



National Oceanography Centre
British Oceanographic Data
Centre BODC

Justin Buck

James Ayliffe

jbuck@bodc.ac.uk jamayl@bodc.ac.uk

AGU community of practice citation use case (NOC BODC)

Introduction - BODC data collections

Series schema

- Oldest schema for research cruise data and trajectories



Samples schema

- For ocean samples



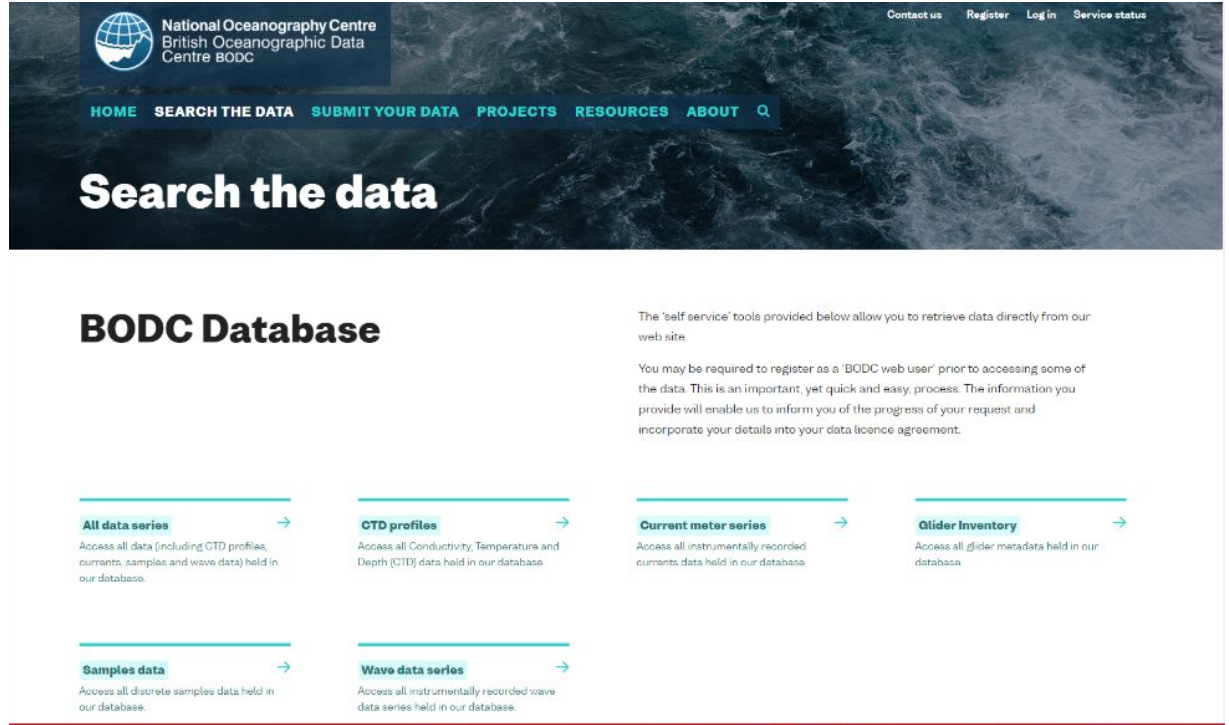
Autonomous platform data system

- New schema for ocean robots



Introduction - BODC data collections

When users download data from BODC collections they need to be able to cite the data.



The screenshot displays the BODC website interface. At the top left is the logo for the National Oceanography Centre British Oceanographic Data Centre BODC. To the right are links for Contact us, Register, Log in, and Service status. Below the logo is a navigation menu with links for HOME, SEARCH THE DATA, SUBMIT YOUR DATA, PROJECTS, RESOURCES, and ABOUT, followed by a search icon. The main heading is "Search the data".

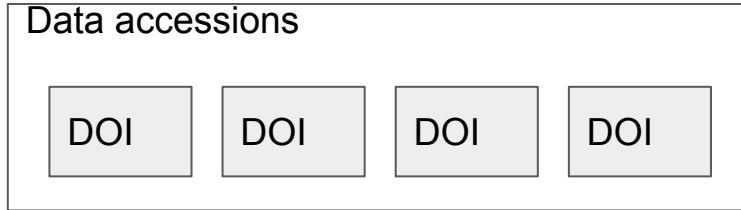
BODC Database

The 'self service' tools provided below allow you to retrieve data directly from our web site.

You may be required to register as a 'BODC web user' prior to accessing some of the data. This is an important, yet quick and easy, process. The information you provide will enable us to inform you of the progress of your request and incorporate your details into your data licence agreement.

- All data series** → Access all data (including CTD profiles, currents, samples and wave data) held in our database.
- CTD profiles** → Access all Conductivity, Temperature and Depth (CTD) data held in our database.
- Current meter series** → Access all instrumentally recorded currents data held in our database.
- Glider Inventory** → Access all glider metadata held in our database.
- Samples data** → Access all discrete samples data held in our database.
- Wave data series** → Access all instrumentally recorded wave data series held in our database.

Use case

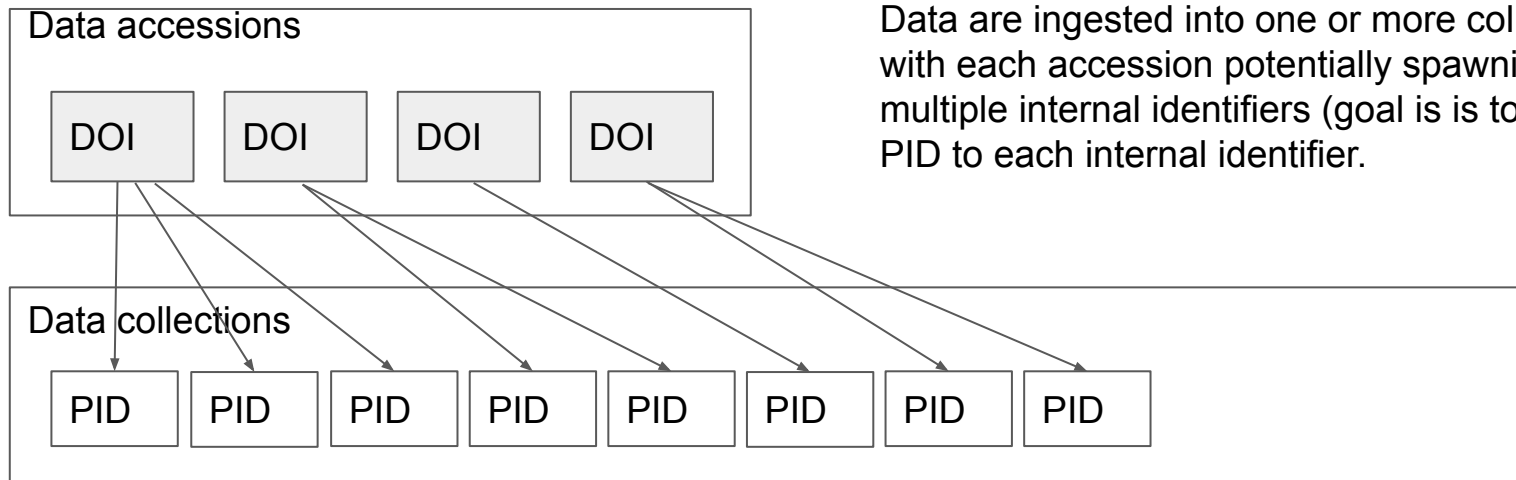


Data from studies are submitted by originators and will have a DOI assigned to the data accession (making the data citable in academic literature).

There are currently 8000 accessions

- Data span sea level data from the 1800's to NRT data up to the present day

Use case

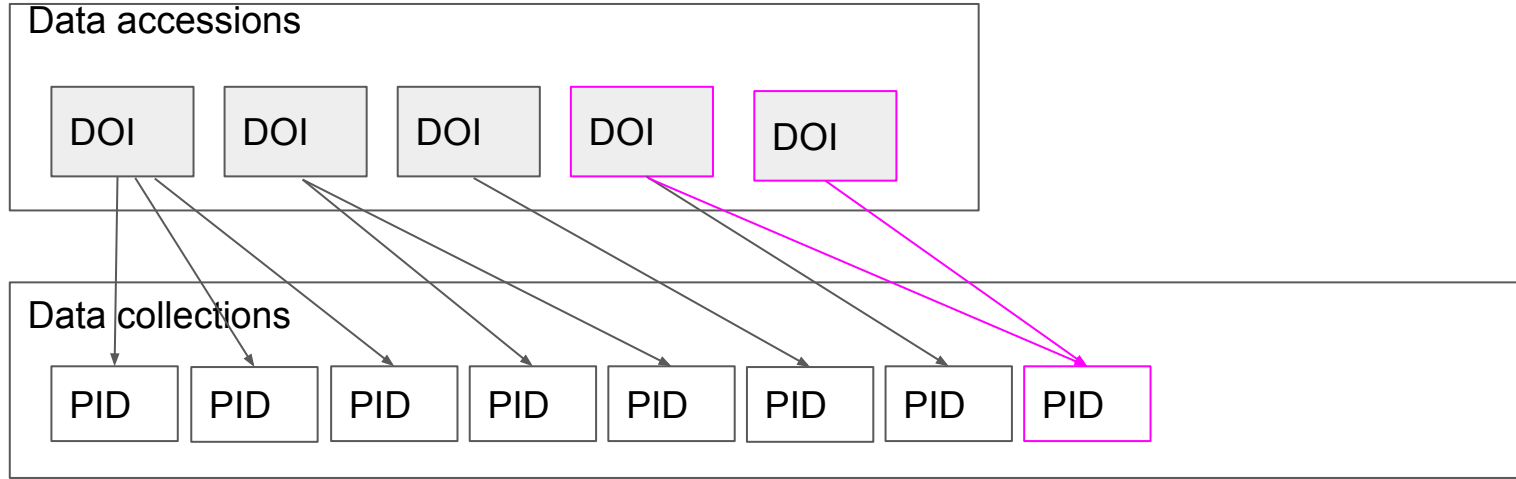


Data are ingested into one or more collections with each accession potentially spawning multiple internal identifiers (goal is to assign a PID to each internal identifier).

Each element within the collections can currently be accessed by a web url that returns it metadata with a download link e.g. <https://www.bodc.ac.uk/data/documents/series/1357426/>
(This is not currently a PID though)

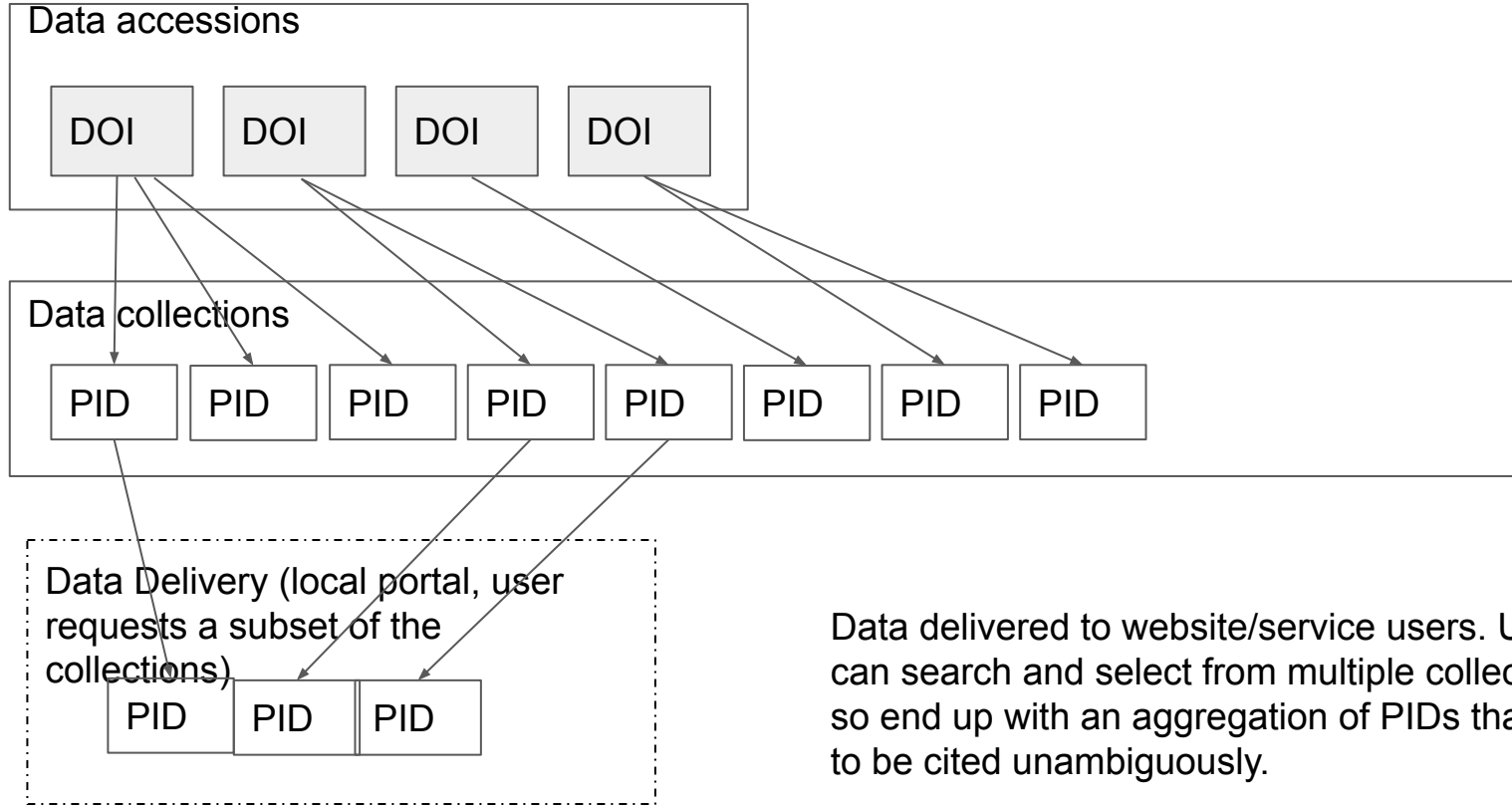
Collections range from 175,000 to 5,500,000+ data granules

Use case



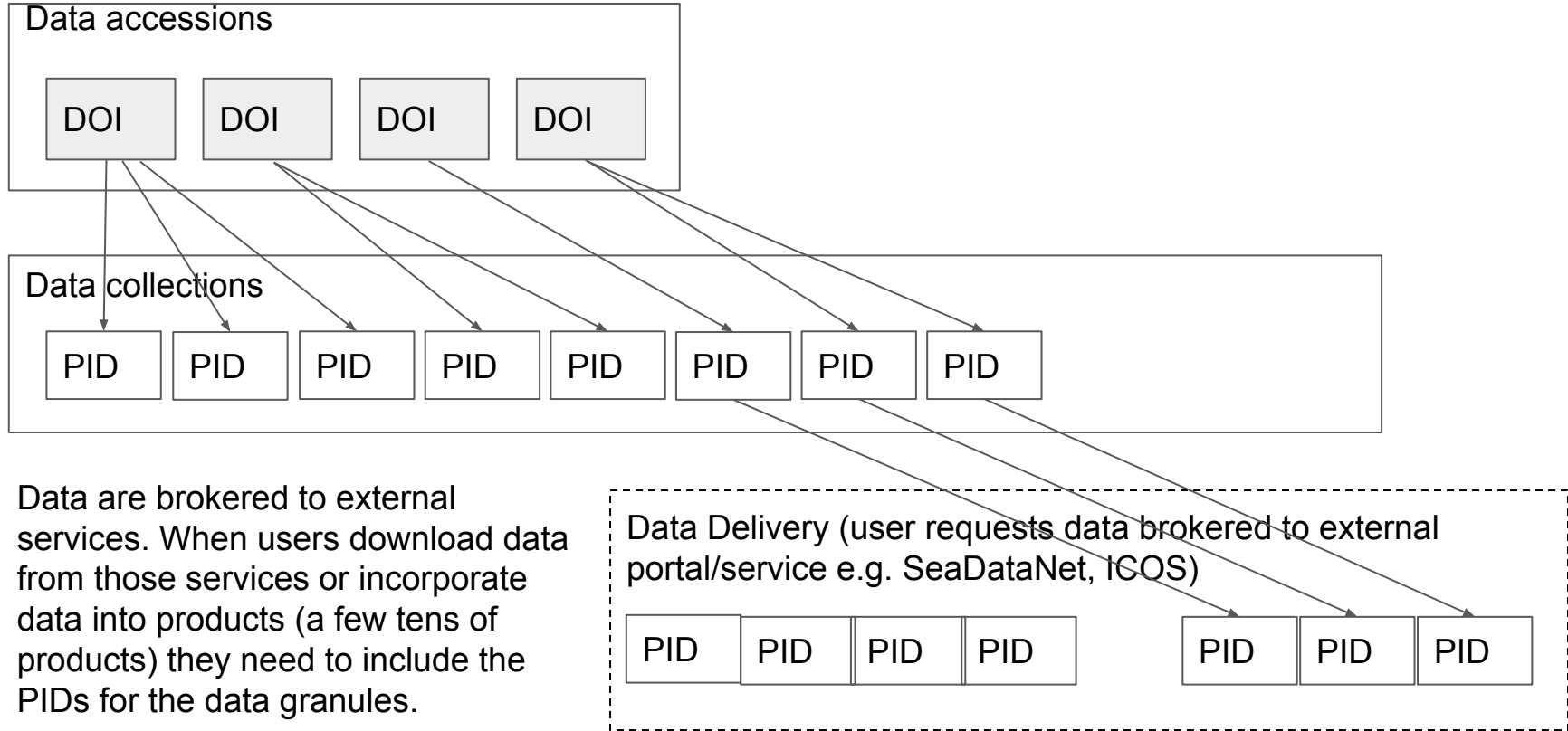
Edge case - for some of our data streams several DOIs can map to a single PID e.g. different PIs QC different variables to form a single granule.

Use case



Data delivered to website/service users. Users can search and select from multiple collections so end up with an aggregation of PIDs that need to be cited unambiguously.

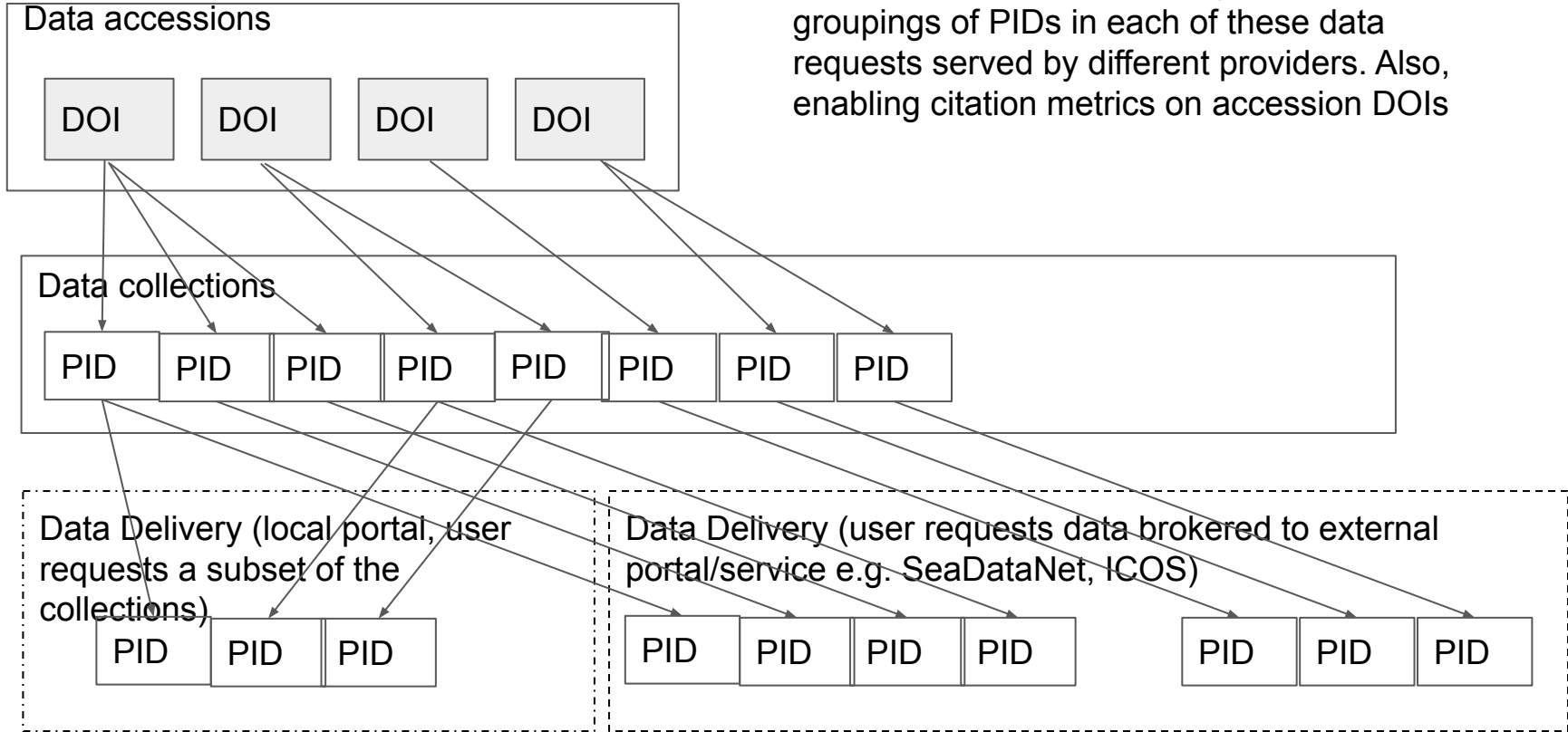
Use case



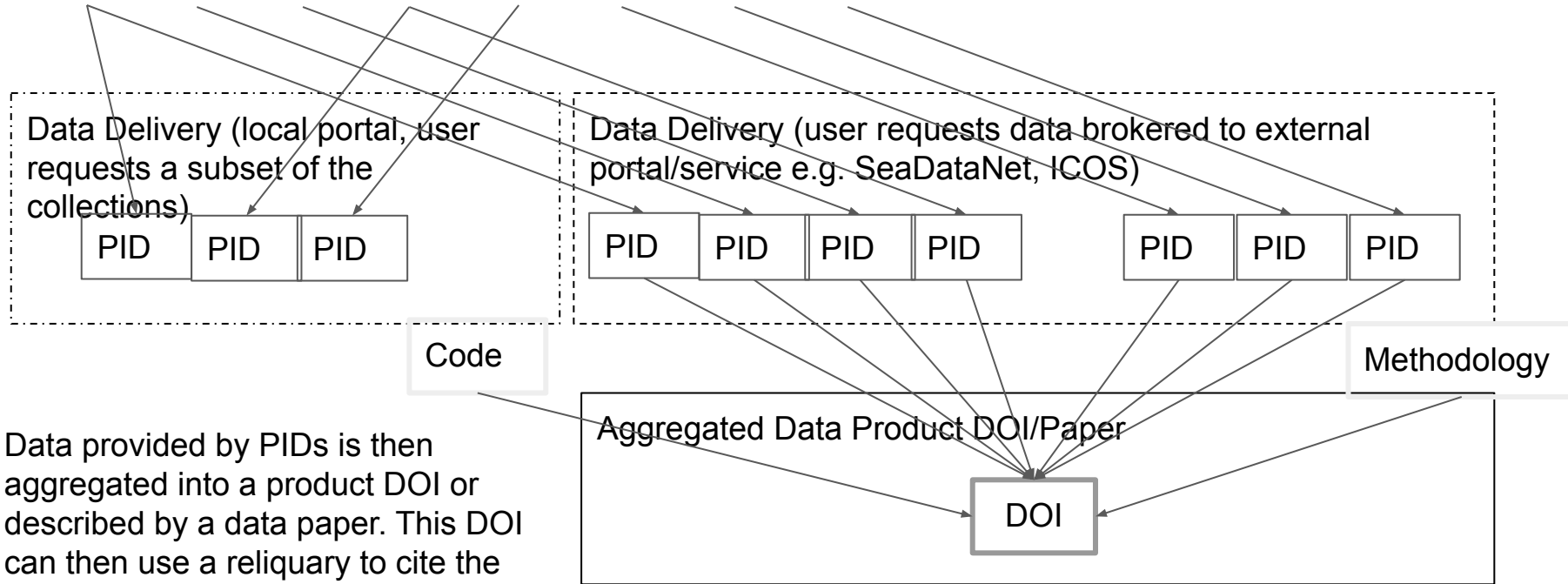
Data are brokered to external services. When users download data from those services or incorporate data into products (a few tens of products) they need to include the PIDs for the data granules.

Use case

Need to be able to accurately cite the groupings of PIDs in each of these data requests served by different providers. Also, enabling citation metrics on accession DOIs



“Aggregate DOI” Inter data services citation?



Solution - primary goal of unambiguous citation

Mandatory
Optional

Data
collection
PIDs



Reliquary

PID, ID_NAME/ID_DESC, UR(L/I/N), COMMENT, RELIQUARY_CREATOR, R_CREATOR_TYPE
PID, ID_NAME/ID_DESC, UR(L/I/N), COMMENT, RELIQUARY_CREATOR, R_CREATOR_TYPE
PID, ID_NAME/ID_DESC, UR(L/I/N), COMMENT, RELIQUARY_CREATOR, R_CREATOR_TYPE
... one row for each PID ...



Each PID points to a landing Page

- Includes link to data
- Link may be brittle if data version is updated (pointer to new version?)
 - Reproducibility is important, current focus is transparency though

DOIs versus other PIDs for identifying granules

DOI

- Metadata rich
- Require significant human interaction to mint (abstracts etc, is this level of interaction sustainable?)
- Strict constraints on (meta)data updates
- Infrastructure cost - 10 cents per DOI

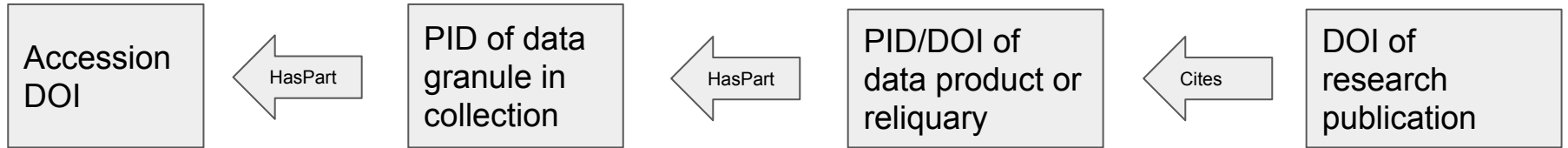
Other PIDs (e.g. Handles, EPIC IDs)

- Light metadata - Can readily automate production
- More sustainable for small granules that are part of broader DOIs
- Infrastructure cost - 1,200 Euros for 100,000 PIDs per year (1.2 cents per PID)
- Need to define minimum metadata model to enable integrity checks on PIDs

Other use case this would facilitate

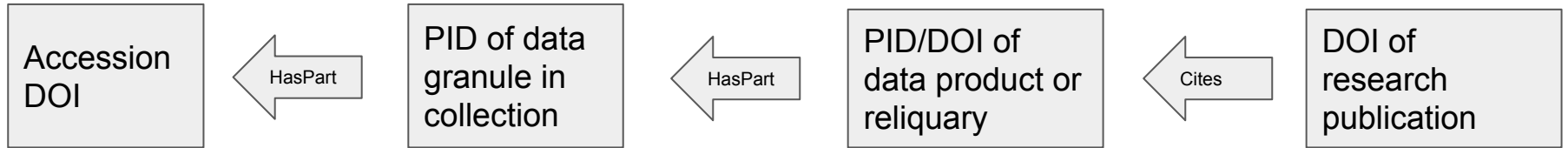
- High volume image data from cameras
 - DOI on deployment, PIDs on each image, 1,000,000+ images per camera deployment
 - Will underpin reproducible image processing workflows
 - First high volume BODC DOI (numerical model data rather than image data)
 - <https://catalogue.ceda.ac.uk/uuid/2e982e6692e3427dbe35e64ad9dee12d>
 - Does not have PIDs on granules currently

Secondary goal PID graph for metrics



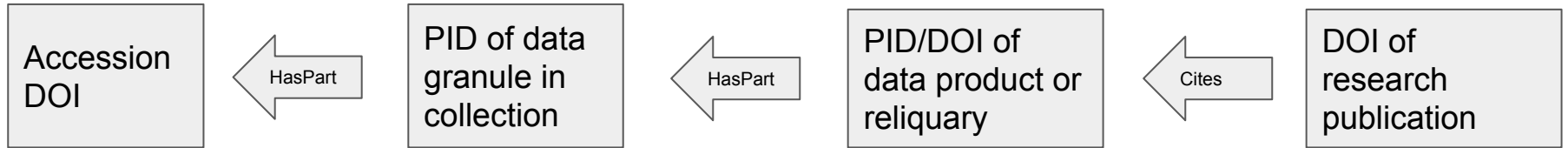
Enables data product citation or reliquary citation to be list against DOI for originators

Secondary goal PID graph for metrics



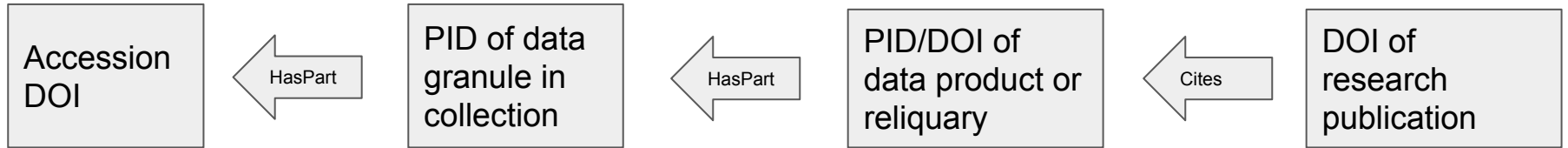
Enables data product citation or reliquary citation to be list against DOI for originators

Secondary goal PID graph for metrics



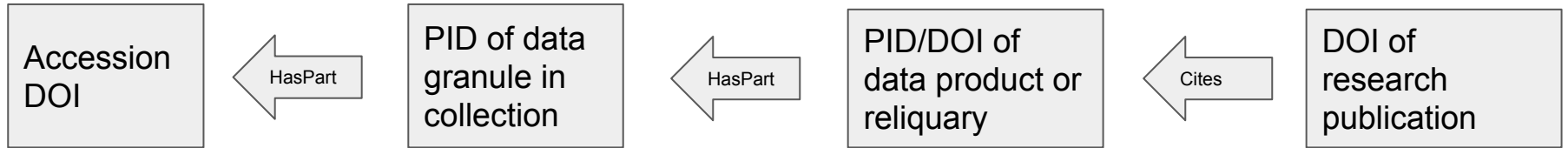
Enables data product citation or reliquary citation to be list against DOI for originators

Secondary goal PID graph for metrics



Enables data product citation or reliquary citation to be list against DOI for originators

Secondary goal PID graph for metrics



Enables data product citation or reliquary citation to be list against DOI for originators