INTRODUCING HELIX: IMPERIAL COLLEGE LONDON'S NEW FAIR DATA REPOSITORY

STEP-UP RSLondon 2025

Christopher Cave-Ayland & Wayne Peters

CIRCA 2021

The College provides an institutional repository service - https://data.hpc.imperial.ac.uk/

It has some pros:

- Innovative approach to data collections and domain specific metadata.

And cons:

- Developed as a partnership between Matt Harvey (previous head of RCS) and Prof Henry Rzepa (Chemistry).
- Arguably over tuned to the needs of specific users giving the perception that it's "for chemists".
- Lacks clear supporting policy and terms of service.
- Heavily used by a small group of users (chemists) but scarcely otherwise.
- Old tech stack.
- Lack of software engineering best practices during development.

PILOT WORK

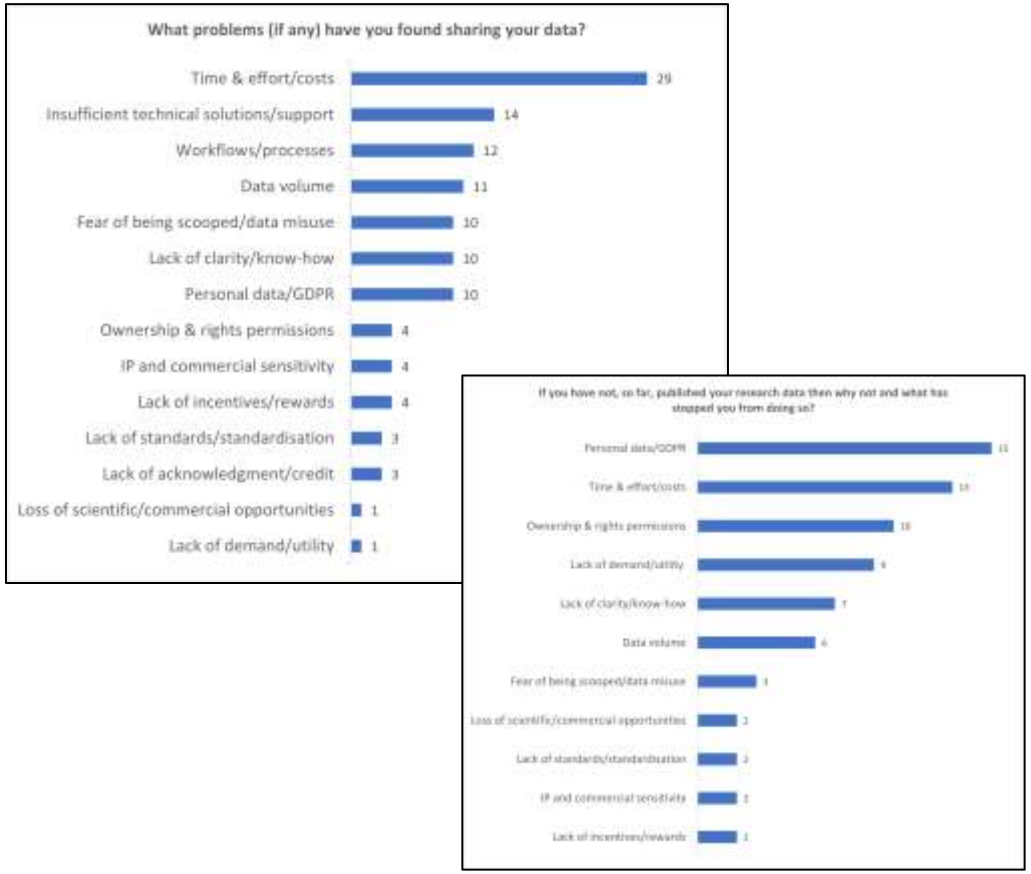
- Funded through UKRI Enhancing Research Culture stream.
- Led by Mike Bearpark and Henry Rzepa from Chemistry.
- Development work carried out by me focused on exploring developing a modern data repository application built on the Invenio framework (not InvenioRDM).
- Whilst some code was developed the most important outcome was development of technical expertise and refining a model for supporting domain specific metadata.

FAIR DATA WORKING GROUP

- Created circa February 2022
- Tasked with revisiting requirements for an institutional data repository in conjunction the RCS FAIR research data strategy
- Stakeholders include RCS/ICT, Faculties, Library Services, Research Office, Archives, and the Data Protection Officer
- Terms of reference included making the case for an institutional data repository ‘based on requirements identified through community consultation’
- Survey of funded PIs conducted in April 2022. The survey had 205 respondents, approximately 10% of University PIs

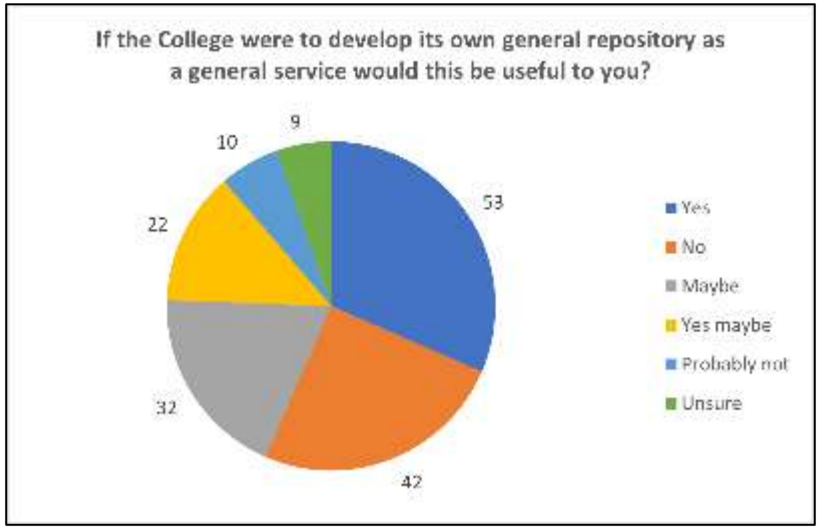
BARRIERS TO DATA SHARING

- Time and effort needed to prepare data for deposit
- Restrictions due to sensitive/confidential data (e.g., personal data, IP, rights permissions)
- Lack of infrastructure/technical support (incl. large data)
- Lack of clarity/know-how
- Fear of data misuse/misinterpretation
- Fear of being scooped/insufficient credit
- Lack of demand/utility within research community



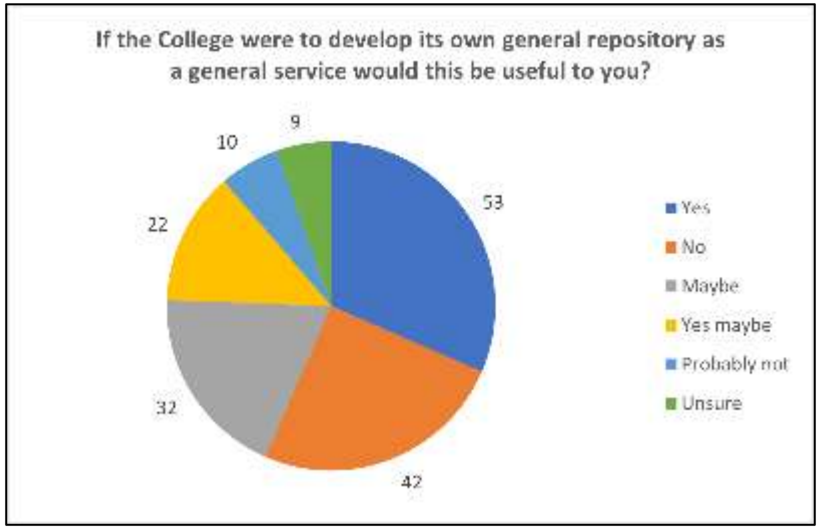
SUPPORT FOR A DATA REPOSITORY

- 52% participants said they would consider using a locally hosted research data repository
- 60% of participants who have previously used a community data repository said they would use - or would consider using - a college data repository
- 47% of participants who have previously used a general data repository said they would use - or would consider using - a college data repository



SUPPORT FOR A DATA REPOSITORY

- Many participants argued that an institutional repository should
  - provide a level of functionality and user experience that is at least equivalent to that provided by existing general repositories such as Zenodo and/or
  - offer types of support not always available from general repositories



ZENODO PLUS

A research data repository for researchers not well served by other domain specific or general-purpose repositories (e.g. large data/sensitive datasets)

This could be described as ‘Zenodo Plus’… and adds value through support and curation

Three-level data service model: Minimum, Medium, Maximum

RDM SERVICE LEVELS: MINIMUM

- The repository is not able to accept datasets that contain personal data or sensitive information
- The repository makes datasets and metadata freely available but with an embargo where required
- There are limits to data that can be uploaded - currently 10 GB for individual files and all files in a submission
- Consistent documentation and guidance are provided by the library/research office/ICT

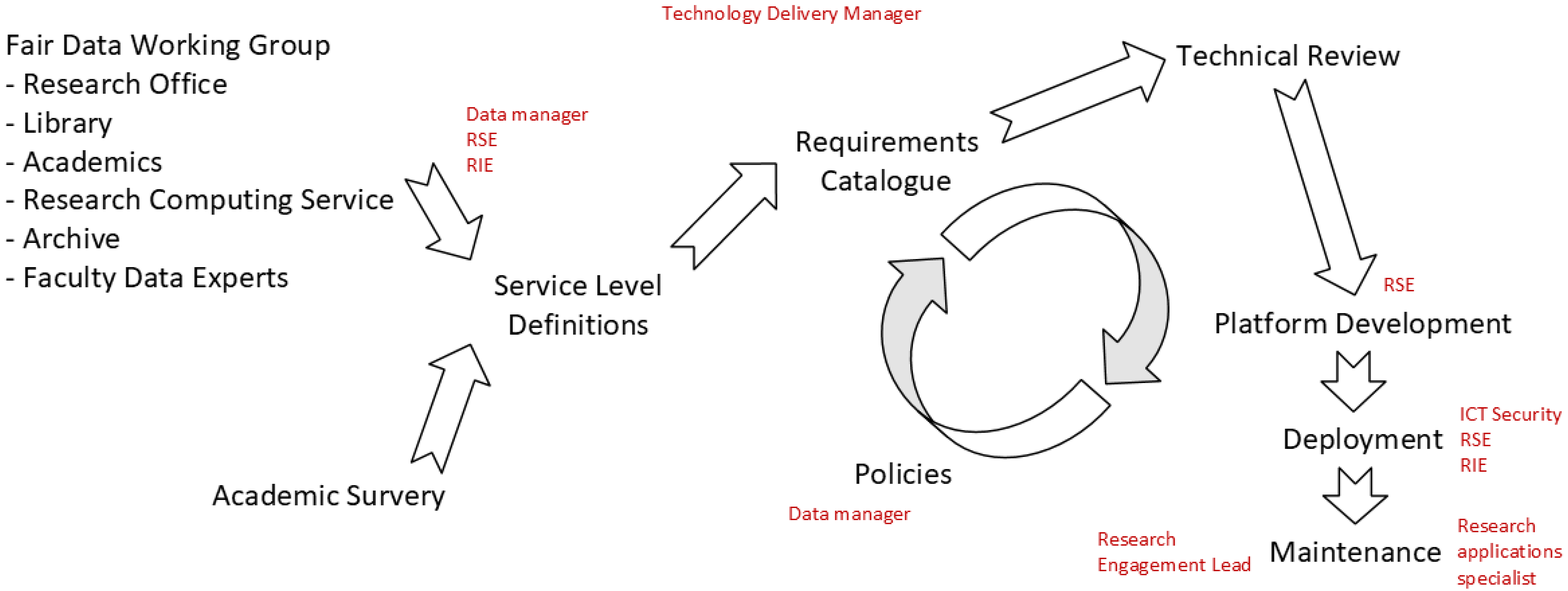
RDM SERVICE LEVELS: MEDIUM

- The repository accepts low-risk sensitive data providing appropriate permissions are in place
- The repository provides managed access to low-risk sensitive data
- The repository provides alternative methods of upload/transfer for datasets/files that exceed stipulated size limits
- There's basic integration with a research data archive and Trusted Research Environment (TRE)

RDM SERVICE LEVELS: MAXIMUM

- The repository can facilitate access to moderate or high-risk sensitive data
- All datasets receive full digital preservation and curation support and there is integration with a digital preservation system (e.g. Archivematica)
- The repository can provide enhanced support for FAIR data / metadata at a disciplinary specific level
- Faculty data stewards/data champions, together with dedicated staff time across library, research office, ICT

A LONG JOURNEY



Fair Data Working Group

- Research Office
- Library
- Academics
- Research Computing Service
- Archive
- Faculty Data Experts

Data manager
RSE
RIE

Service Level Definitions

Academic Survery

Technology Delivery Manager

Requirements Catalogue

Policies

Data manager

Technical Review

RSE

Platform Development

Deployment

ICT Security
RSE
RIE

Research Engagement Lead

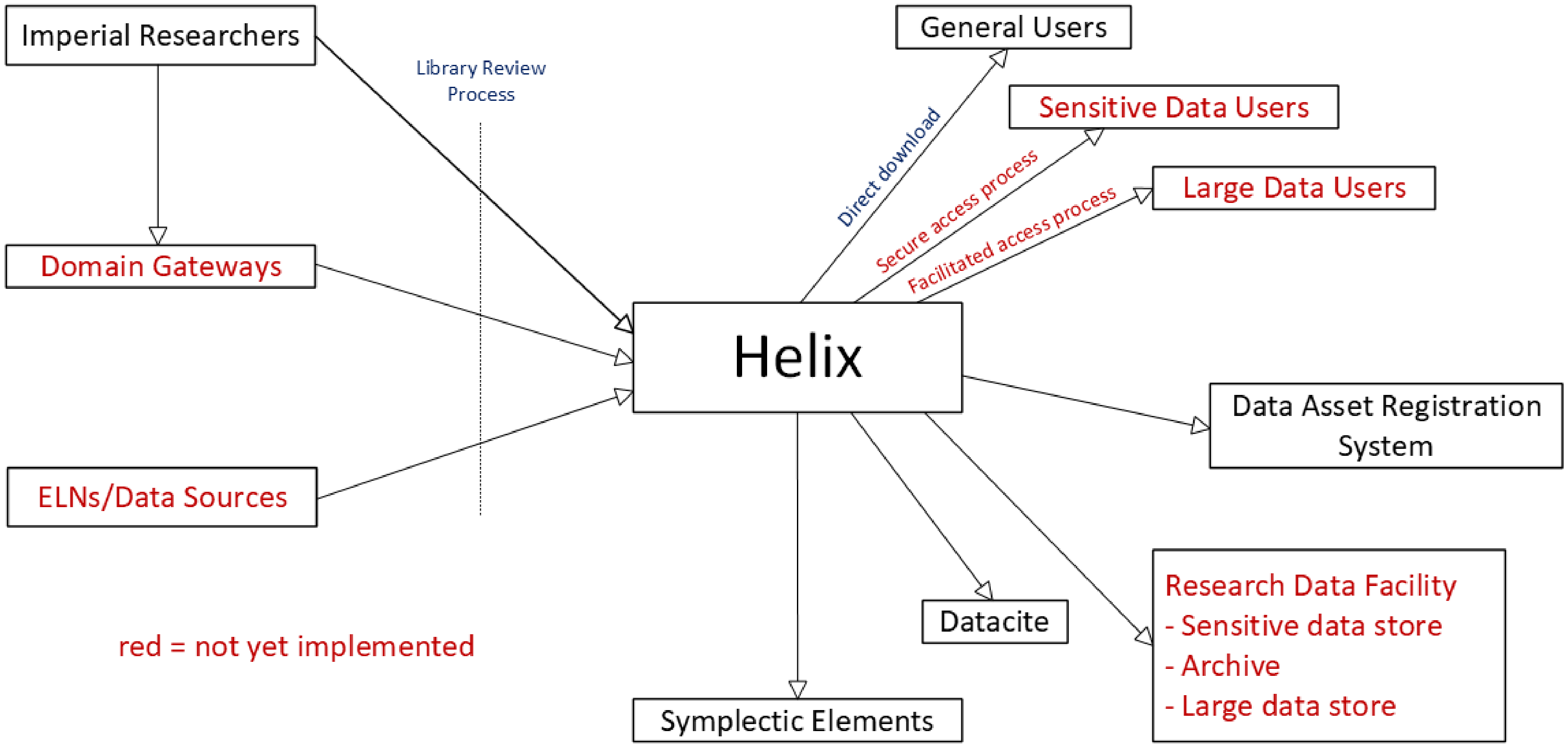
Maintenance

Research applications specialist

A NEW INSTITUTIONAL REPOSITORY SERVICE (HELIX)

- Key strategic goals:
  - A robust, sustainable service for the College.
  - Alignment with the College's research data management policies and FAIR practices.
  - Curated.
  - Support for large (multi-TB) datasets.
  - Support for sensitive datasets.
  - Support domain specific metadata and ingestion routes.
- Built on InvenioRDM.
- Close to Beta launch - late July.

THE PLAN



red = not yet implemented

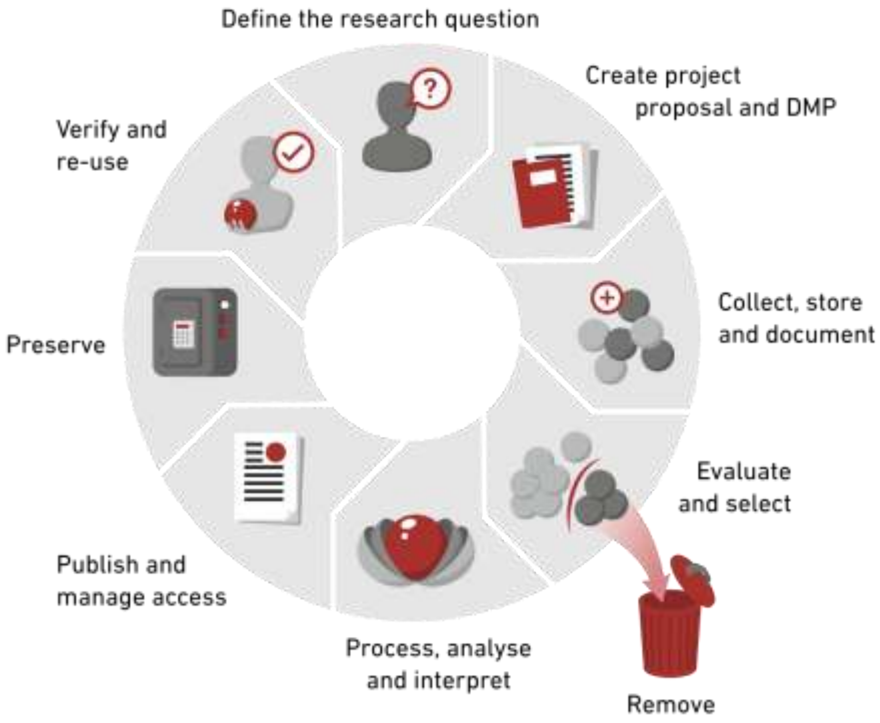
INVENIORDM OVERVIEW

- Proven stack – underpins Zenodo.
- Active development and community supported by Cern.
- Number of existing institutional deployments.
- Datacite based metadata model.
- Customisable – we've provided external integrations, updated the data model and user interface.
- Suitable for modern scalable deployment (Azure Kubernetes Service).

INVENIORDM CAVEATS

- Technically complex.
- Convoluted project structure with many individual packages.
- Sparse documentation.
- More hacky modifications than ideal were required.
- Unit testing was a particular pain point.
- Overly abstracted code base.

CONCEPTUALISING DATA USAGE


Define the research question
Create project proposal and DMP
Collect, store and document
Evaluate and select
Remove
Process, analyse and interpret
Publish and manage access
Preserve
Verify and re-use

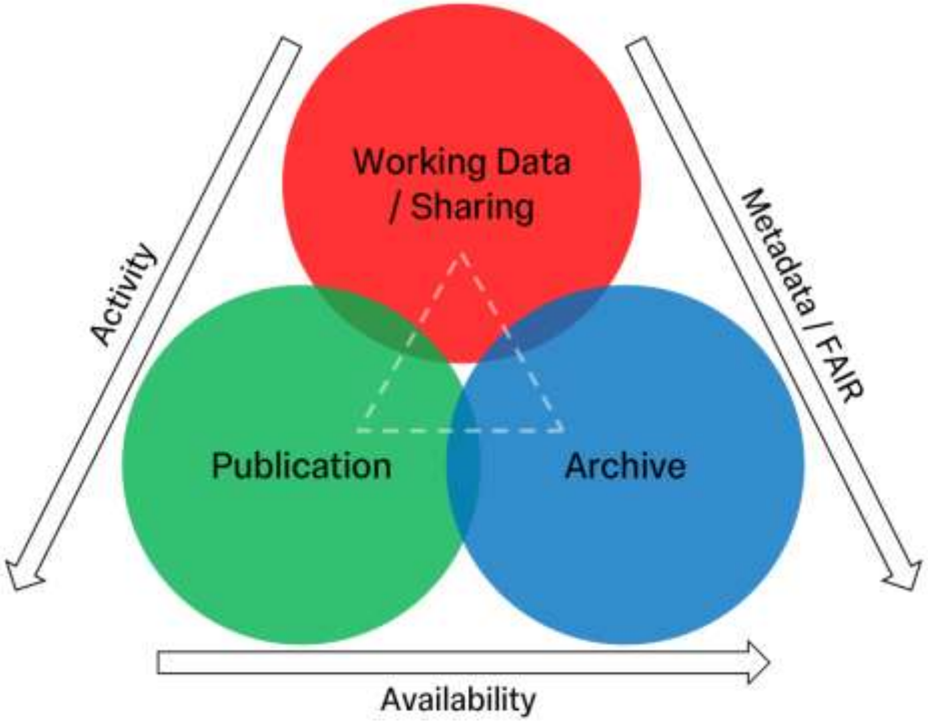
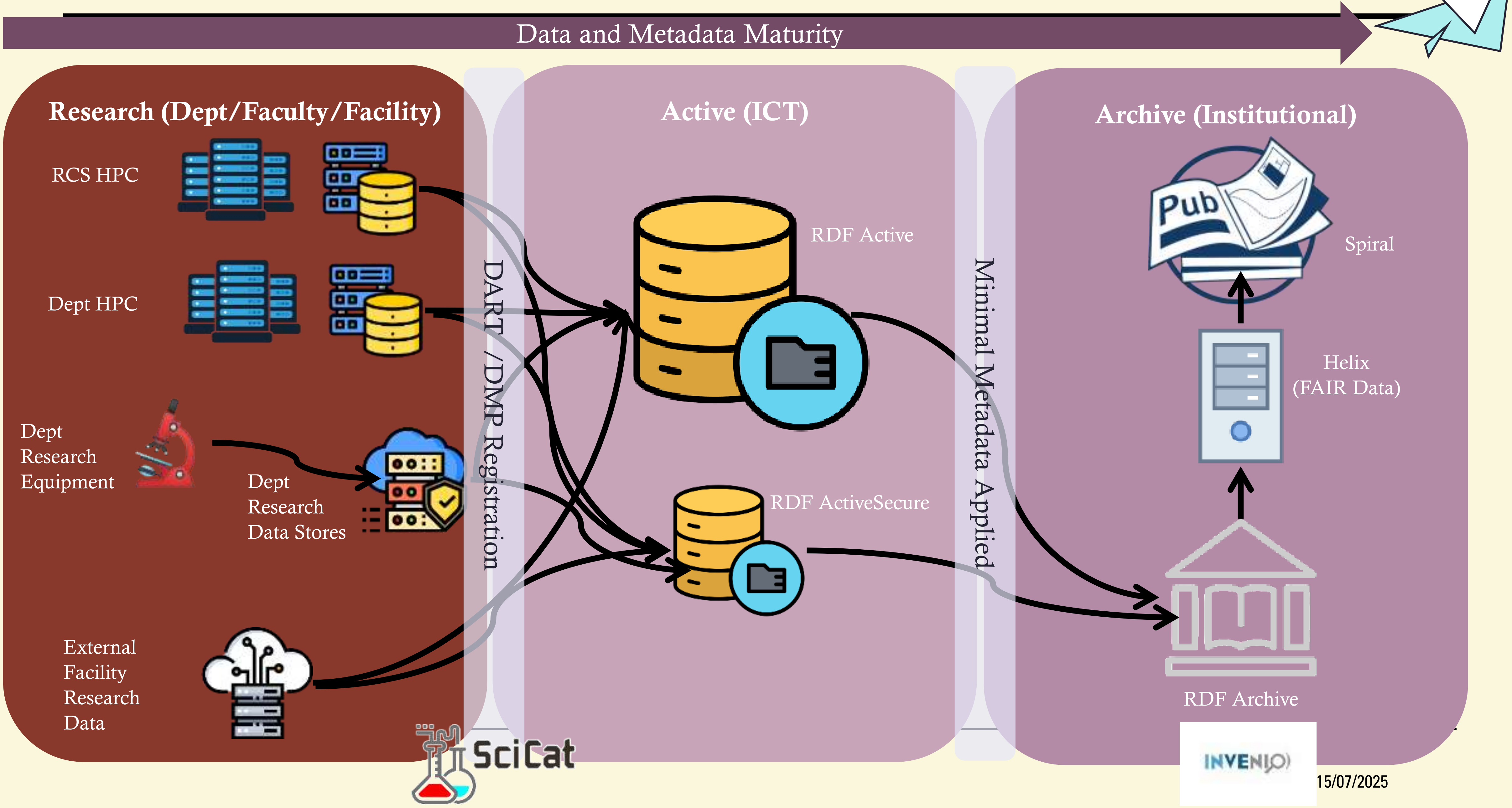


Image Credit: Prof. Mike Bearpark

IMPERIAL DATA FACILITY



15/07/2025

CHALLENGES

- Infrastructure - lack of central storage/archiving solutions for sensitive/large data
- Sensitive data governance – establishing effective workflows/processes for data submission, review and access requests
- FAIR metadata e.g.

  To support FAIR, published metadata must be

  - machine-actionable
  - align with domain-relevant community standards and practices

Draft UKRI research data policy (April 2025)