

Data Retention Project

One Year on

PRESENTED BY

Max WILKINSON
ARDC Data Infrastructure Architect

eResearch NZ 2022



Australian Research Data Commons

Purpose

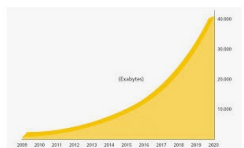
To provide Australian researchers with competitive advantage through data.

Mission

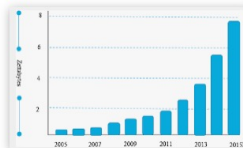
To accelerate research and innovation by driving excellence in the creation, analysis and retention of high-quality data assets.

The Challenge (part 1)

Significant Data Growth



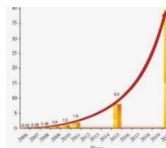
The exponential data growth estimated ...
researchgate.net



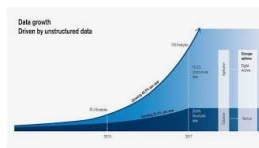
rapid growth rate of data in Zettabytes ...
researchgate.net



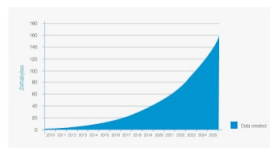
Data growth and expansion (IDC, 2009) ...
researchgate.net



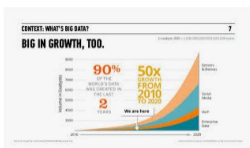
Global growth trend of data volume ...
researchgate.net



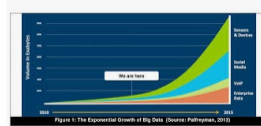
Everything a Data Scientist Should Know ...
kdnuggets.com



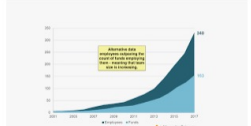
Forecast of exponential growth of ...
reddit.com



Rise of the Data Warehouse | Avora
avora.com



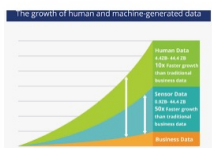
Data Analytics: Concepts, Technologies ...
semanticscholar.org



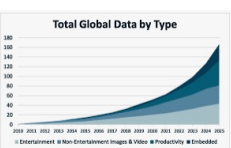
Buy-side Alternative Data Employee ...
alternativedata.org



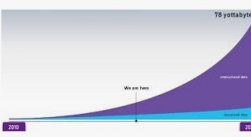
Industry Verticals Tackle Unstructured Data
kevinjackson.blogspot.com



IoT, Big Data and AI - the New ...
business2community.com



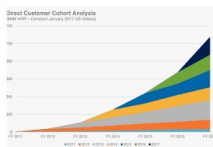
The Data Deluge - Drowning in Data ...
uncommonlogic.com



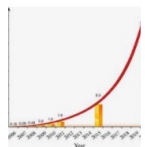
Introduction to BIG DATA: What is ...
bigdatapath.wordpress.com



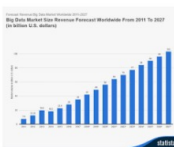
big data growth - Google 搜尋 | Big ...
pinterest.com



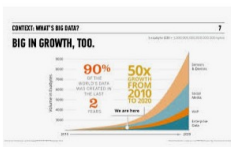
MongoDB: Riding The Data Wave (NASDAQ) ...
seekingalpha.com



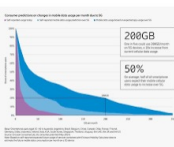
Global growth trend of data ...
researchgate.net



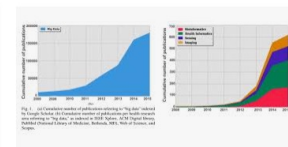
10 Charts That Will Change Your ...
forbes.com



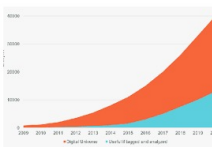
Ensure Business Growth via Big Data ...
promptcloud.com



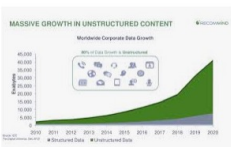
How Much Will 5G Data Usage Increase ...
spectrummattersindeed.blogspot.com



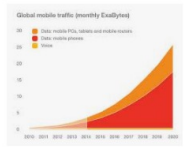
Healthcare Big Data Analytics
healthanalytics.com



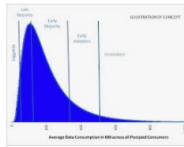
Data growth between 2009 and 2020 ...
researchgate.net



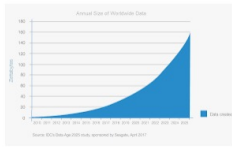
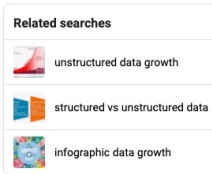
The Massive Growth in Unstructured Data ...
researchgate.net



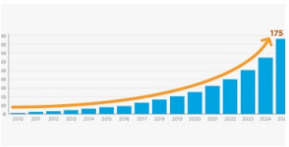
Global Mobile Data Traffic 2010-2020 ...
whatsthebigdata.com



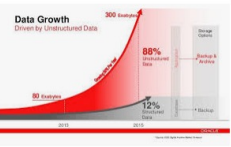
Mobile Data Growth ... The Perfect Storm ...
techeconomyblog.com



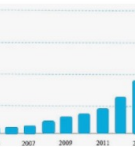
Orchestrating Enterprise Data with Data ...
virtustream.com



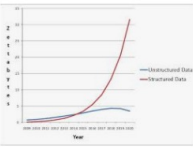
Big data overview | AP CSP (article) ...
khanacademy.org



data growth driven by unstructured ...
pinterest.com



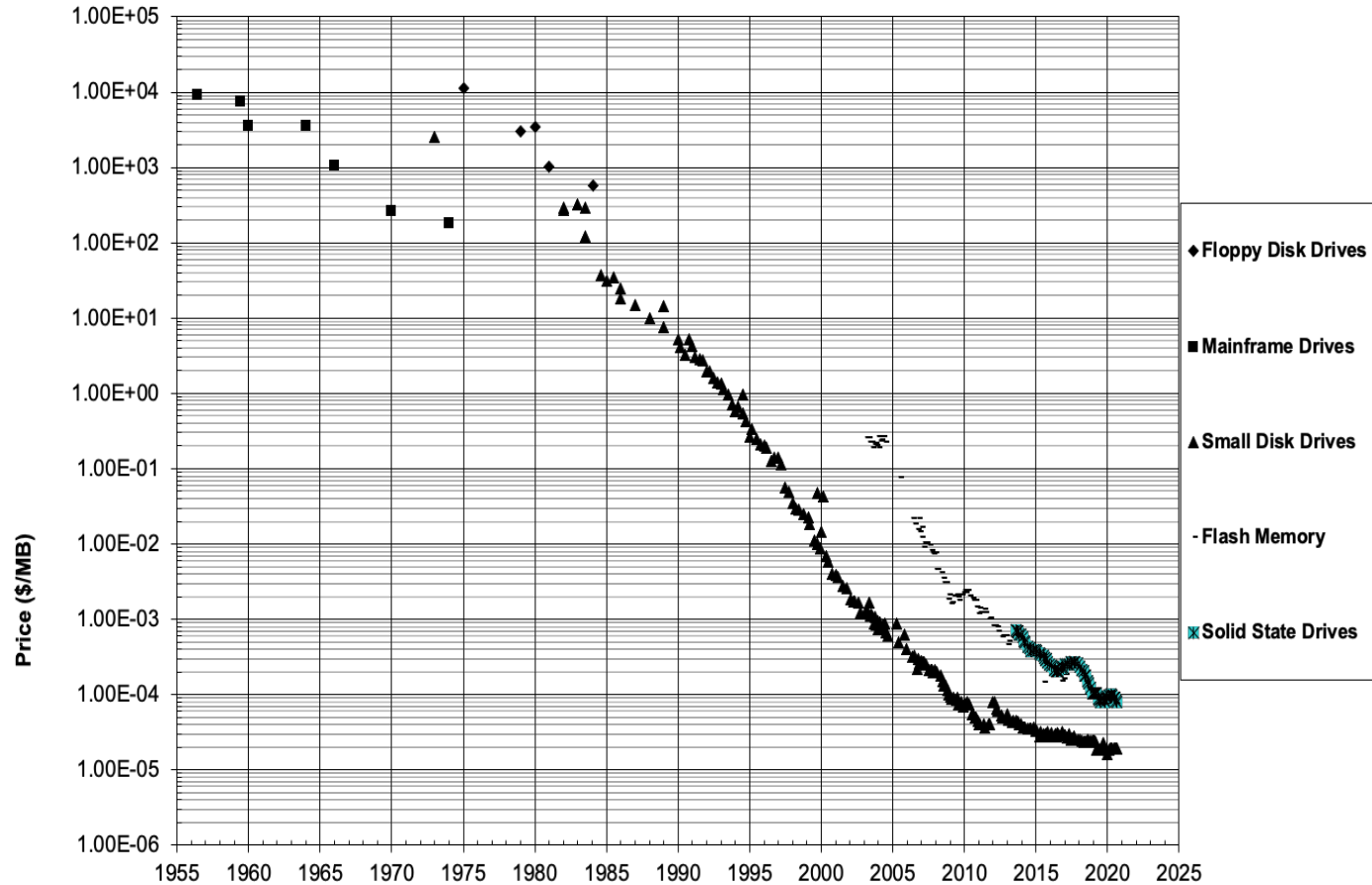
rapid growth rate of data in Z ...
researchgate.net



structured vs unstructured data growth ...
tomkendig.wordpress.com

The Challenge (part 2)

Flatlining Storage costs

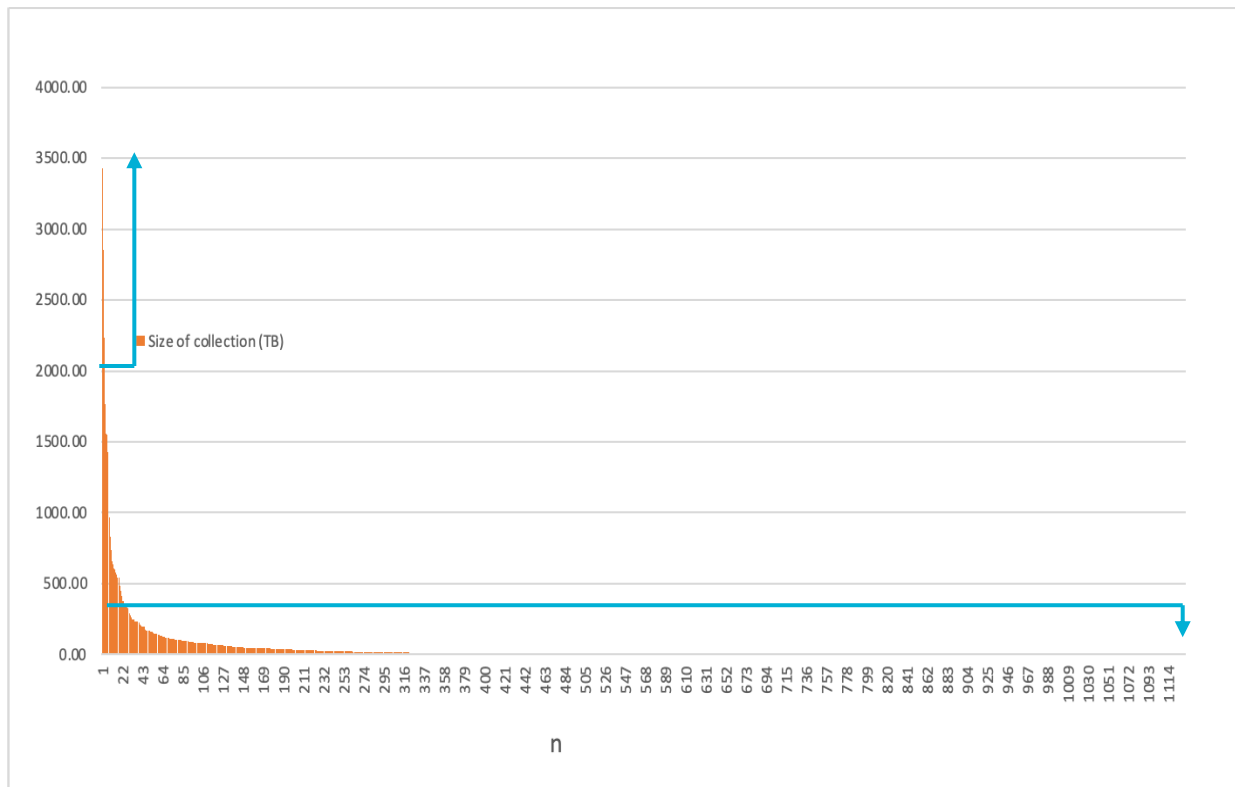


<https://jcmit.net/index.htm>

The Challenge (part 3)

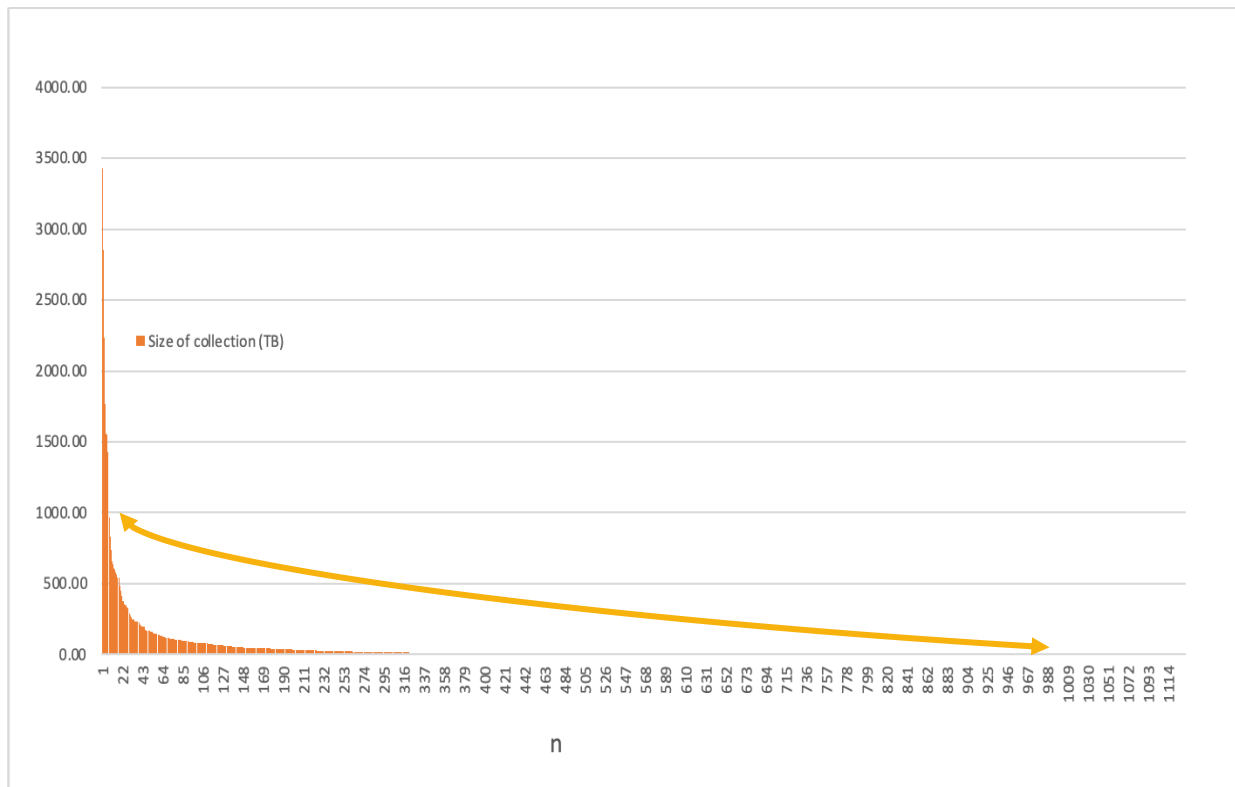
Burden Distribution

Range TB	n	%
0-400	1089	96
>2000	3	0.27



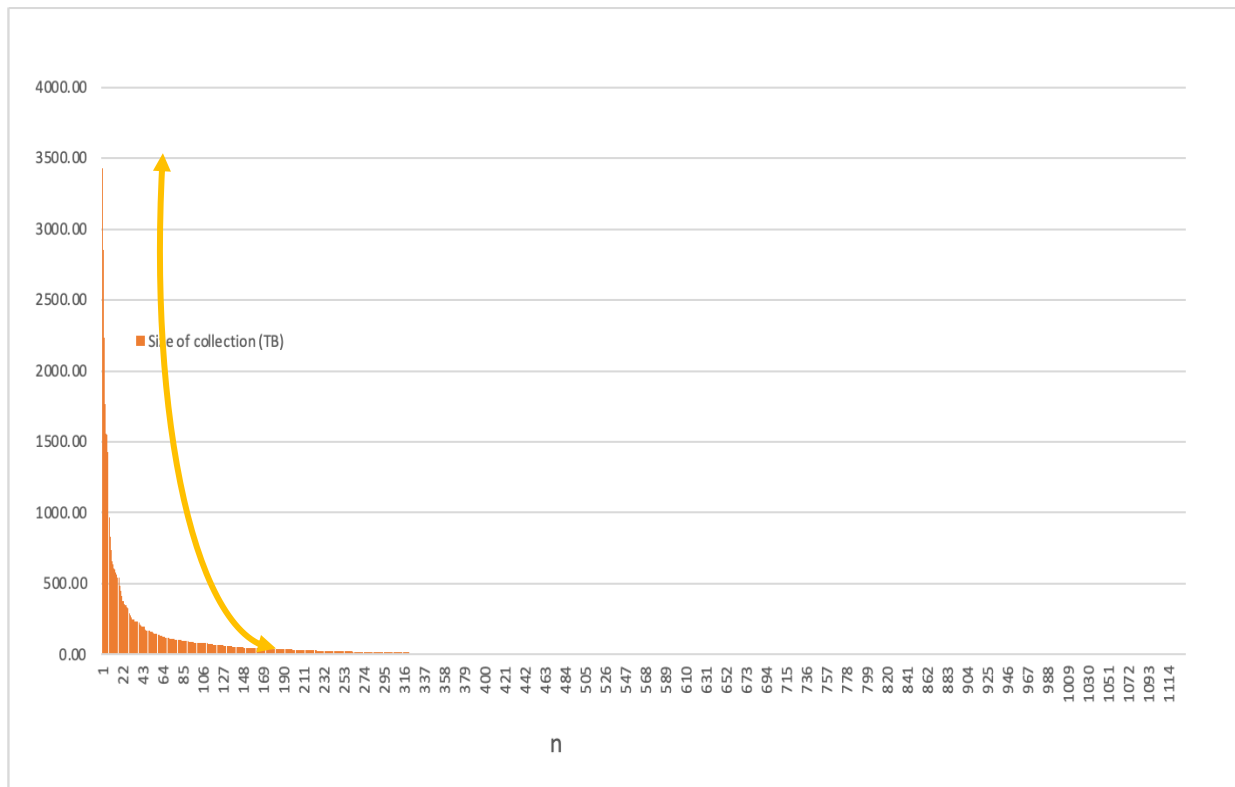
Direction 1 ?

- Increase in small collections



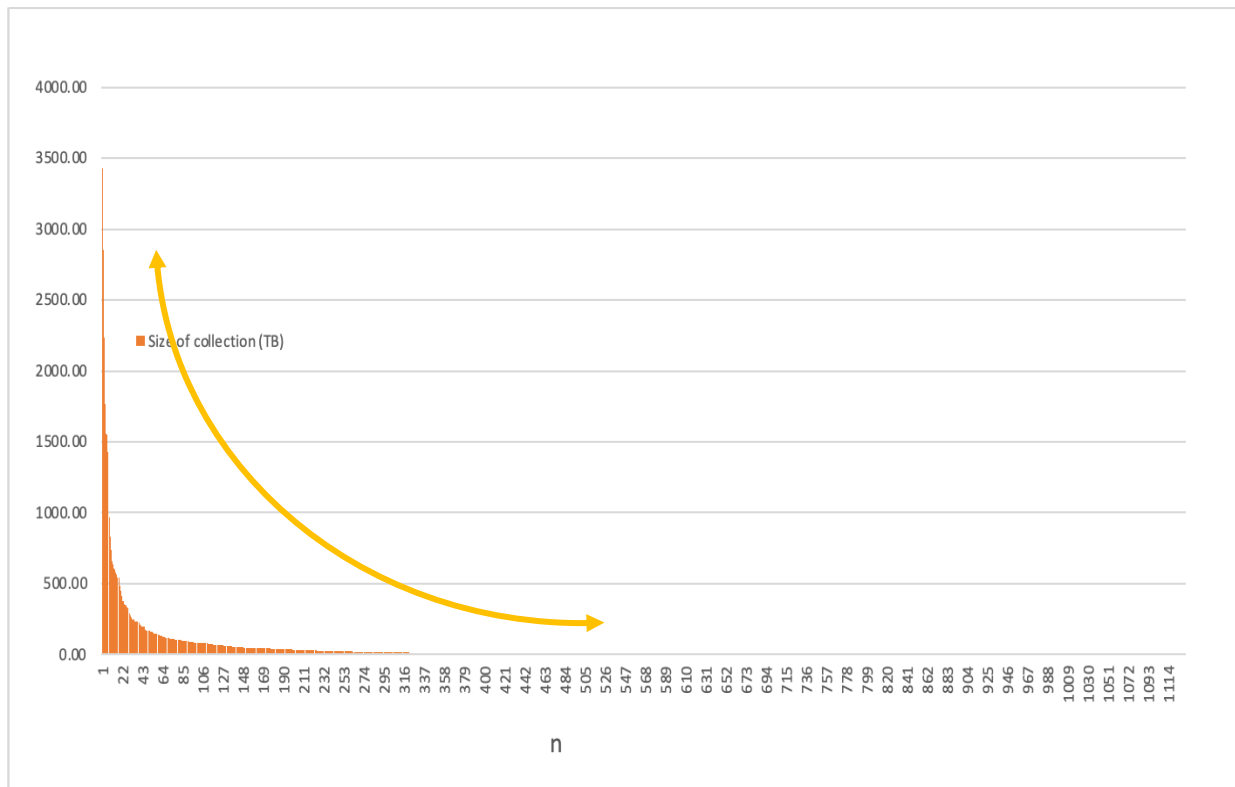
Direction 2 ?

- Increase in large collections



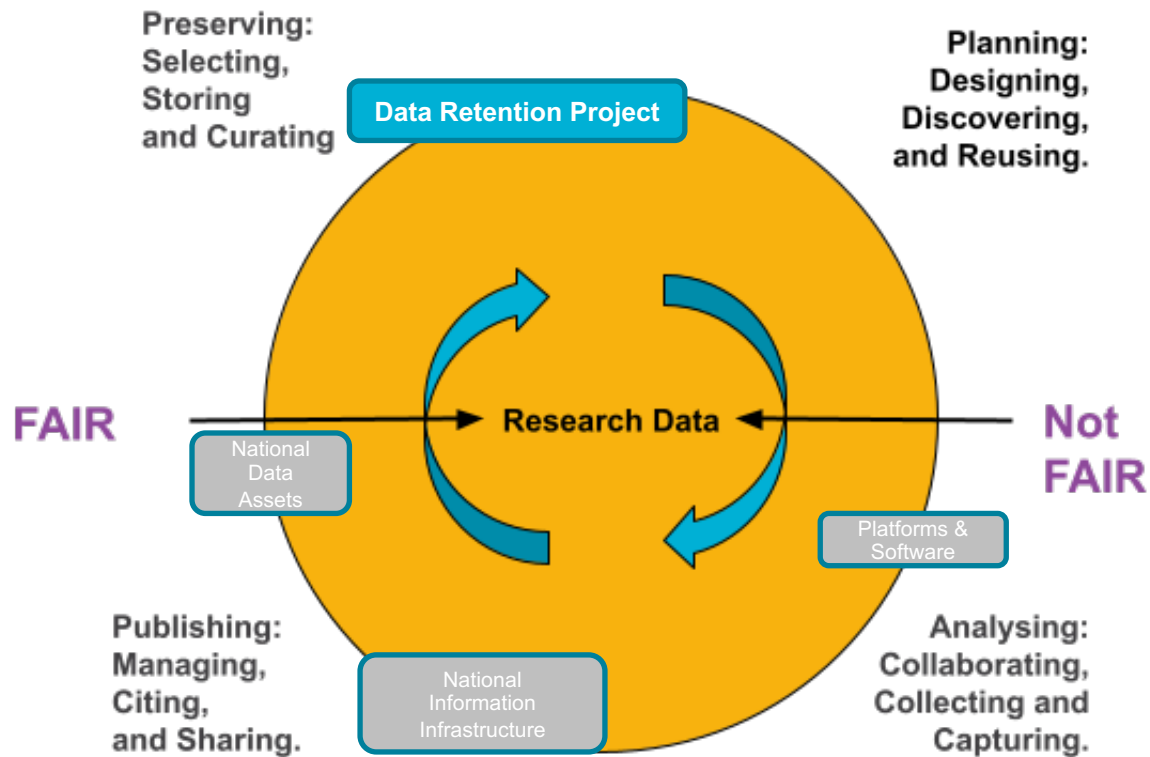
Direction 3 ?

- Increase in all collections



Primary Focus

- Post Project
- Infrastructure
- Enduring change



Data Retention Project Objectives

How to get...

- **Coherent Business Intelligence**
- **Consistent collection characterisation**

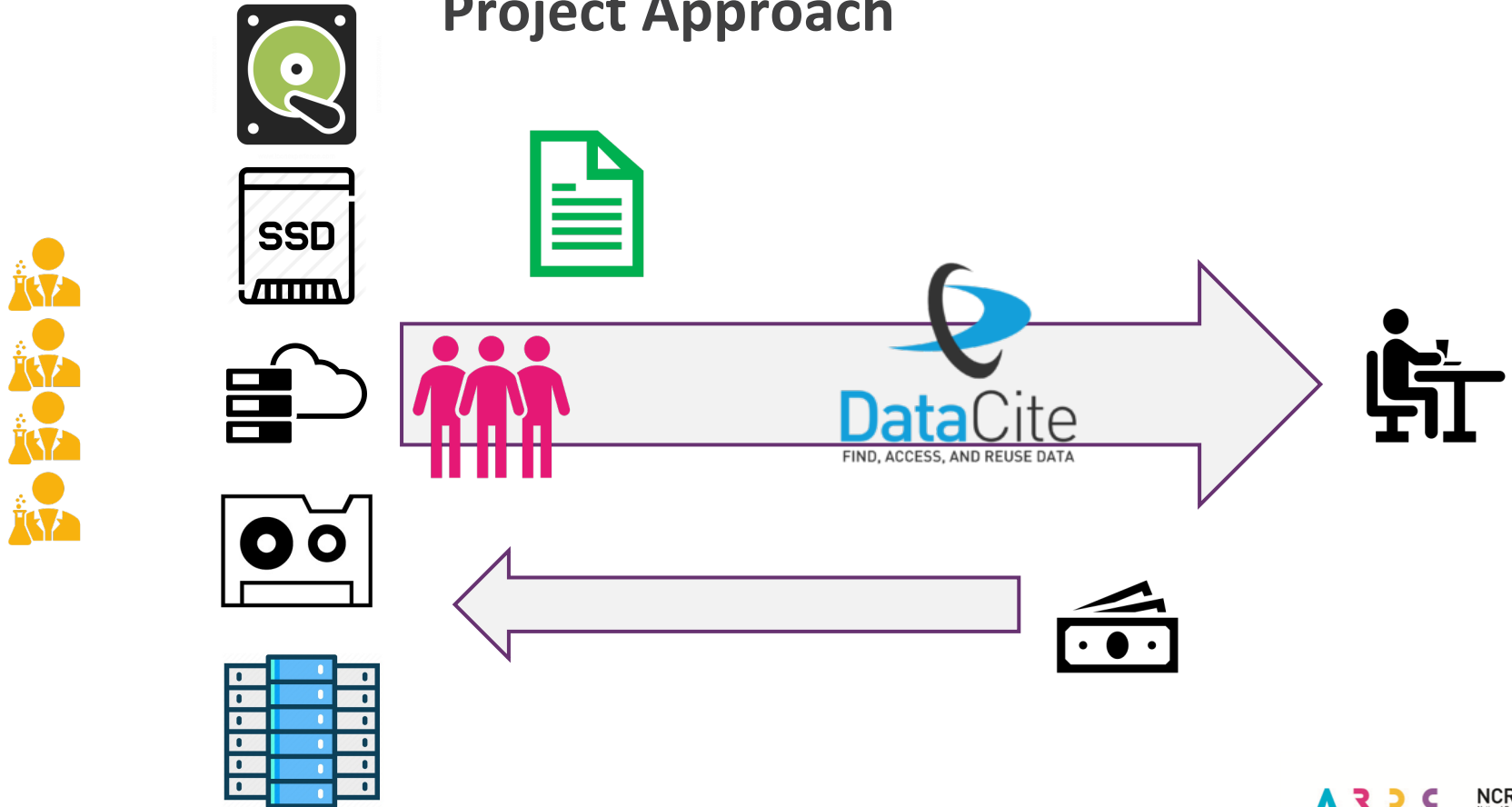
when....

- **Fragmentation remains across**
 - Providers,
 - practice and
 - burden concerns

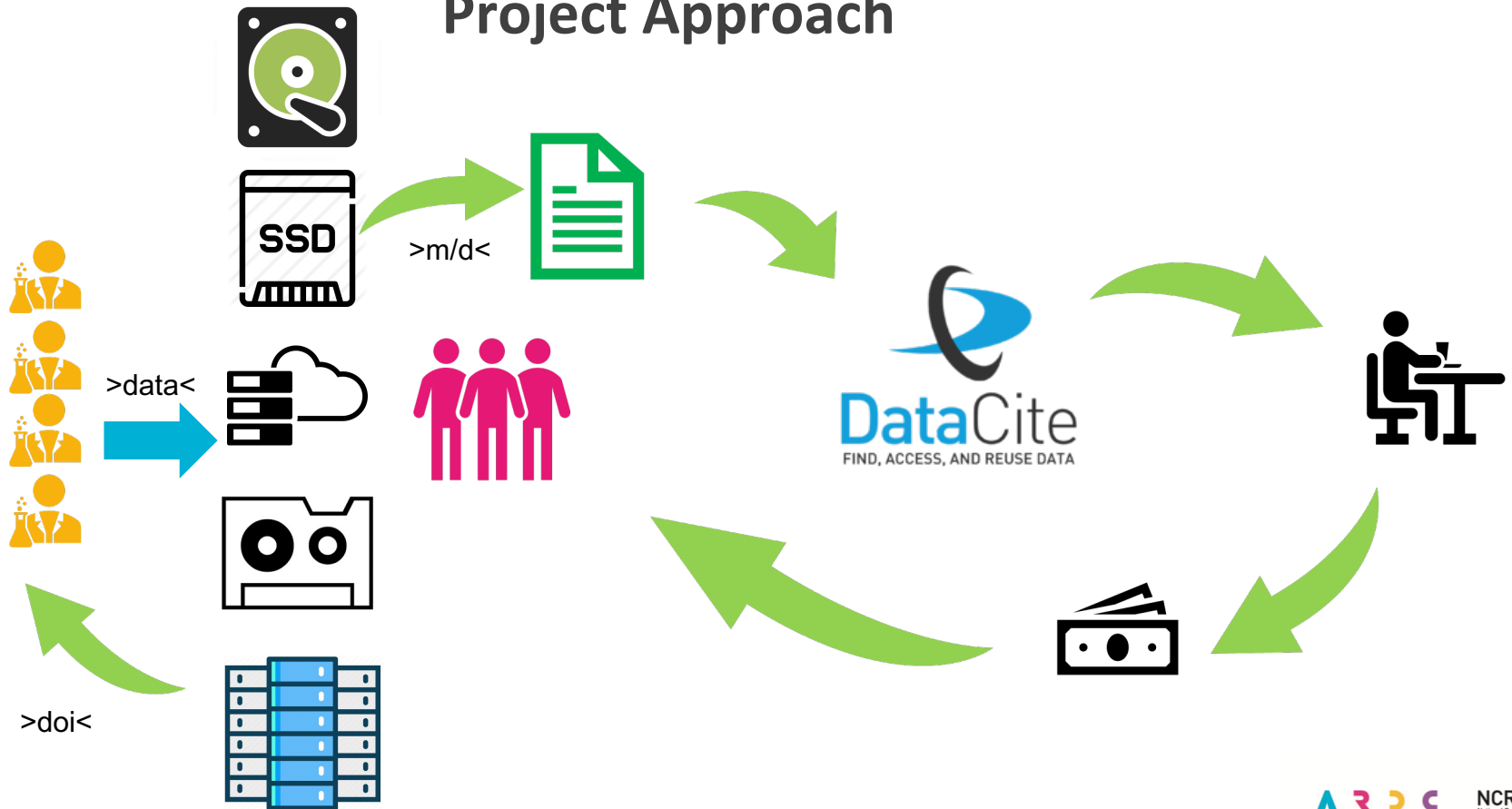
Data Retention Project Strategy

- **Content selection via surrogate value measures**
- **Support focused on delivery**
- **Multi-incentive models**
 - \$
 - International data citation standards
 - Demarcation of responsibility

Project Approach

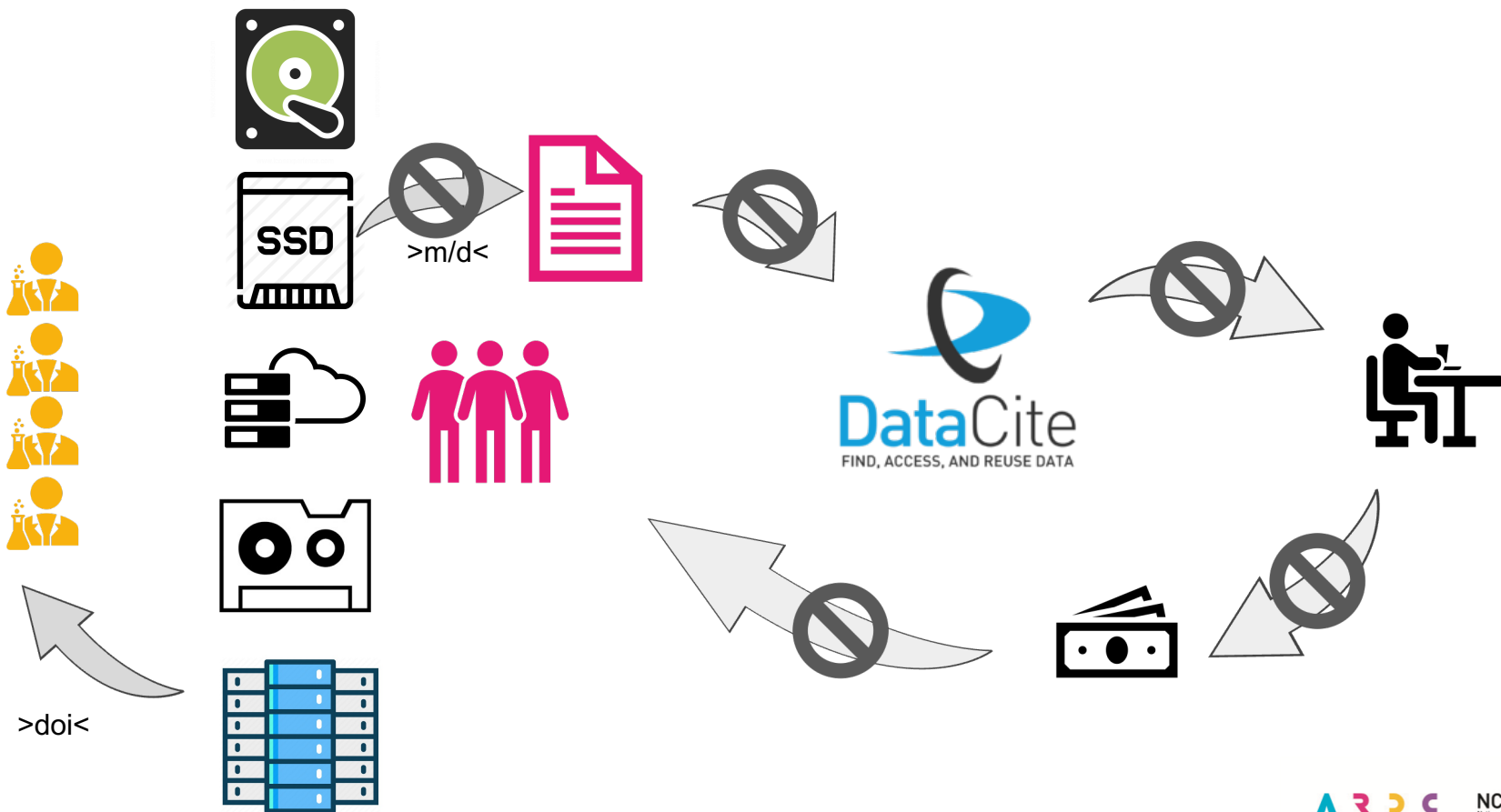


Project Approach



Some Challenges

Issue – Collecting metadata



Issue – Collecting metadata

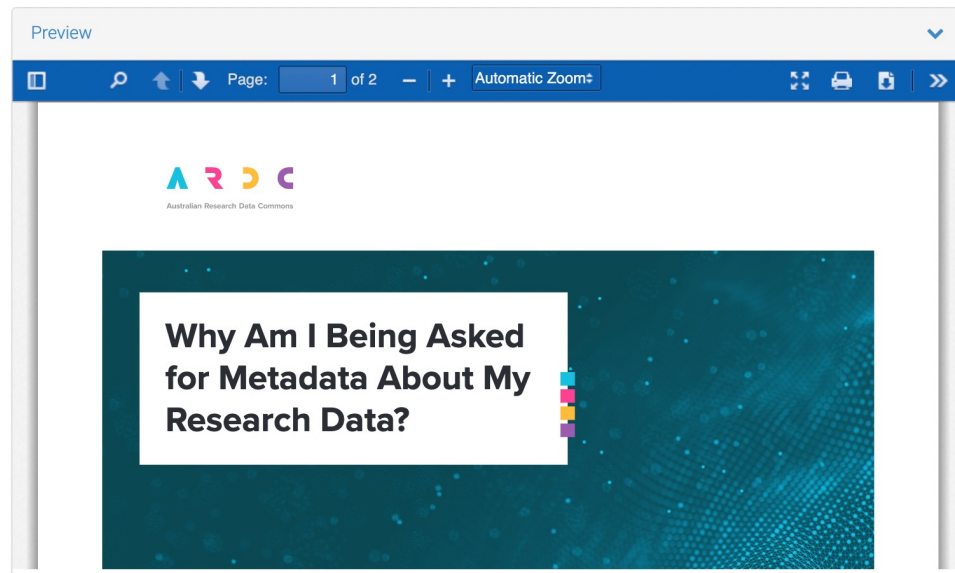
Why am I being asked for metadata about my research data?

Australian Research Data Commons

Find out why metadata are important for your research data collection. This brochure shares the reasons why researchers should use metadata for their data collections.

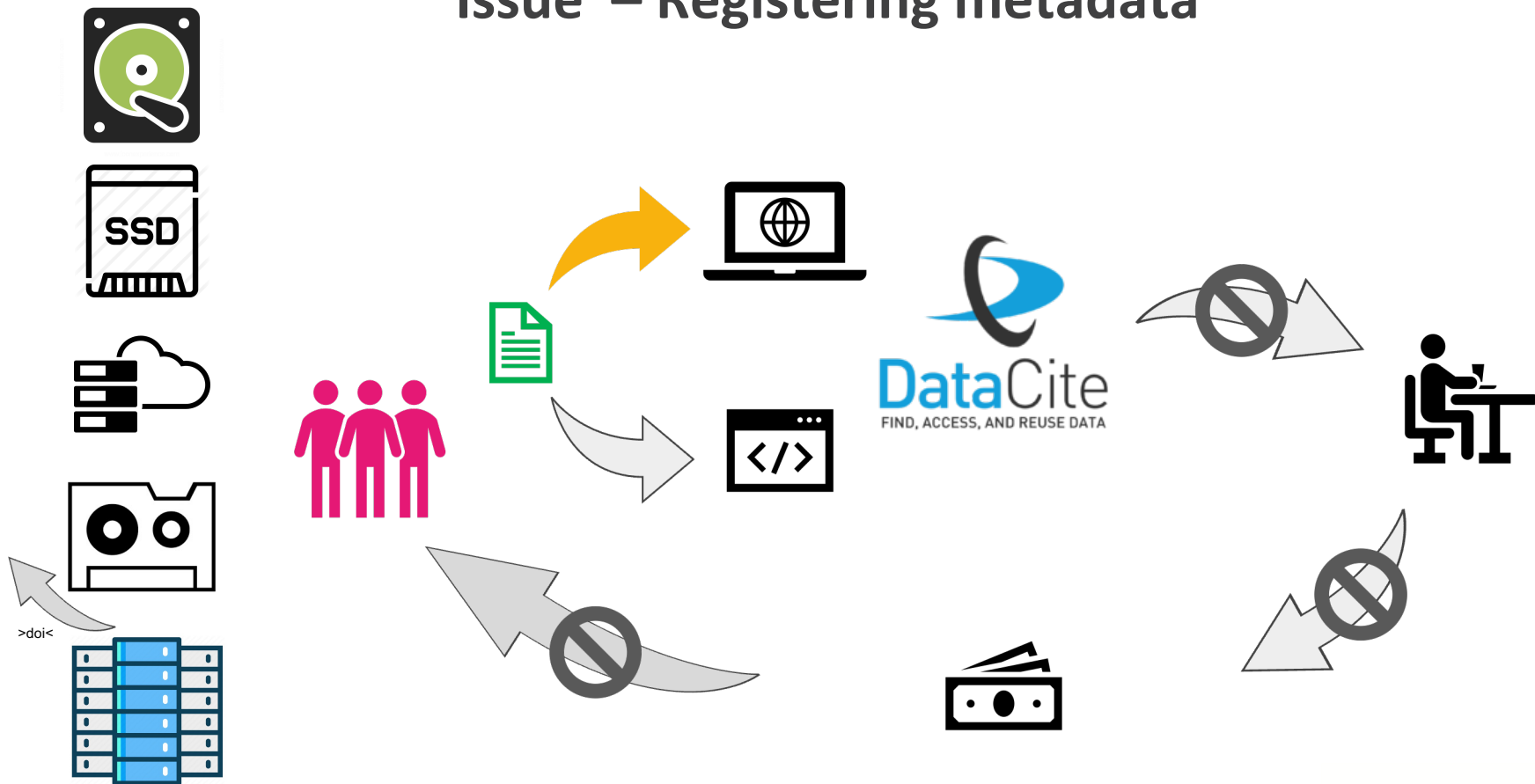
This brochure was prepared for the ARDC Data Retention Project <https://ardc.edu.au/collaborations/strategic-activities/data-retention-project/>.

It is for researchers at any institution in Australia.

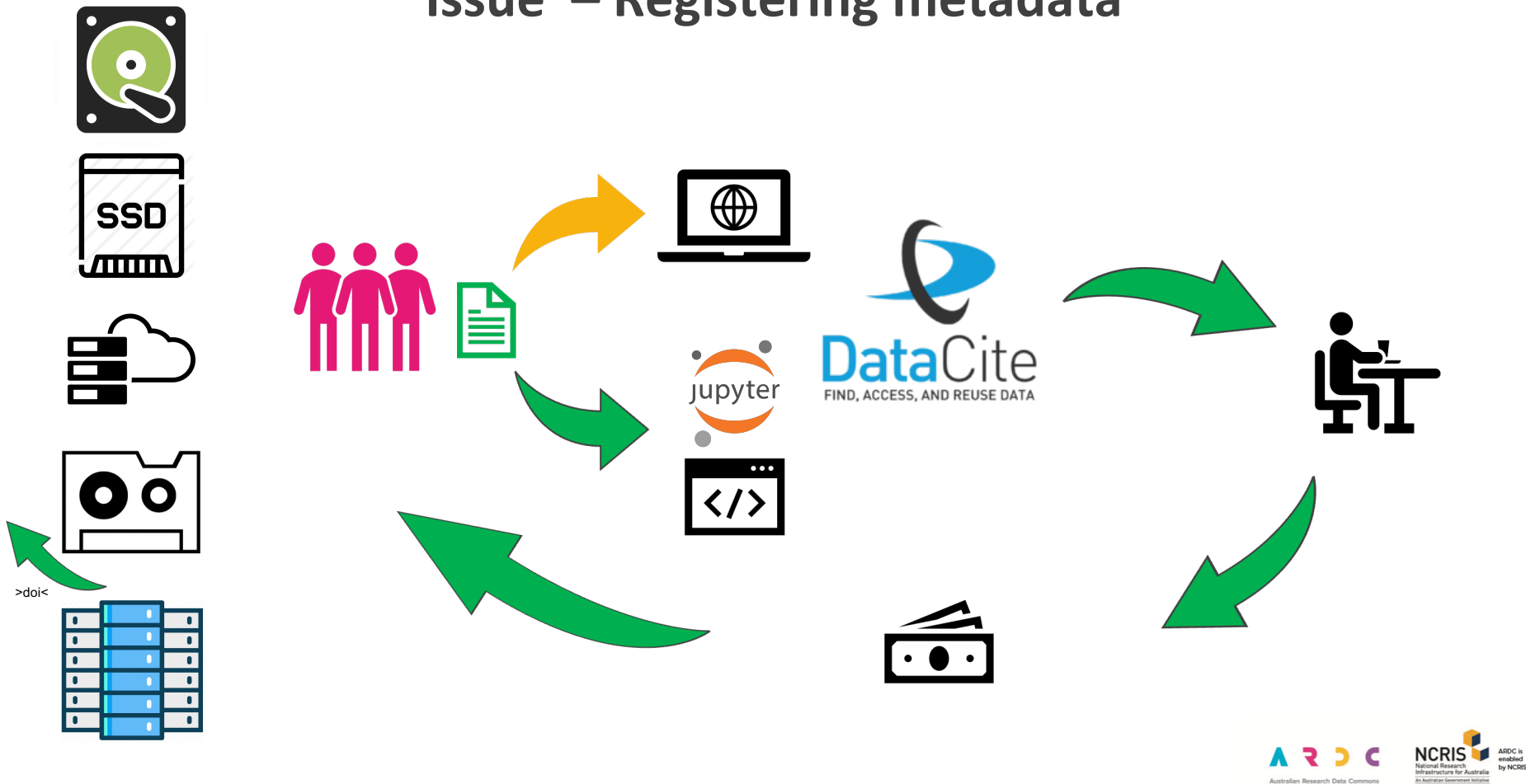


<10.5281/zenodo.5778322>

Issue – Registering metadata






Issue – Registering metadata



Issue – Registering metadata

☰ README.md

 CITATION.cff  passing  Open in Colab

DataCite API notebook

This repository contains a Python Jupyter notebook and associated files that can interact with the [DataCite Member REST API](#) to perform the following functions:

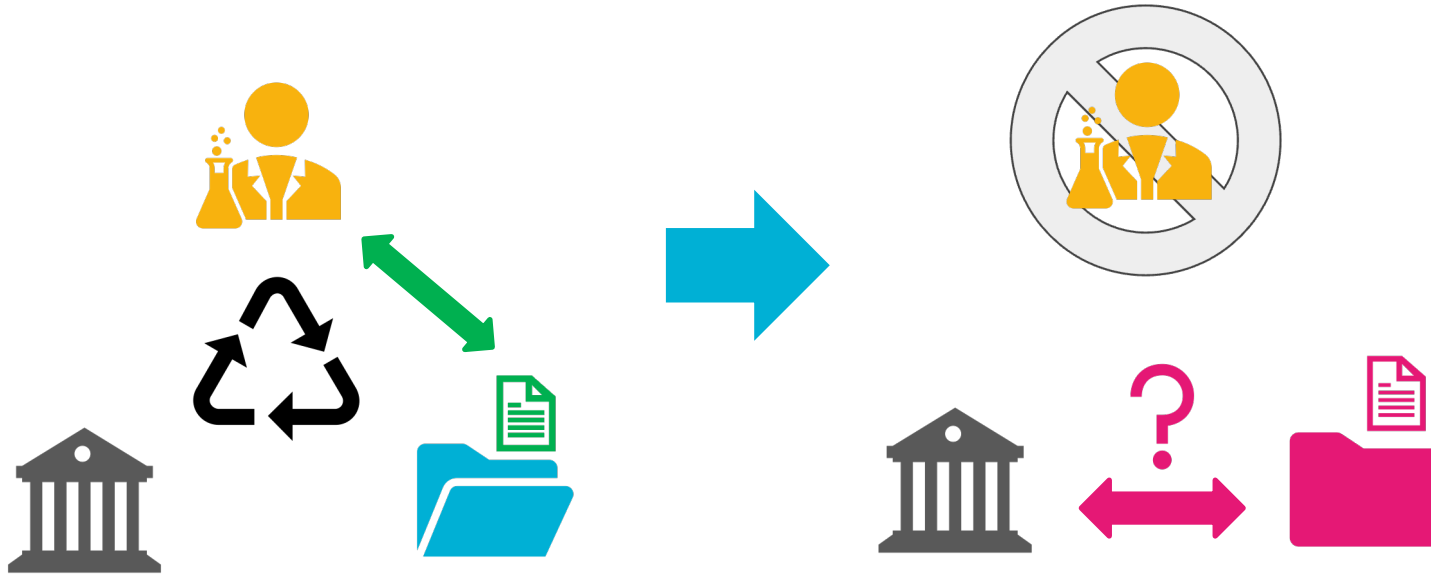
- Mint draft DOIs
- Publish draft DOIs
- Hide published DOIs
- Delete draft DOIs
- Download metadata for existing DOIs
- "Sanitise" downloaded metadata to...
- Update metadata for existing DOIs



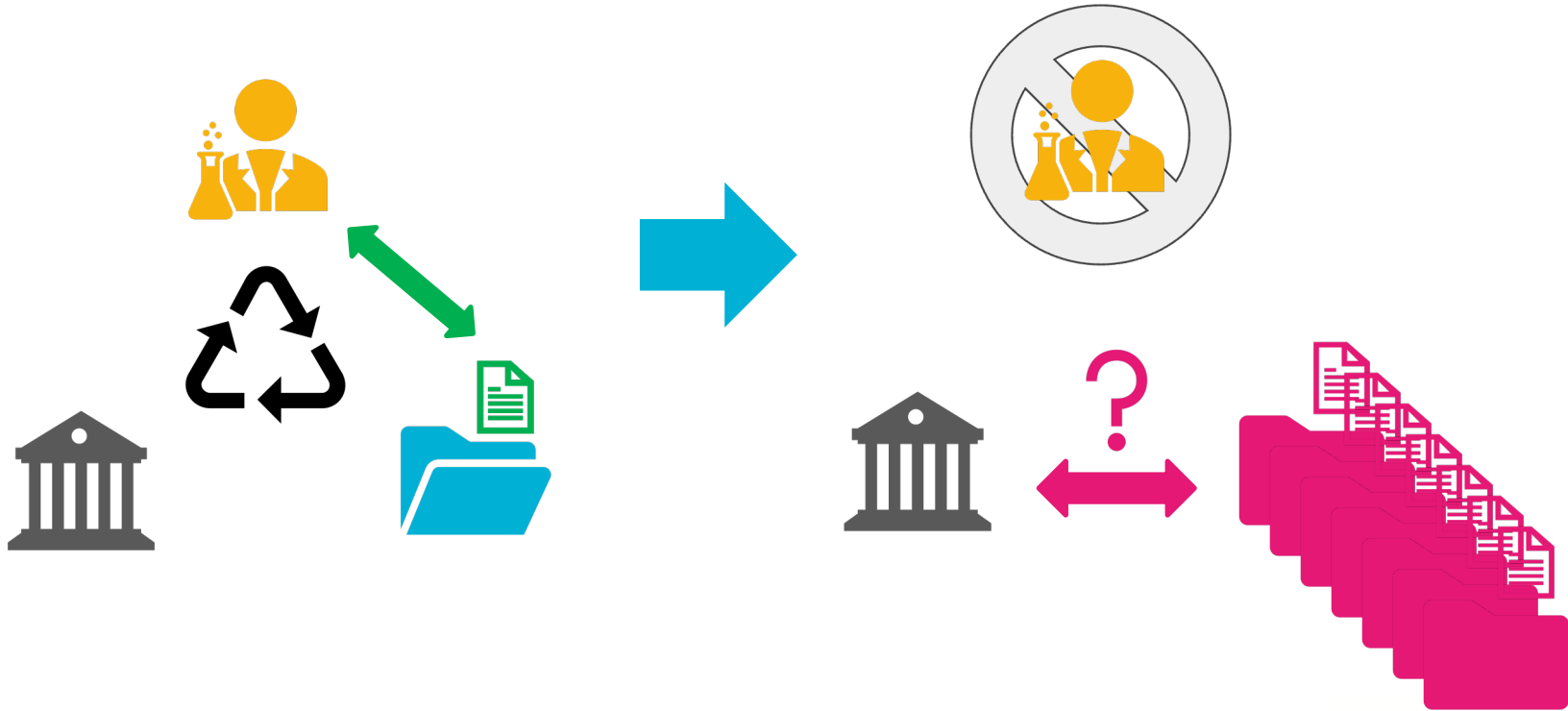
[10.5281/zenodo.5574653](https://doi.org/10.5281/zenodo.5574653)

Liffers, M. (2021). ARDC DataCite API Jupyter notebook (Version 0.1.0) [Computer software].
<https://doi.org/10.5281/zenodo.5574653>

Issue – Extracting metadata



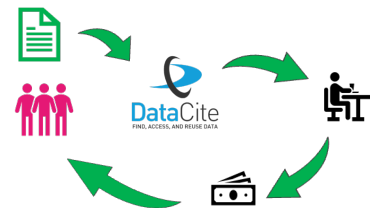
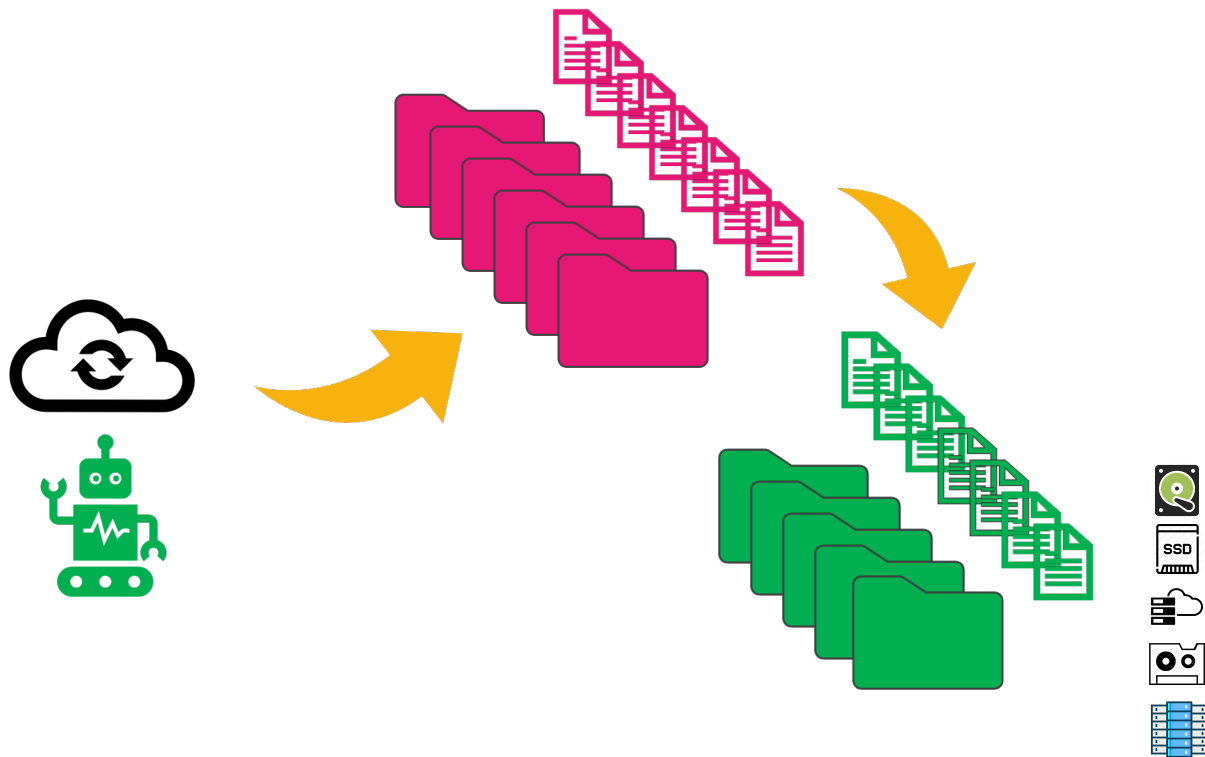
Issue – Extracting metadata



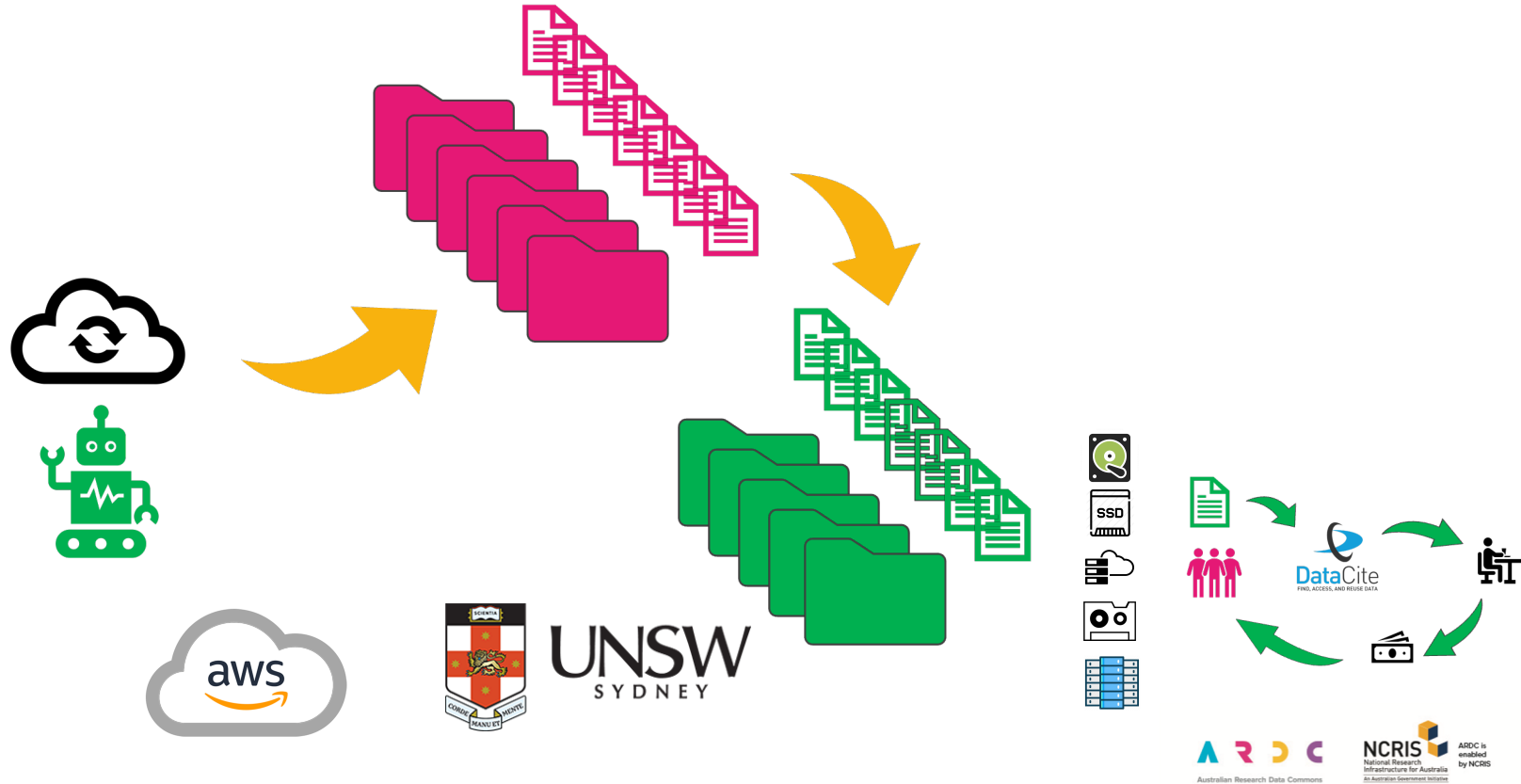
Issue – Extracting metadata



Issue – Extracting metadata

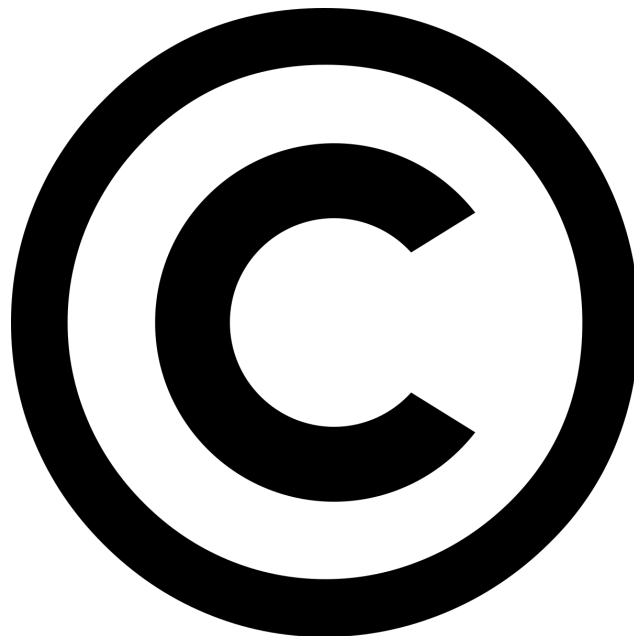


Issue – Extracting metadata



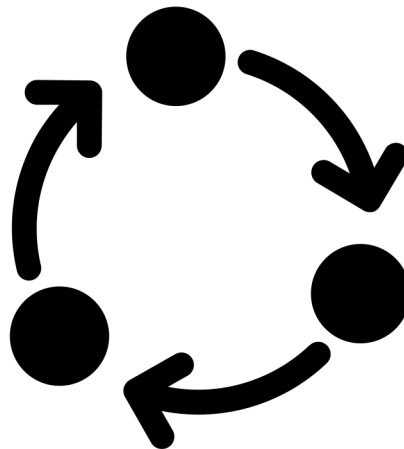
Next Steps

- **Licensing**
- Process Integration
- 'Apex' incentives
- 'Foundation' incentives



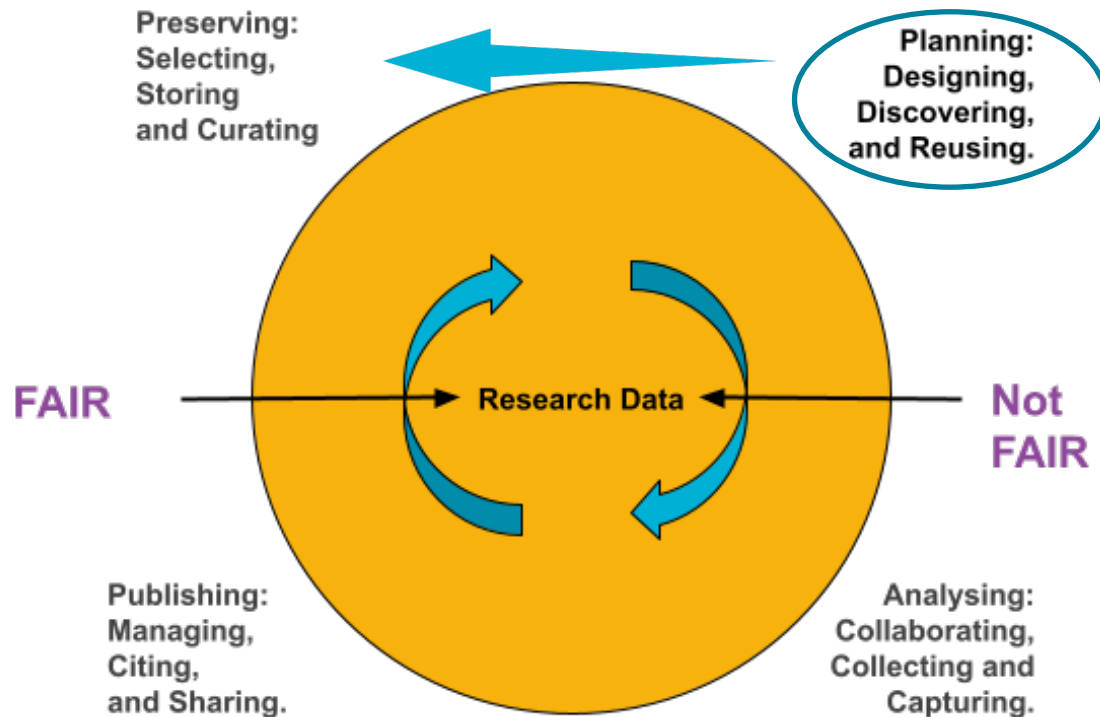
Next Steps

- Licensing
- **Process Integration**
- 'Apex' incentives
- 'Foundation' incentives



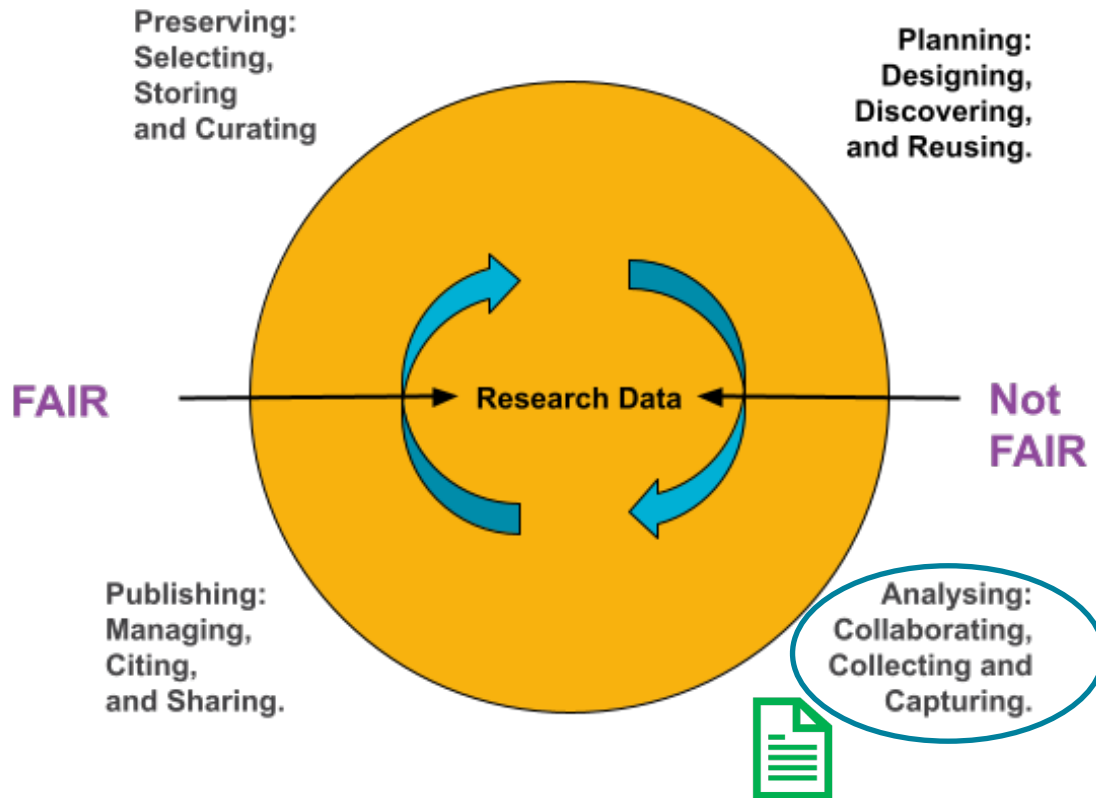
Next Steps

- Licensing
- Process Integration
- 'Apex' incentives
- 'Foundation' incentives



Next Steps

- Licensing
- Process Integration
- 'Apex' incentives
- 'Foundation' incentives



Thank You

- max.wilkinson@ardc.edu.au