# School of Informatics

# Research Committee

# Research Data Management & Infrastructure

Ian Simpson (Deputy Director of Research)

April 2, 2019

## Background

## 1 Background

The user community served by the Informatics Computing Team has grown rapidly in recent years now comprising around 350 staff (250 teaching and research), 440 research students, 715 taught postgraduate students, 1380 undergraduates, and 290 visitors and associates (data from this years' school computing plan). The **research** and **regulatory** environment continues to evolve quickly and we need to make changes to the way in which we manage research data and infrastructure in order to achieve regulatory and funder compliance and to establish a sustainable plan for growth.

Our research compute requirements have outstripped our ability to accommodate them in terms of procurement, maintenance & support, infrastructure and physical space. This is exacerbated by both legacy processes and the absence of new policies and processes to meet new requirements.

We do not have a functioning system in the pre-award phase for linking resource bids in grant applications to CS in a way that would allow them to assess the feasibility of being able to accommodate the proposed resource were the proposal to be funded. We do not have a policy for requiring the production of a Data Management Plan for each piece of research (which has been University policy since 2011). This means that for many if not most research projects the *Data Life Cycle* has not been considered appropriately (and therefore not costed) and as a result key triggers relating to Data Management including satisfying funder requirements, GDPR compliance, Data Security, Ethics and Open Access/Publication requirements are often overlooked entirely. When projects are live they generally proceed

without need for monitoring or intervention, but when problems arise we should provide guidance to help address issues (for example with critical hardware failure, data breach, misconduct). At end-of-grant we need to establish a wash-up procedure to ensure compliance with our obligations such as data-retention and deposition of data & software as well as checking continuation status of projects to keep an accurate record of active ongoing research. Whilst this might seem overly burden-some it highlights areas where we are currently not in a position to demonstrate what we are doing at any given time. When requests come in for information from funders, regulators (such as the ICO) and other interested parties we are often in a position where we cannot respond with confidence or in detail.

This has knock-on effects that leaves the School largely *in the dark* in each of these areas which has the potential for reputational damage and impacts on our abilities to both secure funding and conduct research effectively. For example; we do not currently know:

1. usage levels of servers in the self-storage server rooms
2. what research data the School uses, creates and stores
3. at any given time what sensitive data and/or projects are live
4. what happens to data, software & hardware at the end of projects
5. what our compliance levels are with conditions of funding, GDPR/Ethics & Data Security regulations
6. usage levels of RDS systems such as *Data Store, Data Sync & Data Vault*

As a result of these knowledge gaps our ability to engage effectively in planning both internally and with central services such as the RDS and IS are severely hampered. Several one-off attempts have been made to perform things like data audits in the School, but these are incredibly time consuming, incomplete and out-of-date even whilst they are being conducted. We need to develop a process for capturing crucial information at the inception of research projects and for monitoring key way-points; pre-award, post-award and end-of-grant (for grant funded research). It should be emphasised that this should cover not just grant awards (those these should be the highest priority in the first instance), but also PhD proposals, UG4 & MInf projects and personal (unfunded) research. The level of detail and the nature of those processes will vary significantly depending on the size and scale of the project, but these ideas are already captured in the current University RDS Roadmap (see references). In light of the pressing need for us to progress in this area, several key enabling activities have been incorporated into the current School Computing Plan. Some relevant summarised details are provided for information below.

## Informatics Computing Strategic Plan 2019

Highest priority identified to be *"The production of a strategy for resourcing the compute and data intensive needs of the School"*. A working group has been formed under the Com-

puter Strategy Group that is currently collecting data to support the decision making process.

Several of the key objectives required to tackle the RDM, Ethics, GDPR and Computer Security issues facing the School are already embedded in the School Computing Plan and fed into the College Computing Strategy at the beginning of the year.

**RDM Relevant Goals from 2018/19**

1. *Produce a register of medium and high risk data and a mechanism for users to self populate the register*. Due to slow progress on the College register the decision was made to produce our own simplified system leveraging existing Theon database technology. The expectation is that data entered into this system can be migrated to any future College or University register.

2. *Consideration of how best to make use of the new central RDM services*

3. *Continued consideration of appropriate use of central data storage facilities*

4. *Produce a strategy for resourcing the School's compute and data intensive requirements*. A working group has been formed and the first meeting identified existing issues and various information required to produce a strategy. This information is now being captured by CS.

5. *Consider how to deal with growing server estate, given limited scope for increasing server room space*. We have introduced space management for the server rooms including booking space before procurement, periodically reviewing tenancy and eviction.

6. *Complete the audit of all research data within the School*. A snapshot has been completed. This data will be entered into the School data register once that is complete. The School needs to develop a mechanism to keep this data up-to-date.

**RDM Relevant Goals for 2019/20**

1. Produce a Data Register and a mechanism for users to self populate that Register

2. Populate the Data Register with details of medium and high risk data

3. Improve processes for research data management - including production of Data Management Plans and recording, in the Data Register, details of all research data.

4. Consideration of how best to make use of Information Services' RDM offerings, producing use-case guidance to researchers, in partnership with IS Research

5. Produce a strategy for resourcing the School's compute and data intensive requirements

6. Produce more sustainable and performant compute/GPU clusters - focusing on job management and filesystems for 2019.

Three areas are proposed for discussion for how research data and the associated research infrastructure in the School is to be managed.

# 2 RDM & Infrastructure

## Relationship with and use of Central Research Data Service (RDS) Systems

In recent years the University has made a large investment in developing the environment for research and continues to do so as detailed in the University of Edinburgh Research Data Service Roadmap. The services *DataStore*, *DataVault*, *DataShare*, *DataSync*, *DataSafeHaven* and *ECDF GitLab* are deeply embedded in the University's strategy for research data management and provides Informatics with users a lot of opportunities to remove the burden from School systems, freeing up resource for investment in systems that the University cannot replicate in some of our bespoke research projects. The RDS environment was specifically developed to both facilitate research and help individuals comply with funder and legal regulations. We hosted a presentation by RDS describing these resources this Semester.

It is essential that the School engages more deeply with University RDS services noting that where costs exist they are eligible to be included in grant proposals and a service exists to quote for this at RDS. Whilst it is unlikely that these services will satisfy all of our requirements they need to be adopted more widely as part of the RDM mix to relieve pressure on the School. Our greater engagement would likely open further opportunities to help shape the direction of developments of RDS systems and better highlight where the systems don't meet our requirements.

### Recommendations

- **Recommendation 1** The School should develop ways to ensure that our systems operate well with RDS services and should develop guides and/or training for staff on how to work with those services.
- **Recommendation 2** The School should actively promote RDS services.

## Adoption of Data Management Plans as Standard

The formation of a Data Management Plan (DMP) is increasingly required for grant proposals, but commonly elucidates the general processes rather than providing a detailed *Data Life-Cycle* approach to following RDM through all of the phases of a project with specific costed proposals. A DMP should be created at the beginning of each research project as standard so that necessary consideration is given to the resources, permissions, funds and

compliance requirements needed to enable that research to be successfully and ethically conducted. By developing a policy that requires the creation of a DMP several of the key challenges that we face can be addressed in parallel.

1. DMPs linked directly into WorkTribe and/or Theon or Qualtrics based monitoring/records system
2. GDPR/DPIA and Ethics process compliance can be incorporated by design
3. Computing resource requirements are identified and shared with CS
4. Prospective assessment of future demand becomes possible
5. Live projects, including data details will be captured
6. Identified project way-points can act as trigger points for monitoring and/or reporting events and end-of grant compliance activities.
7. Data will be up-to-date facilitating planning and monitoring
8. Full auditing process no longer needed. DMPs distrbute the workload.

**Recommendations**

- **Recommendation 1** Establish a policy requiring that research projects have a DMP
- **Recommendation 2** Provide use-case examples & DMP templates in DMPOnline
- **Recommendation 3** Provide funder specific use-case examples & boiler-plate extracts in DMPOnline
- **Recommendation 4** Provide guidance and/or training in development and use of DMPs using DMPOnline and for the new School process
- **Recommendation 5** Ensure that specific areas for GDPR, Ethics and Data Security are incorporated and that a process is in place either in WorkTribe, Qualtrics or Theon system(s) to capture and monitor these aspects alongside the DMP itself

## Active Research Data & Compute

The University envisions *DataStore* as the *Active Data* workhorse for University research, but for us this can only be realistic for some of our ongoing research projects. In reality we need to maintain our hybrid approach using internal School servers alongside central services and should evaluate the options for developing service level agreements with commercial cloud compute and storage services.

Our compute infrastructure is currently limited by space, power and support capacity, but we do not have accurate information detailing the usage of our installed systems or their

efficiency in terms of performance and energy consumption. It seems likely that much of the installed hardware is past end-of-life and does not make good use of either space or power and in several cases is little used.

- **Recommendation 1** Audit our server hardware and evaluate options for quantifying how much a piece of hardware is being used, either directly or indirectly (power consumption).

Where a piece of hardware is being used by a research project, but is past end-of-life we need to decide what happens to that hardware and under what circumstances. If such a piece of hardware is decommissioned (to make space for new hardware) what, if anything, should the School put in place to allow that research to continue?

- **Recommendation 2** Review the policy for managing servers in the self-managed server room and communicate any change in policy and any emerging support process for continuing projects that have lost hardware to Staff.

The custom of installing self-managed servers for projects has become somewhat standard, but is not scalable as our research portfolio grows. We need to develop an approvals process for such systems so that space can be properly managed and where possible CS could develop services (such as virtualisation within the School, cloud compute, Research GPU cluster(s)) that can meet many research needs without the need for stand alone systems. This would enable the School to procure compute more effectively and manage installation and maintenance efficiently.

- **Recommendation 3** Conduct a survey to evaluate the breadth and distribution of compute and storage requirements for Research groupings in the School to assess to what extent re-balancing of resources towards School provision of core research compute services is feasible.

Whilst such a provision would not remove the need for bespoke servers, it could make a valuable contribution to alleviating our current pressures, removing the need for research groups to install their own hardware. In order for such a School service to be financially viable, grant application costings would need to include costs for access to such facilities at a *"Guaranteed Service Level"* developed by the School in place of funding for physical hardware. Careful consideration would need to be given to ensure "uncosted" research is not unfairly disadvantaged by any proposed system.

- **Recommendation 4** Explore the feasibility of a facility based costing model for the development of core research compute services in the School.

## 3 Actions

Research Committee members to consider the proposals and provide specific feedback and suggestions to DoR/dDoR. Once the supported components are known they will be developed into policies and implementation plans by the Computing Strategy Committee, Ethics Panel & Research Data Management Teams in consultation with DoR/dDoR. Advice will also be sought from the University Data Protection Officer and Research Data Services. Proposals will then be tabled with the relevant School Committee(s).

## 4 Equality & Diversity

N/A

## 5 Resource Implications

Currently considered to be within the remit of Research Services, Computing Support and DoR/dDoR with support from Computer Strategy Group (including sub-groups) and Ethics Panel members. Proposals developed from this document will specify additional resource needs as required.

## 6 Links to Reference Materials

- Open Data
- Research Data Management
- Funders Data Policies
- University Research Data Management Policy (*)
- University of Edinburgh Research Data Service Roadmap (August 2017-July 2020)
- UKRI Concordat on Open Research Data

*Since 2011, all new research proposals must include research data management plans or protocols that explicitly address data capture, management, integrity, confidentiality, retention, sharing and publication.