evolving tools and specs for data packaging that are interoperable with institutional repositories and digital archive systems.

Cloud storage services can be enhanced to perform as both the conduit for data packages and the repository and the archive.

Watch this space.

THANK YOU