Research Data Priority Paper

Paul Aylin & Dave Collings
Co-Directors of Research Data Strategy

Action Requested

- **An Imperial Secure Research Data Environment should be developed that allows for the storage and processing of different levels of sensitive data. The project group should aim to deliver a basic ISO certified service within 6 months that allows ingestion of smaller research projects to help mitigate the risks of data breaches these projects currently incur.**
- **The College needs to commit resources to provide Information Governance oversight and operational capability to support the design and running of the new environment.**
- **The college run repository needs to be a collaboration between the academic community, the Library and the RCS and should better reflect the principles of findability, accessibility, interoperability, and reusability. The scope and functionality of this repository needs to defined and a realistic time scale on which to provide it identified.**

Executive Summary

The capability to manage, store and analyse increasingly large volumes of data is vital to the success of data-intensive research, particularly in areas such as medical research, biomedical imaging, particle physics, applications of AI, machine learning and big data. A strong Research Computing Service would underpin bids for big data research projects, facilitate research collaboration between faculties, other institutions, the NHS and commercial partners.

A recent audit of Imperial research projects suggest potentially thousands of sensitive databases requiring a secure environment. Without such an environment, the college is leaving itself open to the risk of data breaches and a consequent serious loss of trust of the public and research funding bodies, as well as potentially large fines from the ICO.

Non-sensitive data projects are better served by Research Data Services, and the High Performance Cluster but there is still scope for a more universally appealing environment to get away from a batch computing modality, which is a poor fit for researchers who need to work more interactively with data.

There is an increasing obligation for all data generated through publicly-funded research to be managed under the FAIR principles of findability, accessibility, interoperability, and reusability. Imperial needs to provide a repository for data generated in college that better supports these principles.

Background

The capability to manage and analyse increasingly large volumes of data is vital to the success of data-intensive research, particularly in research areas such as biomedical imaging, particle physics, applications of AI, machine learning and big data. A strong Research Computing Service would underpin bids for big data research projects, facilitate research collaboration between faculties, other institutions, the NHS and commercial partners.

This document is based on the output of a working group, led by the Vice-Provost (Research & Innovation), established to understand and define the academic user requirements of the Research Computing Service and help agree future levels of service from ICT.

A good research data strategy should support the data lifecycle (see below in fig 1). This strategy document focuses on the central capacity to store and manage data. We also recognise that there are existing systems and expertise which can help to inform a more coherent approach to the data lifecycle within Imperial.

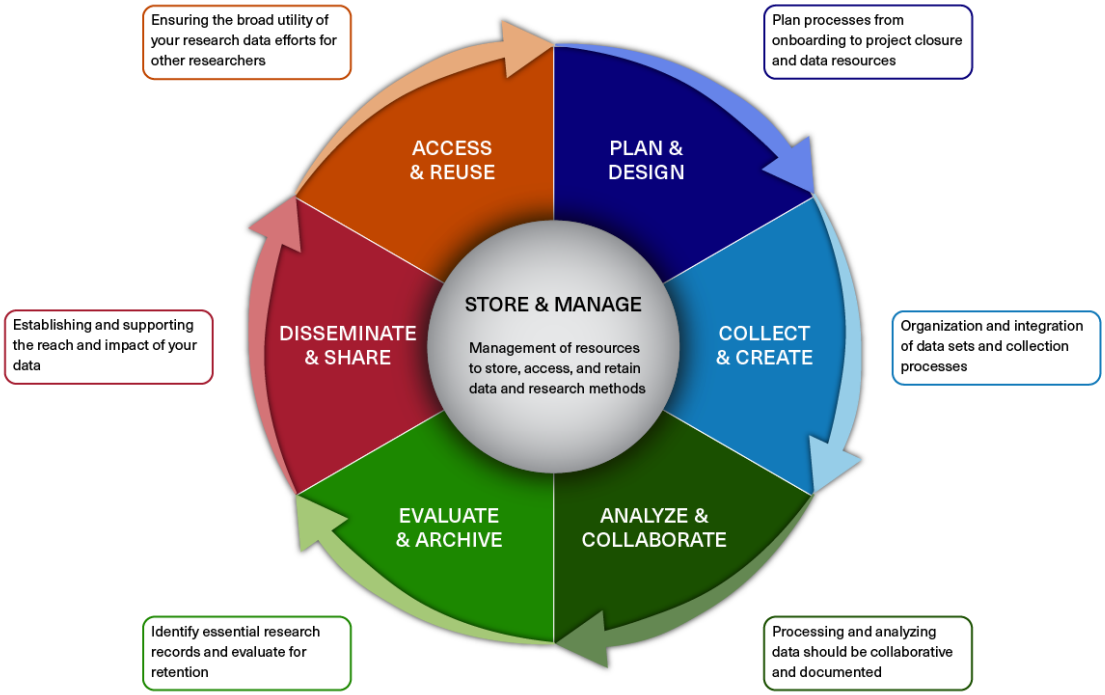

Fig 1: Longwood Medical Area Research Data Management Working Group (LMA RDMWG)

We acknowledge that different types of data may require different systems and approaches. In this document, we consider two crude categories of data, **sensitive** and **non-sensitive** data as well as the overarching activity of **data management.**

The purpose of this document is to set out broad recommendations. The work needs to be driven and iterated by researchers, but it is also recognised that it should fit within a larger framework of a complete College data strategy.

1. Sensitive data

Data can be sensitive for a number of reasons – the most common at Imperial College being because the data are personally identifiable (e.g. medical data) or are commercially confidential. Different levels of Information Governance (IG) are likely to be required for different categories of data.

Having secure data storage is useless unless there are secure mechanisms for analysing these data. Several researchers use the RCS service to work with commercially sensitive data, where the data provider has performed their own due diligence. However, given the multiuser nature of the system, it's likely that the platform will never be appropriate for processing non-anonymised data subject to regulatory controls.

There are currently individual enclaves in college that provide the required level of both security and Information Governance for the datasets that they handle, but each have strengths and weaknesses that mean in their current form, they are not in a position to offer a College-wide solution.

A recent audit of research projects within the FoM suggest potentially thousands of sensitive databases associated with these projects requiring a secure environment. Without a secure environment to support these projects, the college is leaving itself open to the risk of data breaches and a consequent serious loss of trust of the public and research funding bodies, as well as potentially large fines from the ICO.

A review of the current College infrastructure for holding sensitive data found there was strong support for a College-wide system to securely store and process sensitive data. The College Data Protection Officer is fully supportive of a single centralised system that encourages and supports information governance. Key requirements included:

- Strong security and information governance
- Access to high-performance computing within a secure environment
- Ability to share data in a suitably controlled environment across the College and with external collaborators
- Flexibility to accommodate a wide range of users
- Scalability due to increasing demands on storage and processing power. The "super-users" of today will be the standard tomorrow, and this might best be delivered through a cloud based system.
- A tiered service model should be considered which is free to basic users, to encourage migration of high risk small projects into the system. Going forwards, a funding model needs to be established to sustain the service.

A project group has already been established, with members drawn from existing research facilities who handle sensitive data and members of the RCS to create a service that can be used for nearly all sensitive datasets in college. This group is chaired by Professor Paul Aylin.

1.1 Recommendations

- A central Imperial Secure Research Data Environment should be developed that allows for the storage and processing of different levels of sensitive data. The project group should aim to deliver a basic ISO certified service within 6 months that allows ingestion of smaller research projects to help mitigate the risks of data breaches these projects currently incur. At 12 months, the Imperial Solution should aim to comply with NHS Digital DSP (Data Security & Protection) Toolkit compliance (to support processing of NHS data), and larger projects should then be migrated into the environment.
- In setting up such a system, we recognise that both technical and Information Governance infrastructure are equally important. While the RCS can provide technical expertise and resource, the College will need to commit resources to provide oversight and to operate effective Information Governance for the environment.

2. Non-sensitive data

Large amounts of data stored and analysed at Imperial are non-sensitive. Some of these are handled centrally by the Research Data Store (RDS) although other bespoke systems have been created by

research groups to accommodate their requirements, and other researchers use Box. These may pre-date the RDS or contain sensitive data.

Within RDS, projects are allocated a set data volume (2TB but can pay for more space. These volumes are backed up securely in different ways. The RDS is configured in such a way that it can serve data to the RCS compute resources in an efficient way.

RDS is intended for active data i.e. data that are under constant analysis. There is an archive service for inactive data which costs £100/TB/decade (how can this be paid for at the end of a grant period under the research council rules?). These data are slower to access as they need to be brought back from cold storage before use. The RDS is well connected to the data transfer service that currently uses Globus Connect to transfer data between sites.

A document on the RDS website provides a useful comparison between storage options.

The Research Computing Service, provides bulk compute processing capability to the college's research community. The current compute service is presented as multi-user batch processing environment, supporting a diverse set of high-throughput and high-end workloads. Work undertaken in the last few years has substantially modernised this compute infrastructure and its operations, improving its security.

The batch computing modality is a poor fit for researchers who need to work interactively with data, e.g. in a database, spreadsheet or statistical package, and is not fit for sensitive data

2.1 Recommendations

- The levels of resource provided to projects, both free and charging rates, should be reviewed annually to understand if they are at an appropriate level. The free volume of data storage provided should be enough for most projects and should be part of providing a well-found laboratory. Only specialist data projects should need to purchase additional storage. Similarly, the archival storage model and its costs should be reviewed annually.
- Groups using bespoke solutions should be encouraged to use the RDS instead and new bespoke solutions should be avoided wherever possible. However, this policy must not be allowed to hinder science output. There should be consultation with researchers not using the RCS, to ascertain what a more universally appealing environment might look like.

3. Research Data Management

There is an increasing obligation for all data generated through publicly-funded research to be managed under the FAIR principles of findability, accessibility, interoperability, and reusability. While many communities have specific national and international repositories that are suitable for this purpose, Imperial needs to provide a repository for data generated in college which does not belong in any of these. Imperial does have such a repository. This was originally a prototype produced by the Chemistry Department but which has been made more general to form a general repository.

The current repository is very limited in scope – it has no searchable web interface; it has no way of embargoing data until a certain date; there are no ways to automatically verify the meta-data associated with each dataset; etc

The Library is responsible for the Research Data Management (RDM) policy but the resources and infrastructure are provided by the RCS. The Library advises on Research Data Plans (and planning),

provides much training on RDM that is available to researchers and the Library expertise is a vital part of the whole RDM process within the college.

3.1 Recommendations

- The college run repository needs to be a collaboration between the academic community, the Library and the RCS and needs to better reflect the FAIR principles. The scope and functionality of this repository need to defined by representatives from these three communities in a separate document and a realistic time scale on which to provide it identified. The Archives and Corporate Records Unit (ACRU) should be linked to any archive environment
- An archiving and deletion service should be implemented for sensitive data. The scope and delivery of this service including ISO and DSP compliant archive storage should be considered by the group considering sensitive data storage. This would make asking researchers to register and archive old research data much less administratively burdensome for them.