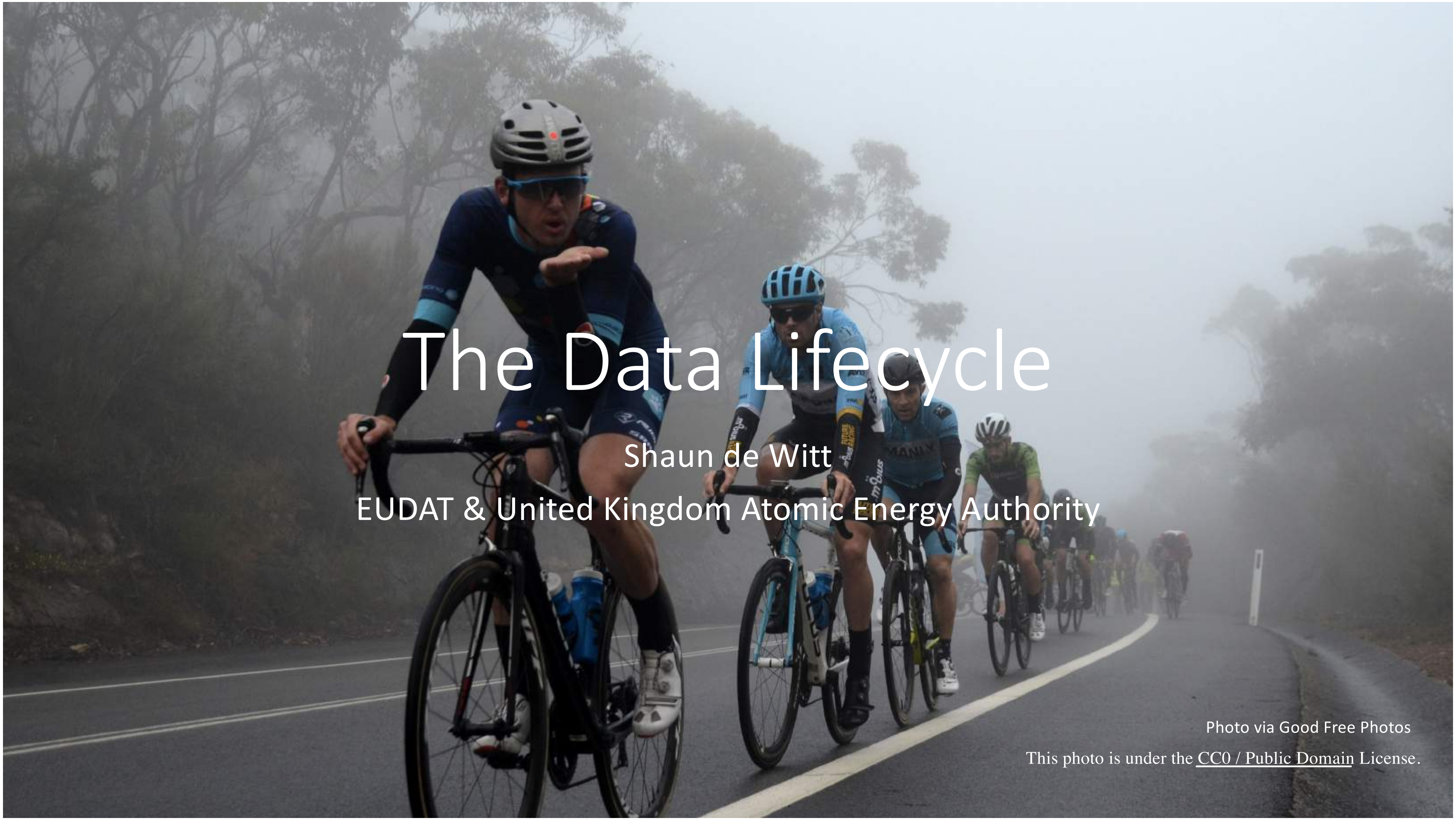


The Data Lifecycle

Shaun de Witt

EUDAT & United Kingdom Atomic Energy Authority

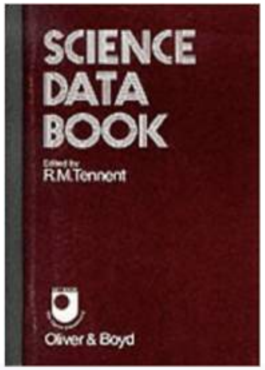
Photo via Good Free Photos

Stuff you might learn…

- Isn’t data lifecycle just recording how many k’s I have cycled??
  - Examples of data lifecycles
- Planning??? But I’m a student!!!
  - Planning to manage your data through its lifetime
- Data Lifecycle for the Real World
- How EUDAT and PRACE Services Fit into the Data Lifecycle

What is Research Data - Sources

👍



👍

👍

CMS Photo: Maximilien Brice/CERN

👍

👍

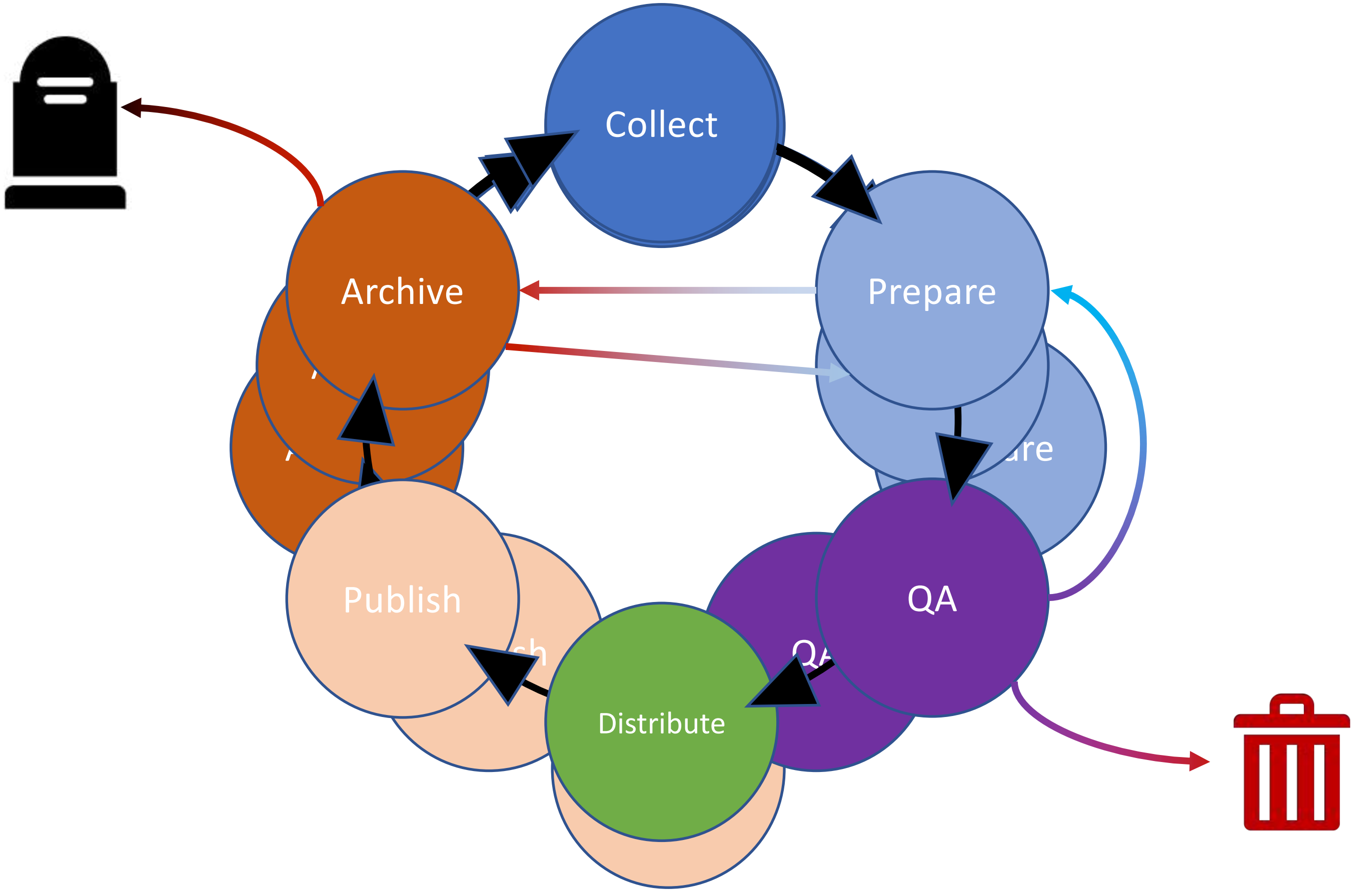
👍

👍

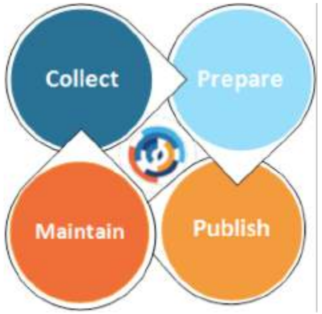
👍

👍

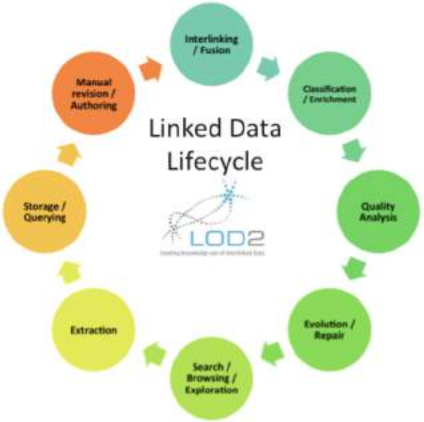
Evolution of the Data Lifecycle



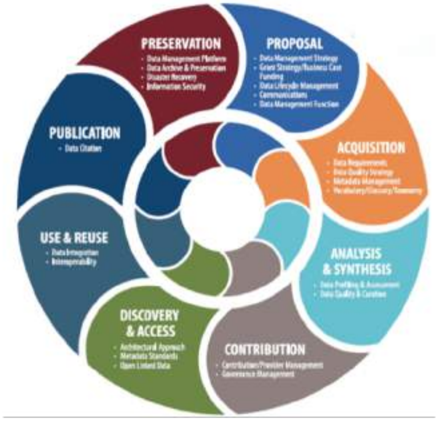
Data Lifecycles… Simple to Complex



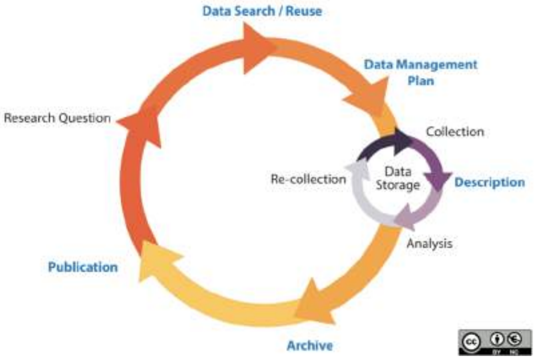
European Data Portal



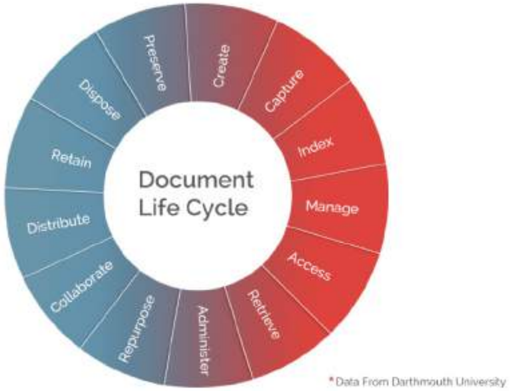
Sören Auer (2011) "The Semantic Data Web"

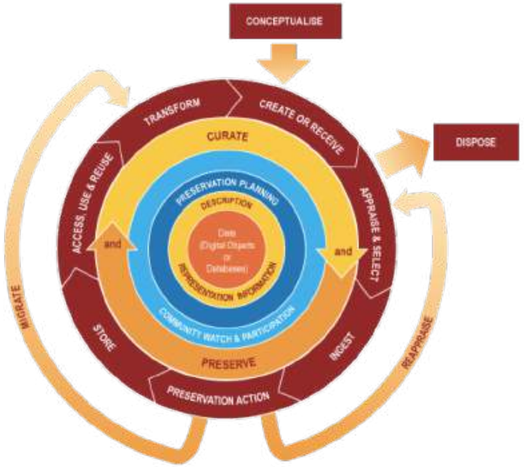


AGU Data Maturity Model



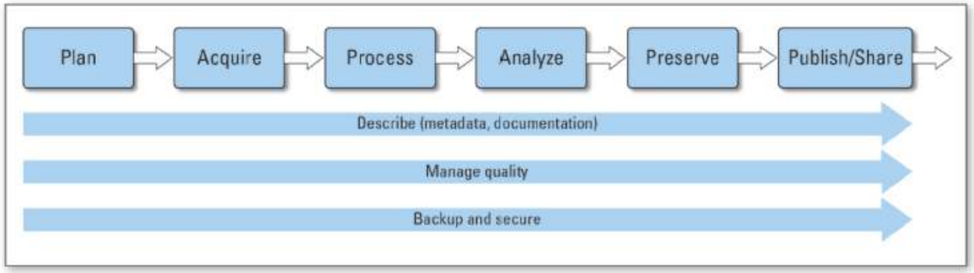
UCSC Data Lifecycle



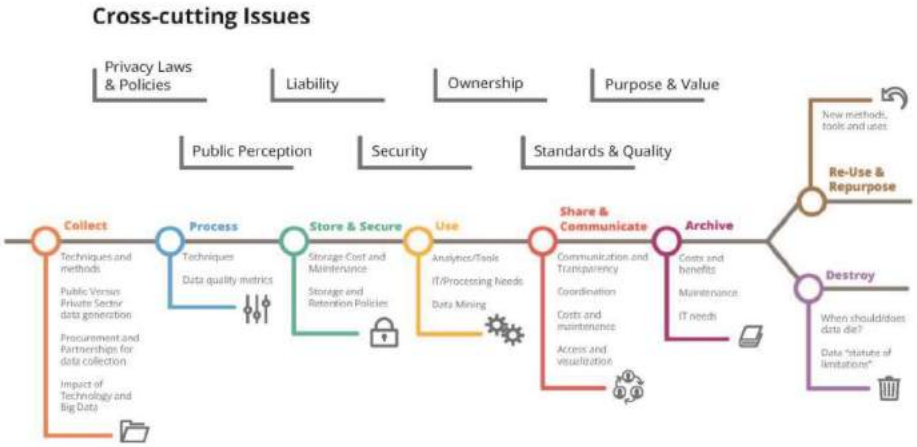


DCC Data Lifecycle

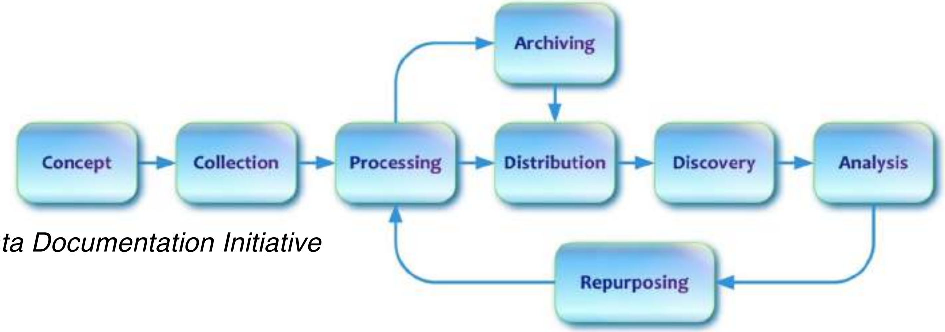
Non Cyclic Data Lifecycles



The United States Geological Survey Science Data Lifecycle Model



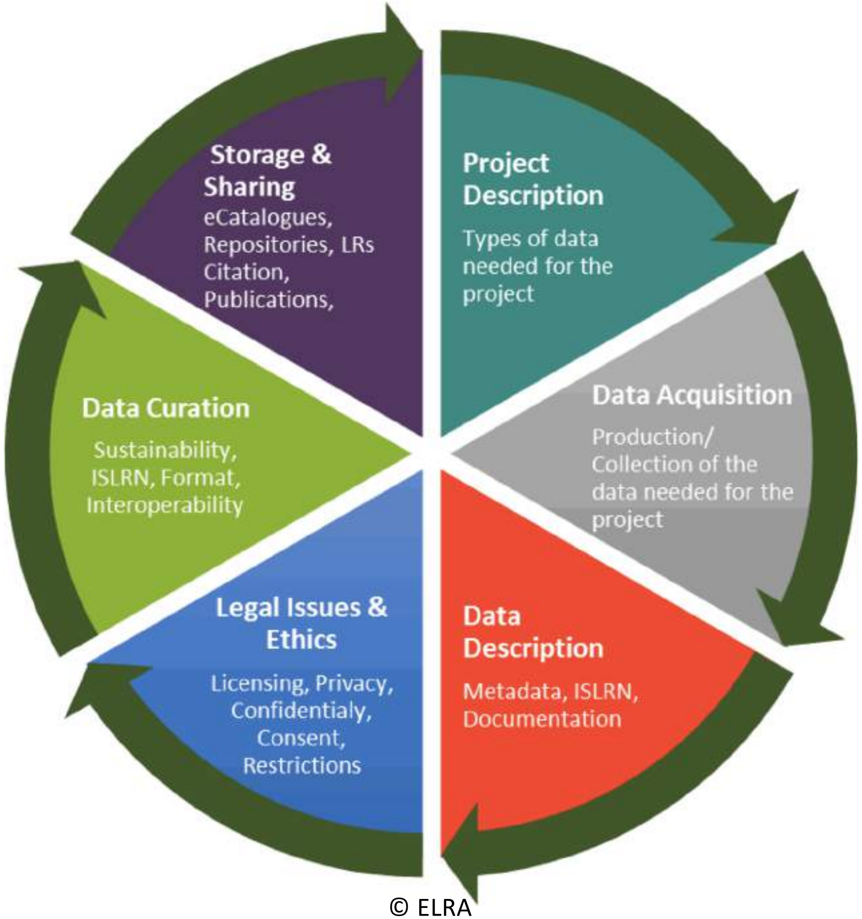
Miller K., Miller M.,Moran M.,Dai B.,
Texas A&M Transportation Institute, PRC 17-84F,
March 2018



Data Documentation Initiative

Data Lifecycles - Planning

- **<u>Data Management Plans</u>**
  - What type of data
    - Content rather than format
  - How will it be acquired
    - Raw sensor, survey, harvesting,…
  - Description of data
    - Metadata, formats, volume, …
  - Legal, ethical and commercial considerations
    - Licensing, embargo periods,…
  - Data curation
    - Longevity, ongoing costs,…
  - Storage & Sharing
    - Cataloging, location, accessibility,…



Project Description
Types of data needed for the project

Data Acquisition
Production/ Collection of the data needed for the project

Data Description
Metadata, ISLRN, Documentation

Legal Issues & Ethics
Licensing, Privacy, Confidentialy, Consent, Restrictions

Data Curation
Sustainability, ISLRN, Format, Interoperability

Storage & Sharing
eCatalogues, Repositories, LRs Citation, Publications,

© ELRA

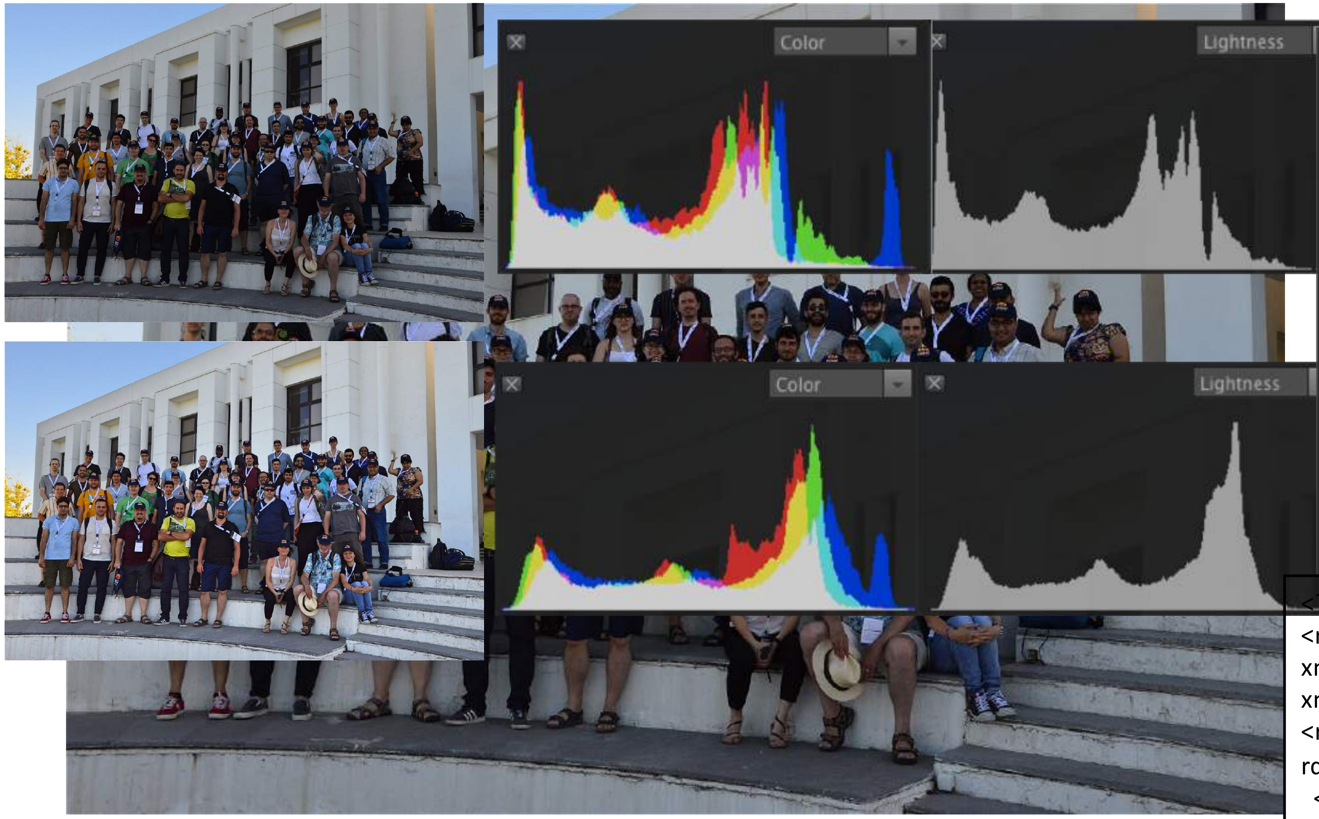
Data Lifecycle – Practical Example (1)

Process…

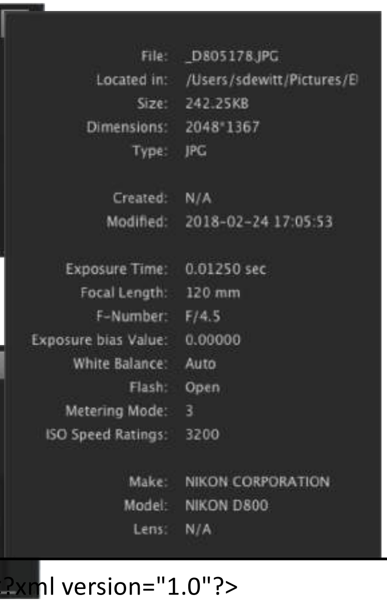
Create…

Analyse…

Data Lifecycle – Practical Example (1)



Add Metadata



File: _D805178.JPG
Located in: /Users/sdewitt/Pictures/E
Size: 242.25KB
Dimensions: 2048*1367
Type: JPG

Created: N/A
Modified: 2018-02-24 17:05:53

Exposure Time: 0.01250 sec
Focal Length: 120 mm
F-Number: F/4.5
Exposure bias Value: 0.00000
White Balance: Auto
Flash: Open
Metering Mode: 3
ISO Speed Ratings: 3200

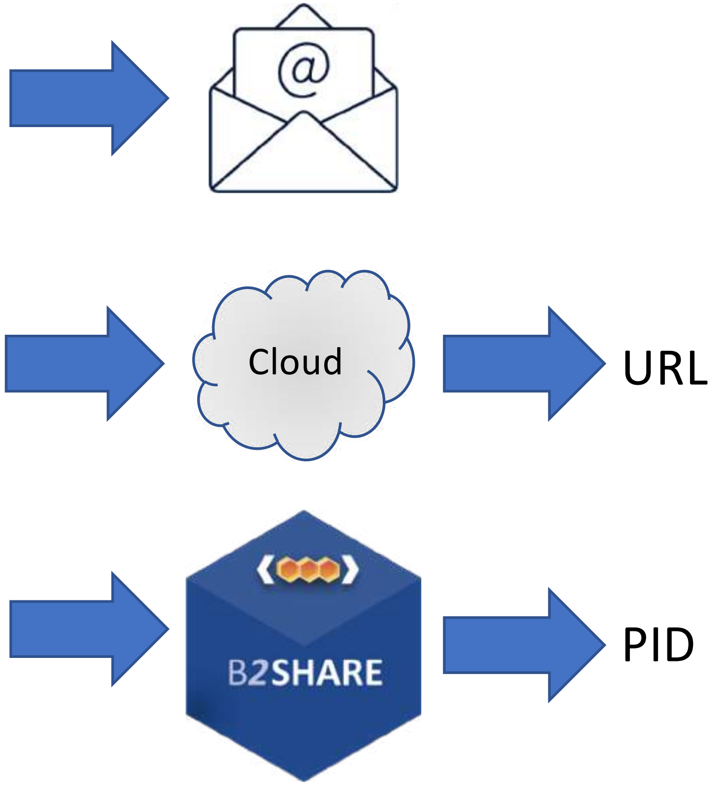
Make: NIKON CORPORATION
Model: NIKON D800
Lens: N/A

Enrich Metadata

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:photo="http://www.photos.fake/photos#">
<rdf:Description
rdf:about="http://www.photos.fake/photos/EUDAT">
  <photo:location>Heraklion</photo:location>
  <photo:event>Summer School</photo:event>
  <photo:year>2017</photo:year>
  <photo:license>CC-BY-NC-SA-4.0</photo:license>
</rdf:Description>
```

Data Lifecycle – Practical Example (3)

Image + Metadata



Cloud

URL

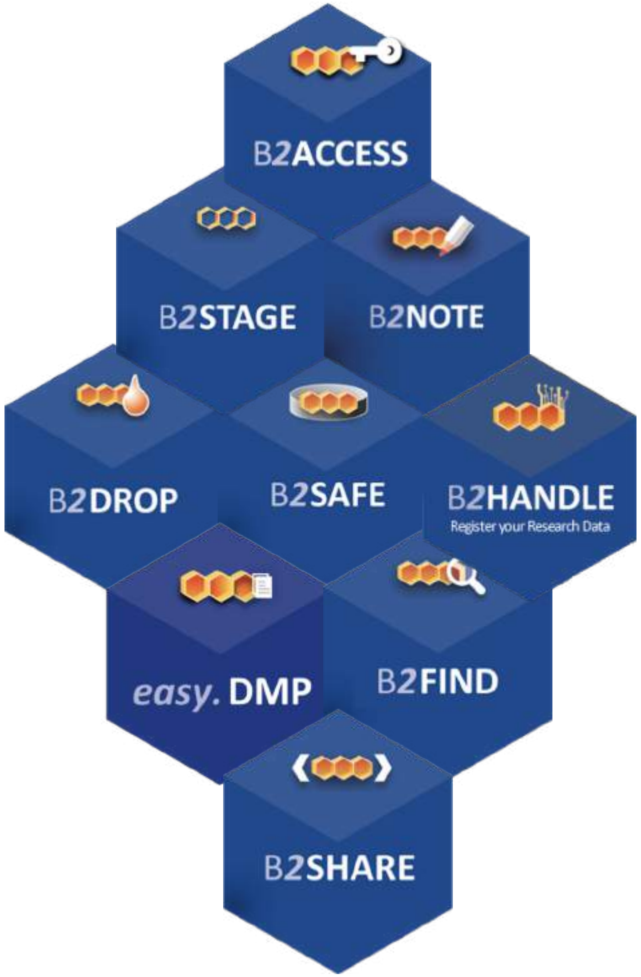
B2SHARE

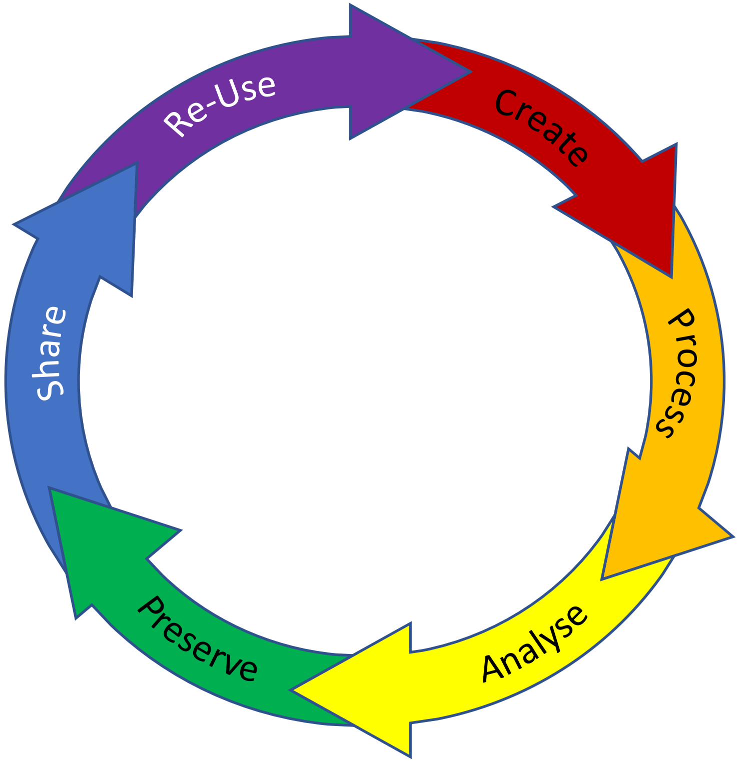
PID

Intermezzo

- The Rules of Data Lifecycles
    - There is **no one** data lifecycle
        - There is **no right** data lifecycle (but there are many wrong ones)
    - Sometimes the data lifecycle is **not cyclic**
    - The data lifecycle is documented in a **Data Management Plan** (DMP)
        - And most **funding authorities** make you write one
    - **Don’t roll your own** – Use institutional or community ones where they exist
    - The **DMP is an output** of research – it should follow it’s own rules
- We will do an exercise on data management planning later this week

EUDAT & PRACE in the DLC

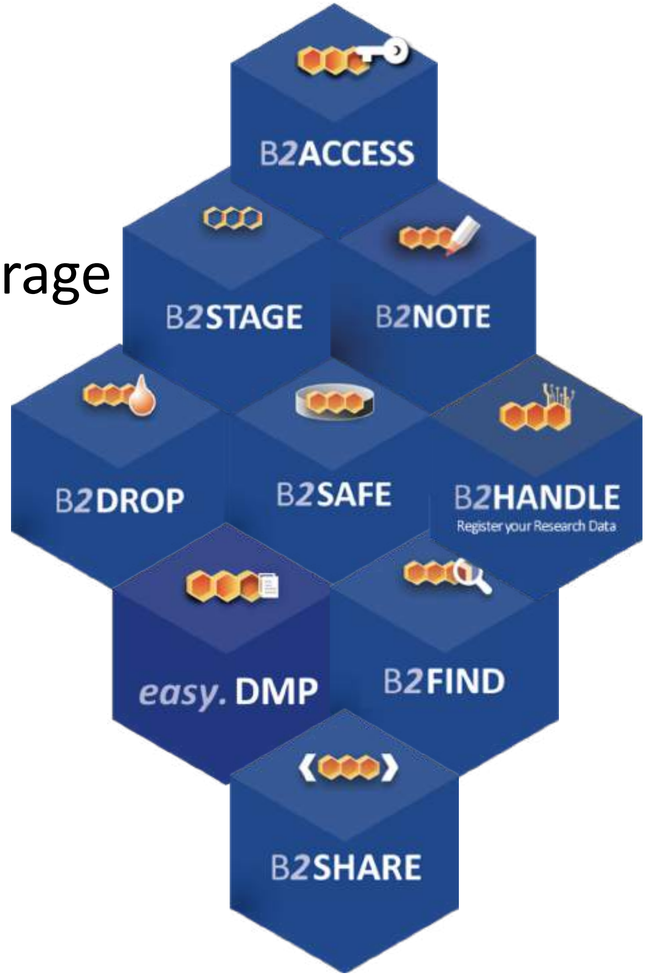




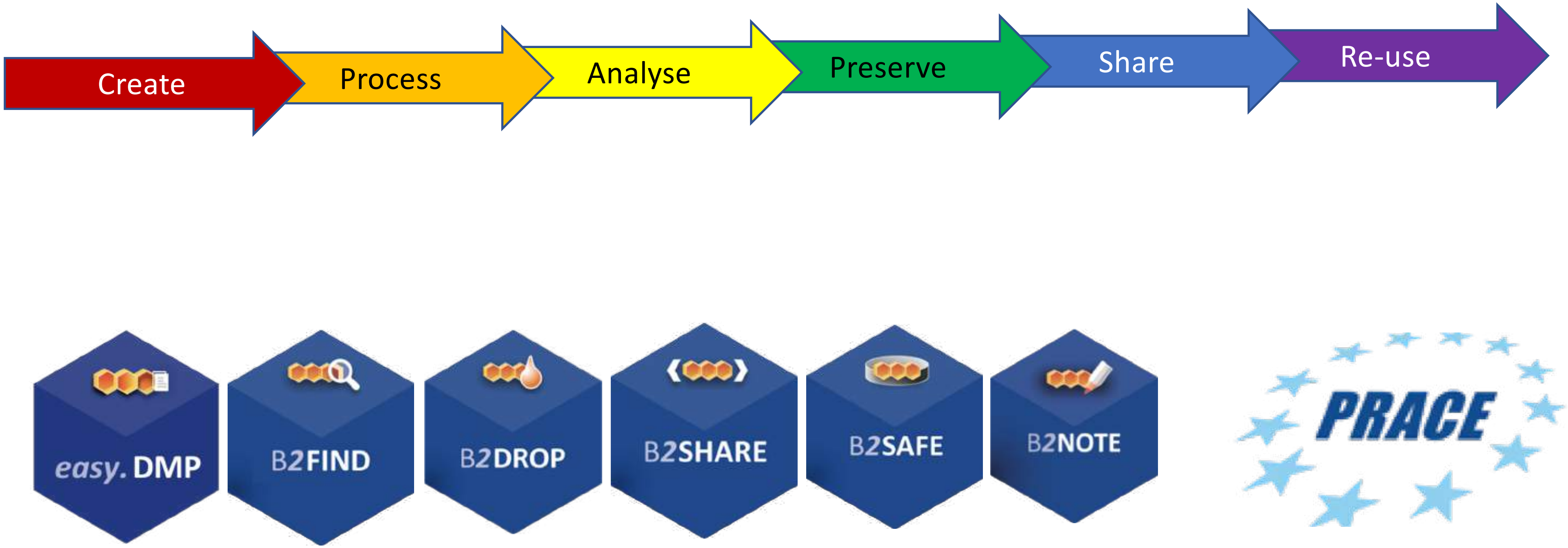
PRACE

EUDAT Services – 1 Line Summaries

- **B2ACCESS** – Authentication and Authorisation
- **B2DROP** – Data Workspace
- **B2SAFE** – Distributed, Secure Policy Based Data Storage
- **B2SHARE** – Searchable Data Repository
- **B2STAGE** – High Performance Data Movement
- **B2FIND** – Searchable Metadata Aggregator
- **B2HANDLE** – Persistent Identifier Provider
- **B2NOTE** – Semantic Metadata Annotation
- **easy.DMP** – Data Management Planning Assistant



Service Mapping



Create

Process

Analyse

Preserve

Share

Re-use

easy.DMP

B2FIND

B2DROP

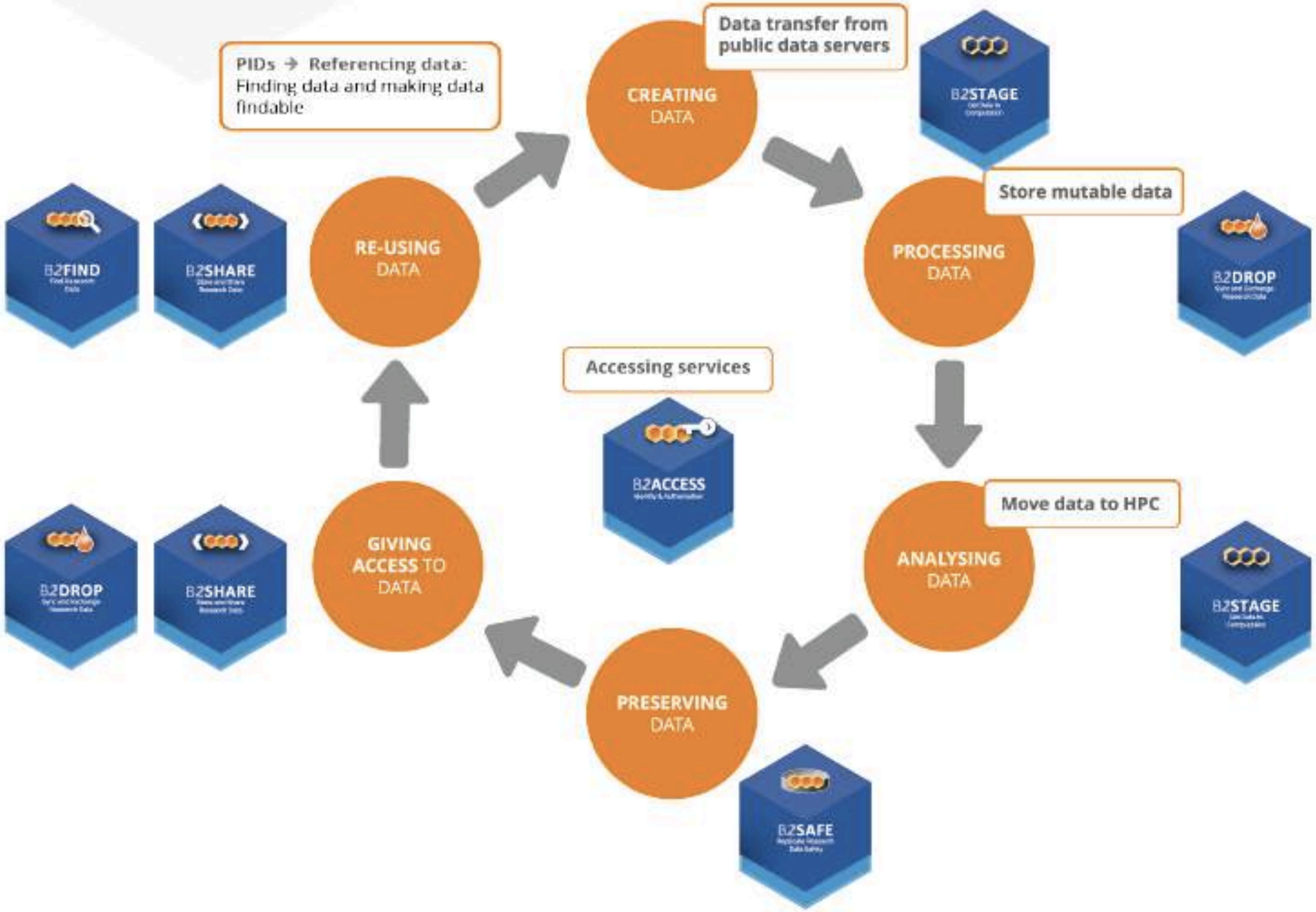
B2SHARE

B2SAFE

B2NOTE

PRACE

EUDAT Services & the Data Lifecycle – Simplified



THAT’S ALL FOLKS

- EUDAT & PRACE Offer Services Supporting **any** Data Lifecycle
- Services are **generic** and not aimed at any one science
- Services are **defined by scientists** who understand their data lifecycle
- Services are **run by scientific institutes** for scientists
- Services are supported by a quality **service management framework**

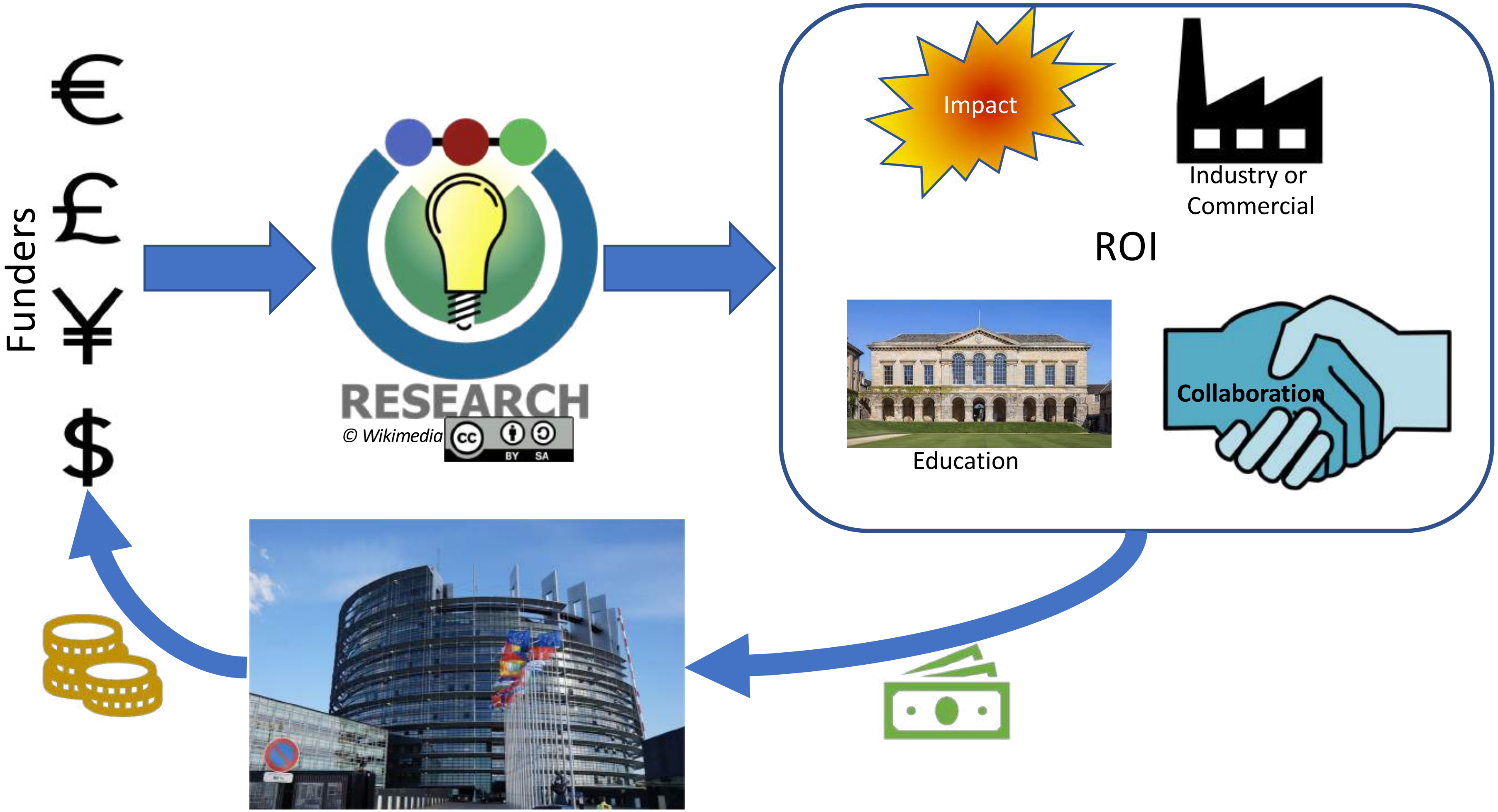
EUDAT Collaborative Data Infrastructure

https://www.eudat.eu/eudat-collaborative-data-infrastructure-cdi

PRACE

http://www.prace-ri.eu/

Why do funders care?



Funders

€

£

¥

$

RESEARCH

© Wikimedia

Impact

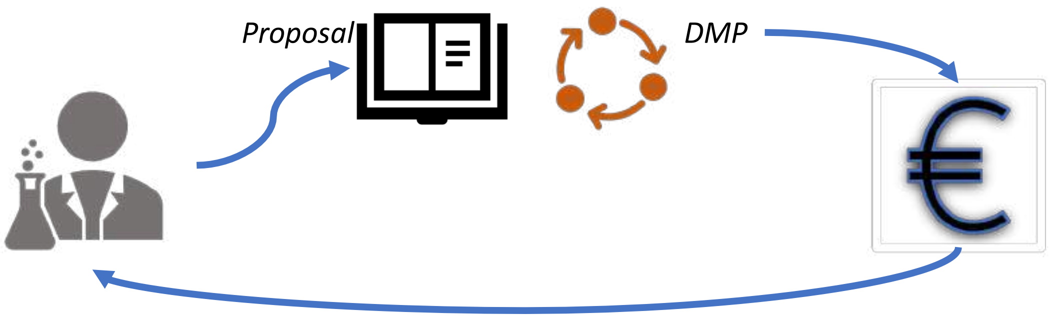
Industry or Commercial

ROI

Education

Collaboration

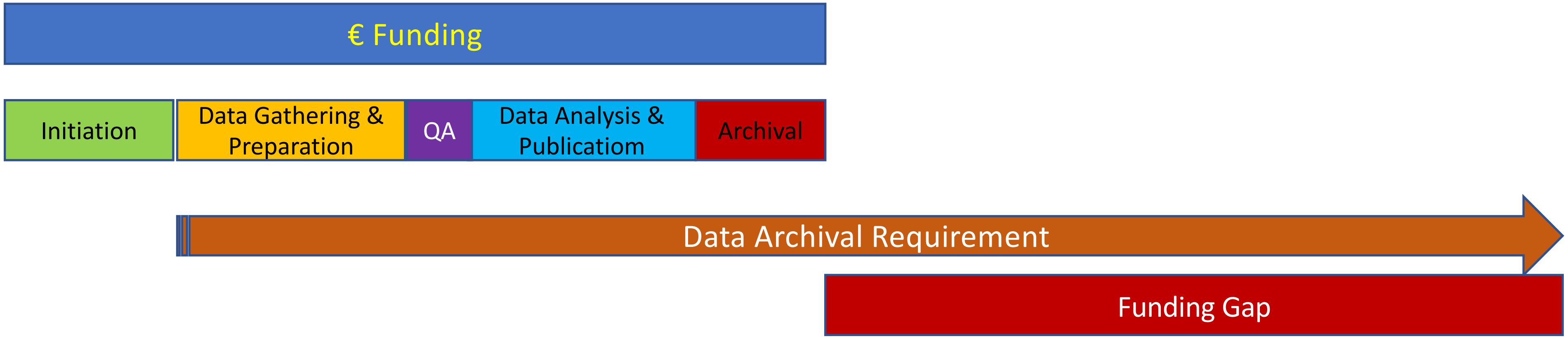
Data Lifecycle – the PI’s View



Proposal

DMP

The Data Lifecycle Problem



€ Funding

Initiation

Data Gathering & Preparation

QA

Data Analysis & Publicatiom

Archival

Data Archival Requirement

Funding Gap

Cost Estimation Game - TV

- Lets estimate the price of the following 10 years ago…

50-54” 4K colour TV

Price 2019: $470

Higher

Lower

Price 2010: $2500

Source: Statista 2019

Cost Estimation Game – Cup of Coffee

Price 2019: $2.10

Higher Lower

Price 2010:	$1.50

Cost Estimation Game – Storage per GB

Price 2019: $0.022/month

Pure storage costs on large cloud providers. Does not include networking or transaction charges

Higher Lower

Price 2010: $0.15/month

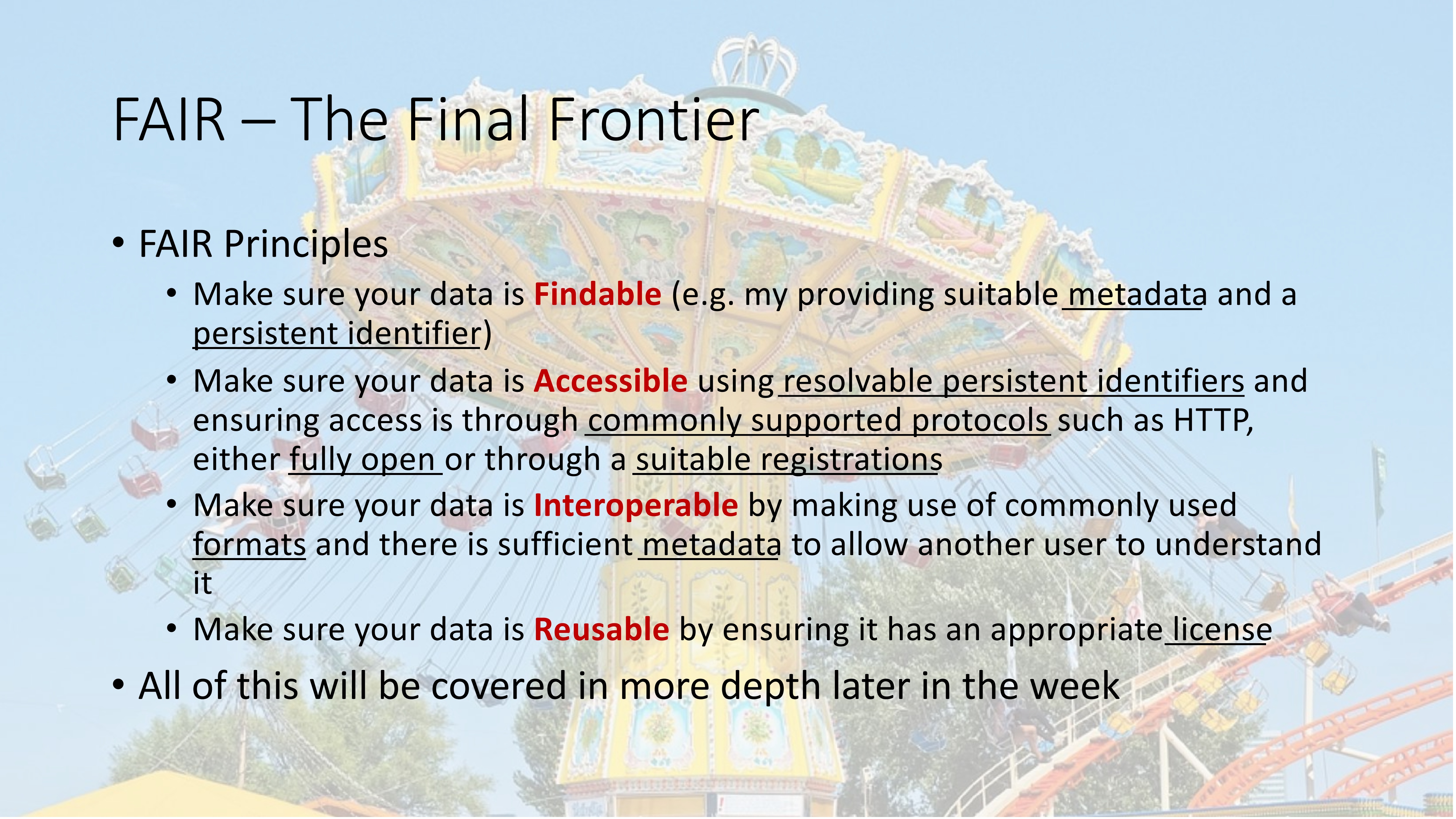
Source: https://www.nasuni.com/57-whats_the_cost_of_a_gb_in_the_cloud/

Price 2019: $0.024

Higher Lower

Price 2010: $0.06

Source: https://jcmit.net/diskprice.htm



FAIR – The Final Frontier

- FAIR Principles
  - Make sure your data is **Findable** (e.g. my providing suitable metadata and a persistent identifier)
  - Make sure your data is **Accessible** using resolvable persistent identifiers and ensuring access is through commonly supported protocols such as HTTP, either fully open or through a suitable registrations
  - Make sure your data is **Interoperable** by making use of commonly used formats and there is sufficient metadata to allow another user to understand it
  - Make sure your data is **Reusable** by ensuring it has an appropriate license
- All of this will be covered in more depth later in the week

Conclusions

- Data can come from **many different sources**
- Data has a lifecycle covering **generation, processing, archiving and re-use**
  - While all lifecycles take a similar form, there may be specifics for your research
  - **EUDAT** Services support the management of data
- The data lifecycle is documented in a **Data Management Plan**
- The DMP needs to consider the cost of **long term archival** or **curation**
- The DMP should aim to make data **FAIR** to support it's future use