Developments in the University of Auckland's Instrument Data Service

The move towards a production ready service.

Chris Seal, YongJe Kwon, Noel Zeng, Libby Li, Mike Laverick and Yvette Wharton

Outline

1. Context
   - UoA Research Data Management programme
   - Separation of data and metadata streams
2. Modifications - implemented
   - Previously discussed
   - Identifiers
   - Data classifications
   - New Instrument Data Wizard
3. Modifications - development/testing
   - Project-specific storage and tiering of data
   - RO-Crate
   - Globus integration
4. Changes planned

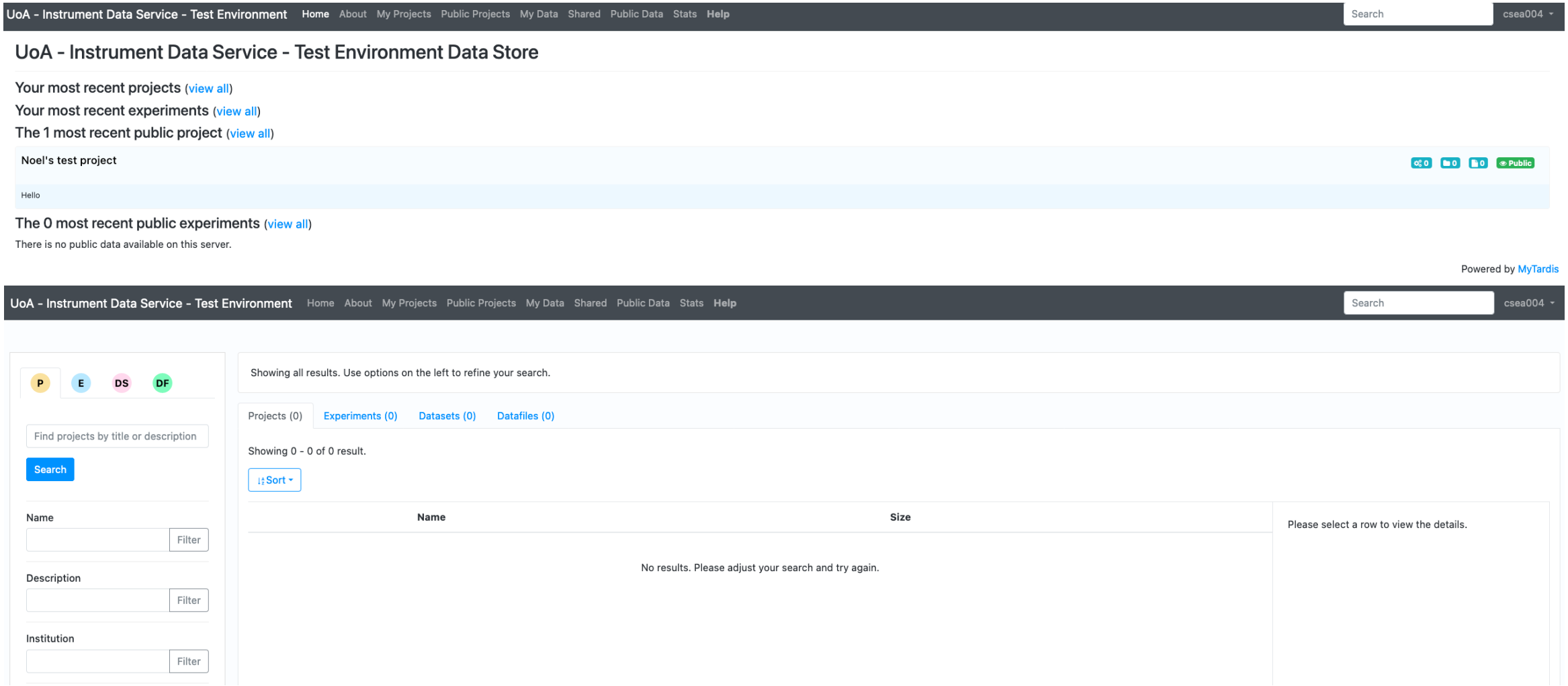
RDM initiatives

1. Secure Research Environment (SRE)
   - Enables secure upload, storage, processing and analysis
   - Supports robust governance and management
2. A machine-actionable Data Management Planning solution (maDMP)
   - Connected tool that facilitates dynamic data management planning
3. Persistent Identifiers (PIDs)
   - connecting our research ecosystem
   - A linked research ecosystem enabling information flow between university systems.

Separation of data and metadata flow

1. Restricted access to some local computers
   - No single approach to collecting metadata
   - Precludes use of MyData
2. Data generated is often large and predicted to increase in size
   - Favours asynchronous approaches to data movement
   - Future-proofing
3. Standard data movement tools can be used
   - Rsync
   - Globus
   - SCP/SFTP

MyTardis at UoA



PIDs

1. Identifiers for any object in MyTardis
   - Enforced uniqueness within an object
   - Multiple identifiers
   - Acts as a filter to locate an object in MyTardis

2. RDM Initiative 3 (PIDs) - integrates

Data Classifications

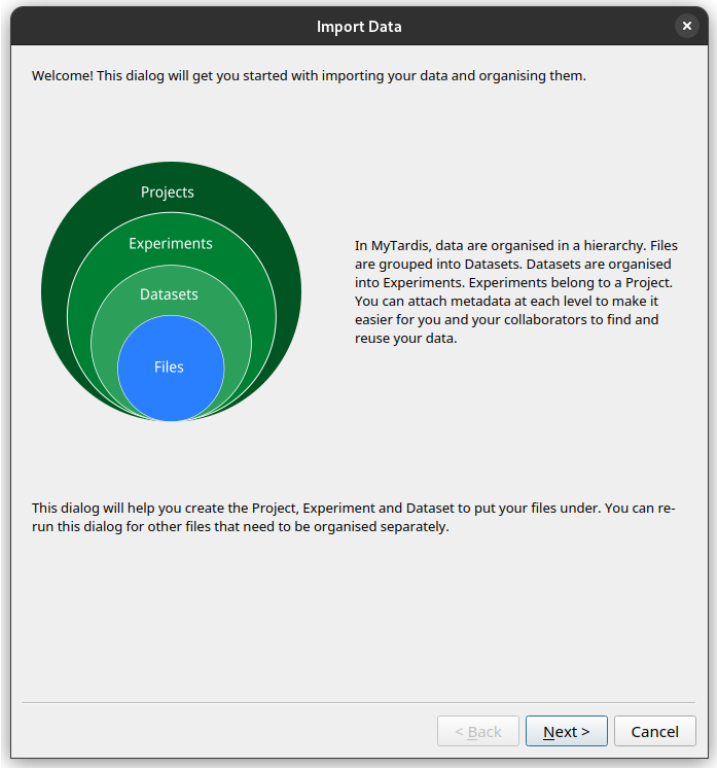
1. UoA Research data security classification
2. Integration into MyTardis
   - Hook for future actions – such as restricting features for Sensitive and Restricted data
3. Sensitive = default
4. Cascades from parent object
   - Project -> Experiment -> Dataset
   - Can be altered on an object-by-object basis
5. RDM programme links
   - Initiative 2 (ma-DMP)
   - Initiative 1 (SRE)

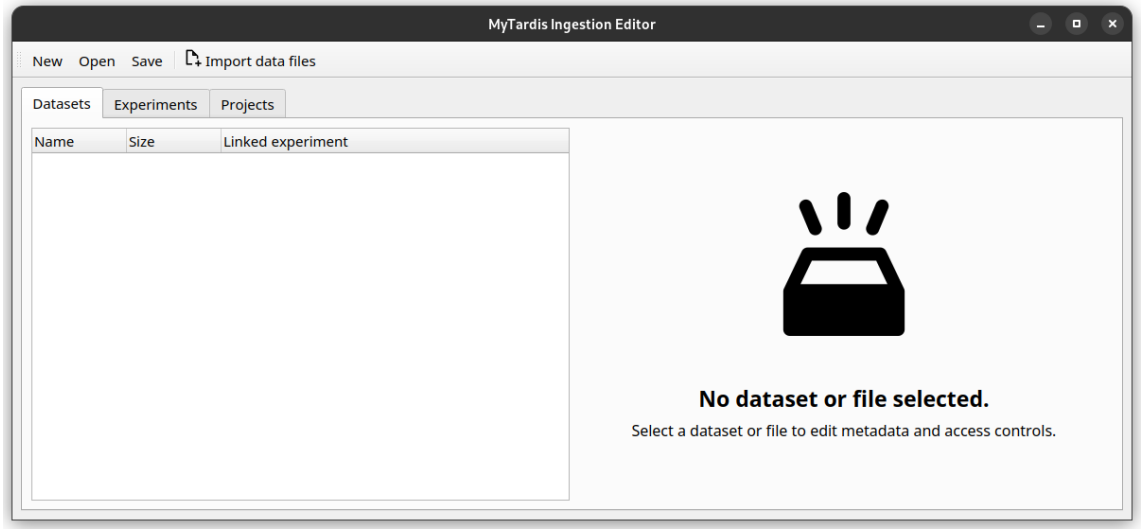
Research data security classification

Classification	Description	Research data example
Public	These data are public and do not have a restricted audience. Disclosure of these data to an unauthorised party is not likely to adversely affect the interest/reputation of the University of Auckland or the privacy of any natural persons.	Published research data.
Internal	These data have a restricted audience. Disclosure of these data to an unauthorised party would likely impede the effective operation of the University of Auckland, adversely affect the privacy of natural persons or could otherwise potentially disadvantage the University of Auckland.	Preliminary research data that are intended for publication at a later stage. Data that are subject to an expedited ethics approval for a low-risk application.
Sensitive	These data have a highly restricted audience. Disclosure of these data to an unauthorised party would be likely to cause serious damage to the interest / reputation of the University of Auckland or endanger the safety of any natural persons.	Sensitive research data includes data that: are commercially sensitive, including data classified as confidential information in commercial research and consulting contracts are or may be the subject of a patent application or other application for intellectual property protection are subject to the New Zealand government export control regime are subject to full human ethics approval processes (excluding expedited reviews for low risk applications) are subject to animal ethics approval processes are otherwise subject to a dual use / sensitive technology risk rating by MBIE or other funder.
Restricted	Access to these data is limited to specific individuals approved by the University Custodian (or nominee). Disclosure of these data to an unauthorised party would likely cause severe harm to the University of Auckland and adversely affect the national interest. Note: This classification of data as Restricted Data is made by the University Custodian.	Restricted government and/or commercial data held at the University as part of a sensitive research programme.

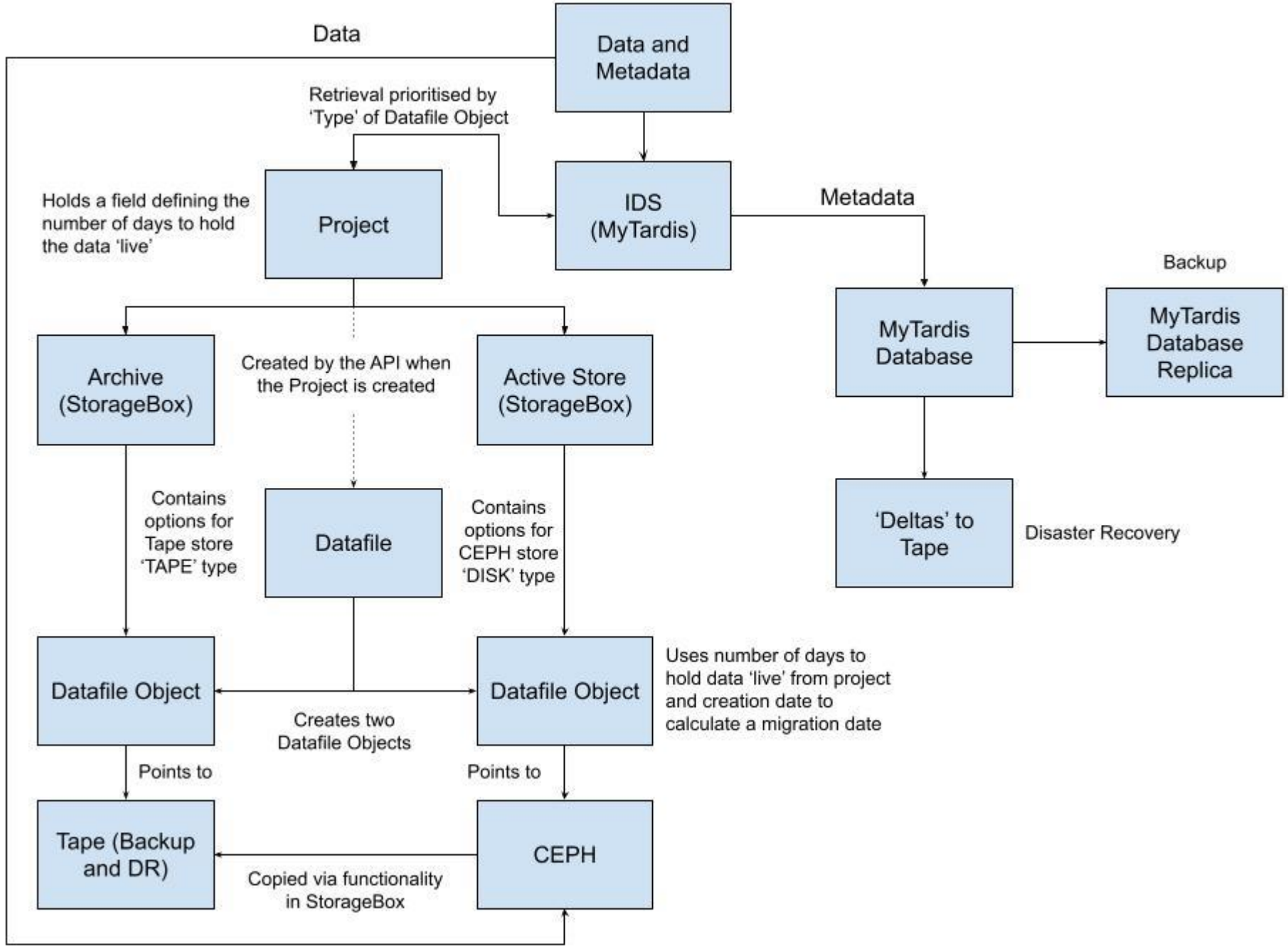
The Instrument Data Wizard

1. Guided way for researchers to add metadata to the ingestion pipeline
   - YAML templates were originally provided
2. Data is ingested from University storage that researchers can access
   - IDW generates YAML file, which is stored alongside data for ingestion
   - Allows researchers time to add metadata outside of instrument usage time

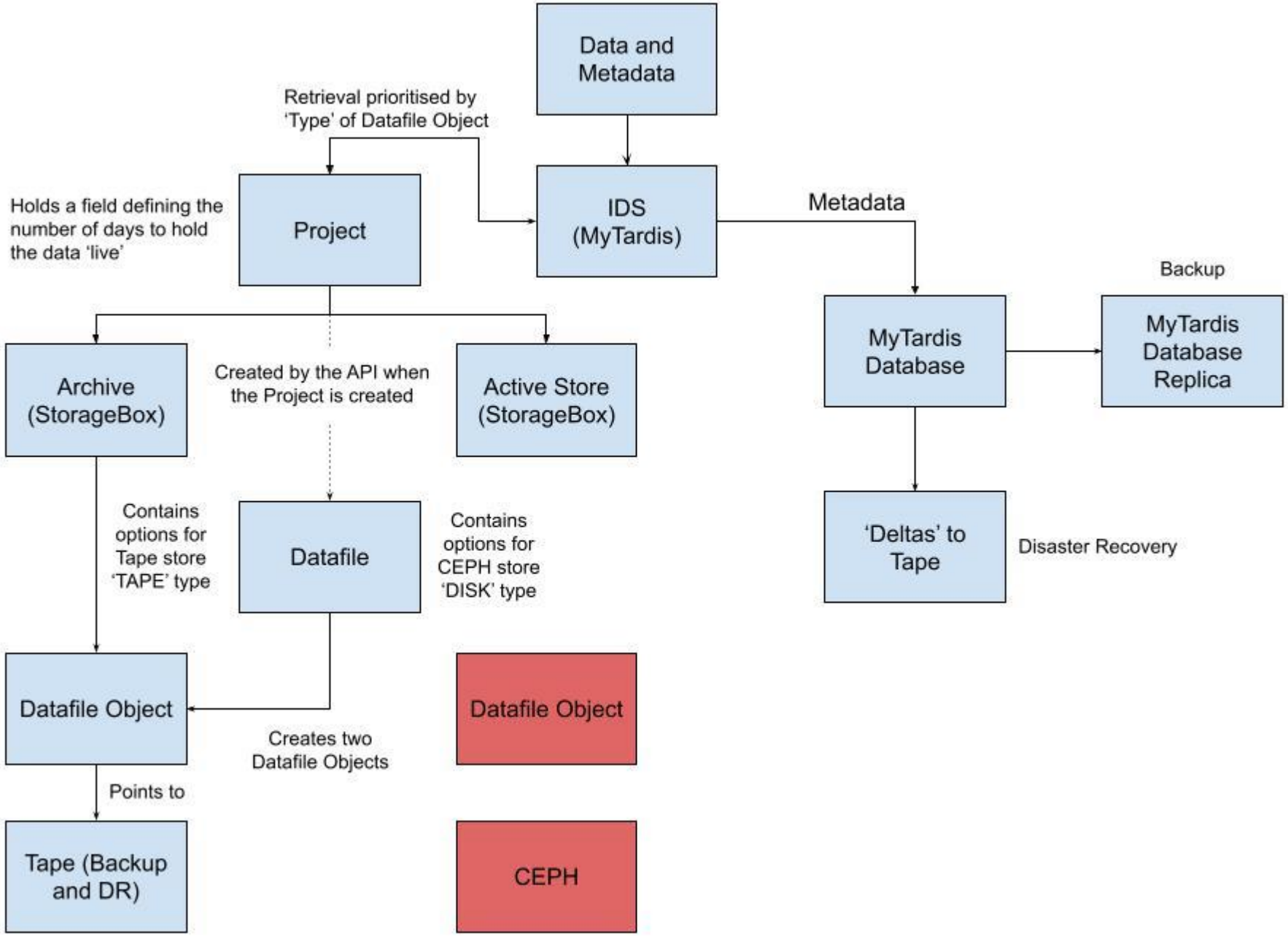




Project-specific storage

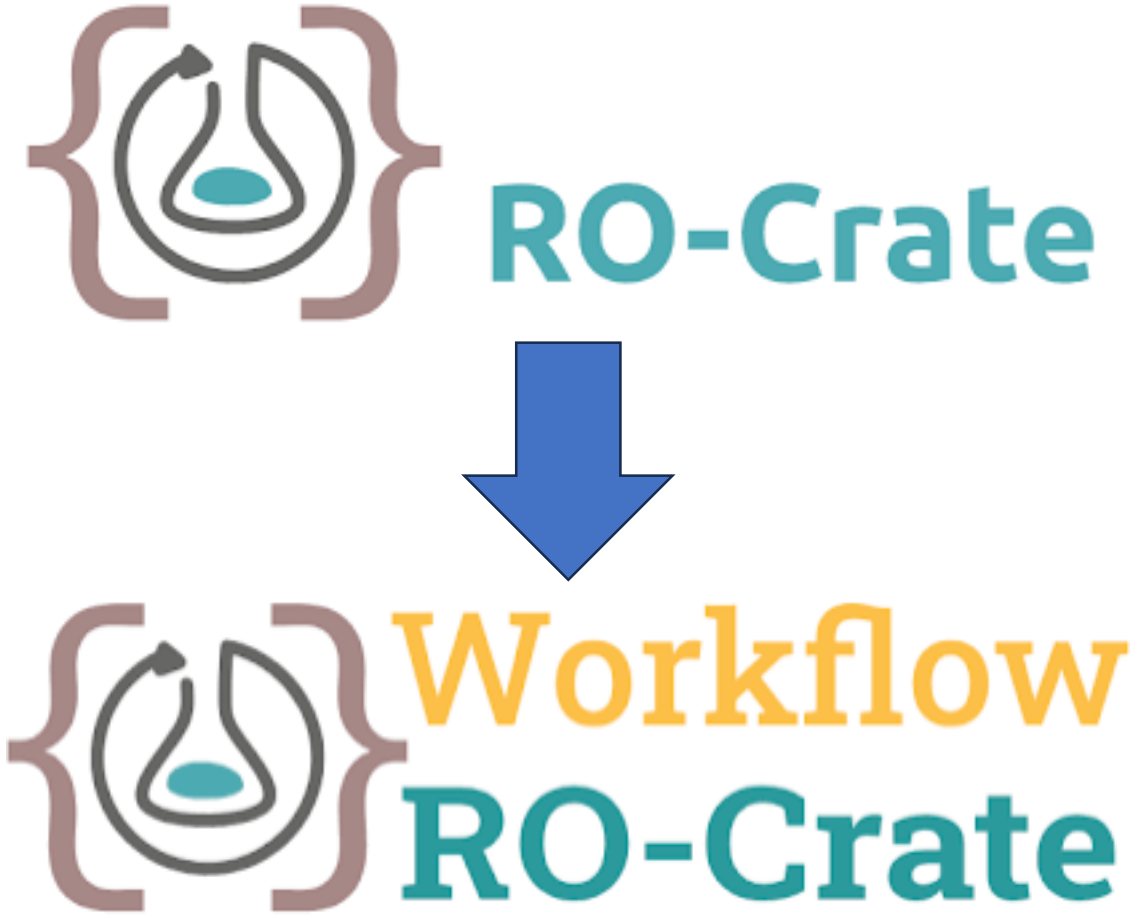


Project-specific storage



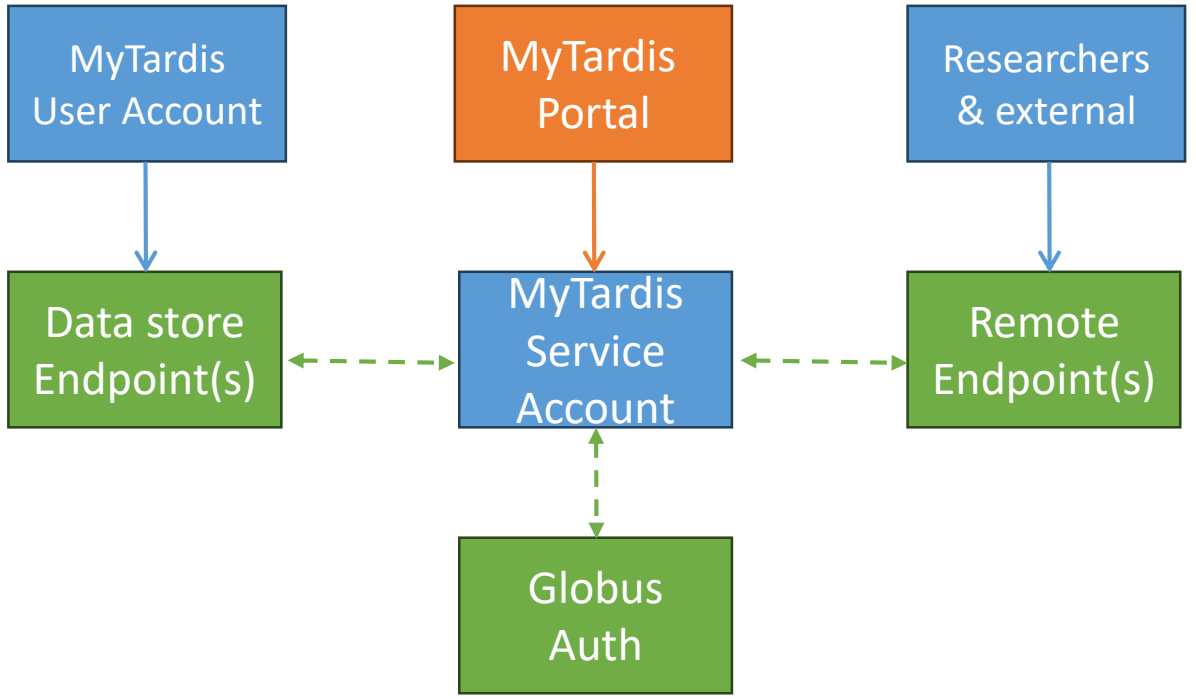
Incorporation of RO-Crate

1. Create RO-Crate JSON-LD during ingestion. Store with datasets
   - Provides additional metadata resilience
   - RO-Crate is planned for use in archiving data
   - Underpins future plan for developing integrated and automated workflows
2. How to handle sensitive metadata?
3. MyTardis group-centric, RO-Crate user-centric



Globus integration

1. Integration into ‘in-development’ shopping cart
2. Project contains known endpoints
   - Likely to interact with data classification
   - Process for adding new endpoints needs service development
3. Second phase development -using Globus for ingestion
   - Leverage platform for large data transfer

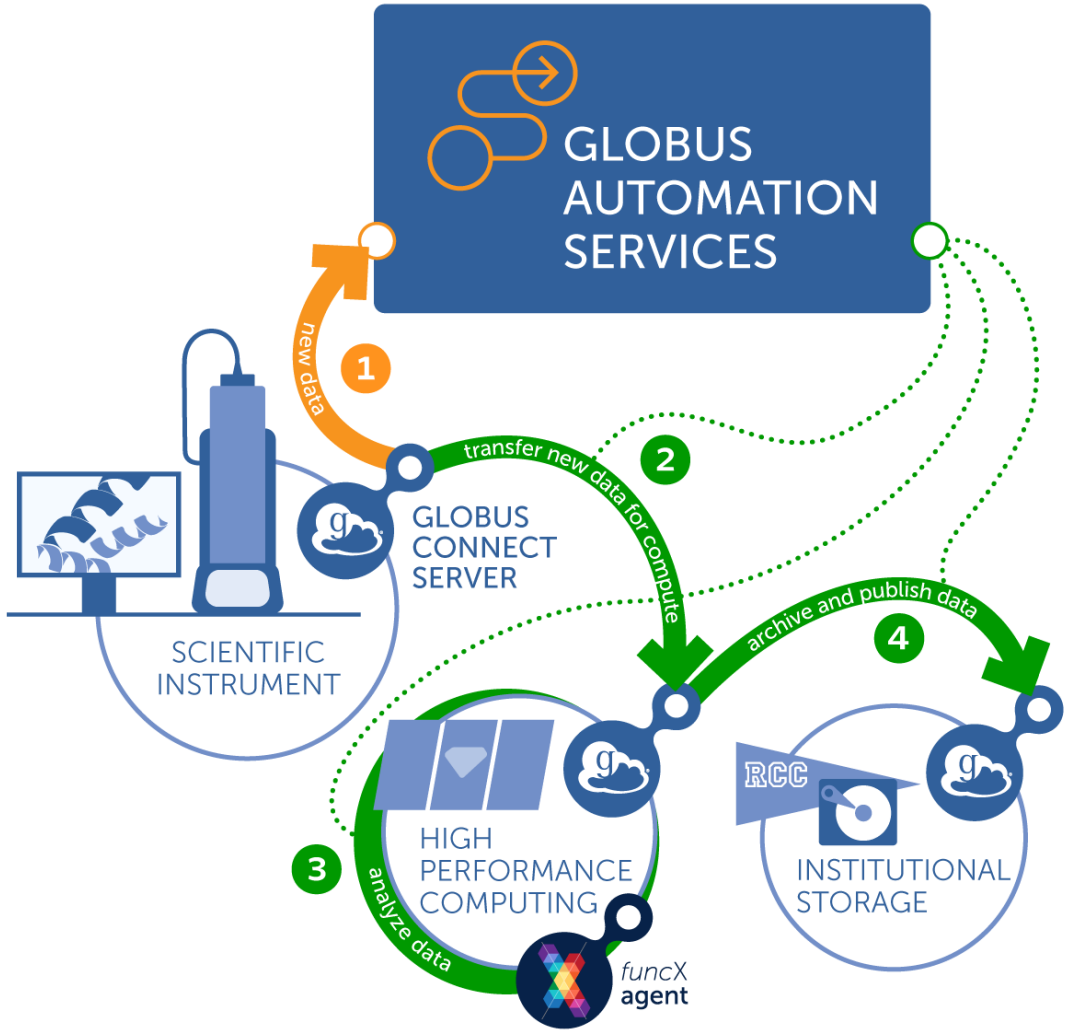


Near future – server-side validation

1. Current ingestion process 'brittle'
   - Parent objects need to be created before an object
   - Needs client-side object collision checking – fails if there is a collision detected
   - Scripts can't create an 'incomplete' object i.e. we need to collect the minimum required metadata at time of ingestion
2. Moving server-side - more robust process
   - Server-side staging objects can be used to hold incomplete objects and prompt for user intervention
   - Object collision can be determined without preventing the ingestion
   - Child objects can remain in staging area until parents have been created
3. Need to associate the data with a person or group in MyTardis
   - Can use instrument-facility relationship to associate the data to facility managers in the absence of user/group ownership

Extended future – integrating workflows

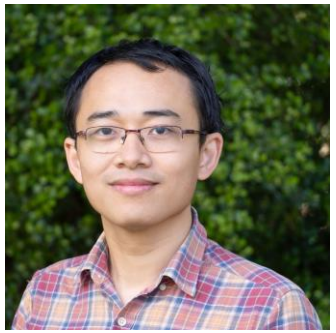
1. Uses RO-Crate + workflow engine
   - MyTardis as researcher's instrument data *home-pa*
2. Gladier leverages Globus development
3. Service design
   - 'Trusted' workflow definition
   - How do we associate workflow with project?
   - Safeguards to ensure workflow does not exceed compute quota?



The rest of the IDS Team at UoA

Yvette Wharton

Libby Li



Noel Zeng

Mike Laverick

Andrew Wilson