Institutional support for best practice in Research Data Management (and funder compliance)

NFAIS Virtual Seminar on Open Data, 16 June 2016

Dr Torsten Reimer, Scholarly Communications Officer
http://orcid.org/0000-0001-8357-9422 / @torstenreimer
Imperial College London
"At a research intensive university like Imperial it is hard to do anything that doesn't involve data."

James Stirling, Provost

"Data is at the heart of the human condition."

Joanna Shields, UK Minister for Internet Safety and Security

Data Science hub and KPMG Data Observatory launch at Imperial College London, 04 November 2015
… and yet we are losing research data

“In their parents’ attic, in boxes in the garage, or stored on now-defunct floppy disks — these are just some of the inaccessible places in which scientists have admitted to keeping their old research data.”

http://www.nature.com/news/scientists-losing-data-at-a-rapid-rate-1.14416
Isn’t research meant to be reproducible?

The results of only 6 out 53 ‘landmark’ studies were found reproducible.

Drug development: Raise standards for preclinical cancer research. DOI: doi:10.1038/483531a

“Several recent publications suggested that the seminal findings from academic laboratories could only be reproduced 11–50% of the time. The lack of data reproducibility likely contributes to the difficulty in rapidly developing new drugs and biomarkers that significantly impact the lives of patients with cancer and other diseases.”

A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. DOI: 10.1371/journal.pone.0063221
## Funder reaction: UK requirements for Research Data

<table>
<thead>
<tr>
<th>Research Funders</th>
<th>Published outputs</th>
<th>Data</th>
<th>Time limits</th>
<th>Data plan</th>
<th>Access/sharing</th>
<th>Long-term curation</th>
<th>Monitoring</th>
<th>Guidance</th>
<th>Repository</th>
<th>Data centre</th>
<th>Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>BBSRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>CRUK</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>EPSRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>ESRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>MRC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>NERC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>STFC</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>○</td>
</tr>
</tbody>
</table>

![Image](http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies)
Institutions in receipt of EPSRC funding must meet requirements like:

- A **data management plan** (DMP) exists for every project.
- All publications to include **data access statement**.
- All data required to validate the findings should be **retained for 10 years** after the date of last access.
- **Metadata** describing the data held must be made **freely available** on the internet normally within 12 months of the data being generated.
  - Metadata must be sufficient to allow others to **understand** what research data exists, why, when and how it was generated, and how to **access** it.
  - Where access to the data is **restricted** the published metadata should also give the reason and summarise the conditions which must be satisfied for access to be granted.
- Monitored through periodic "dip-stick" testing.
Faculties of Engineering, Medicine, Natural Sciences and the Business School

Ranked 3rd in Europe / 8th in the world (THE 2015-16 rankings)

Net income (2015): £969m, incl. £428m research grants/contracts

~15,000 students, ~8,000 staff, incl. ~3,900 academic & research staff

Staff publish 10-12,000 scholarly articles per year

Largest data traffic into Janet network of all UK universities
Process of RDM policy development

- Set up a governance structure, coordination across College
- Aim: guide academics through funder requirements and to best practice
- Policy not be implemented until College can support compliance
  - Lack of reliable data (on data storage needs, scale in particular)
  - Concerns about cost of maintaining infrastructure
  - Concerns about uncertainties and changing market / policy landscape
- Approach
  - **RDM Green Shoots**: 6 bottom-up, academic projects (2\textsuperscript{nd} half of 2014)
  - RDM investigation (Oct 2014-Jan 2015)
    - Online survey (academics; 390 responses), in-depth interviews with academics (~40), workshops (academics & data managers)
    - Investigation into flexible, cost-effective infrastructure components

⇒ Deliver a solution that’s good enough for the 80% who (usually) don’t have specialised requirements
Online survey – where does active data live?

Use of different types of storage in %

- College computer: 70%
- External/portable storage: 50%
- Cloud storage: 40%
- Personal computer: 30%
- Departmental/group storage: 20%
- College H drive: 10%
- ICT central storage: 10%
Example “Green Shoots” project

Research Data Management in Computational and Experimental Molecular Science (H. S. Rzepa, M. J. Harvey, N. Mason & A. Mclean)

Address sustainability and scalability of a hub interfacing e-lab notebooks with HPC resources and repositories

• Data can be deposited from HPC cluster
• Integrate ORCID, incl. login option
• DOIs, DataCite compliant metadata
• Curated 170,000 older datasets from Cambridge, adding metadata
• Developed a procedure using DOIs for directly retrieving data from a digital repository and displaying it using Javascript components

http://doi.org/10042/a3v1w
Some findings

- RDM principles are considered to be sound but not fully practised
- Concerns around back-up and archiving, esp. considering data volume
- Concerns about effort to make shared data discoverable and reusable
- Popularity of cloud services (as opposed to College storage)
- **60-100% of grant required to re-generate data used in publications**
- % of data that needs retaining to support publications: ~60%
- Data storage capacity will have to grow significantly

⇒ Researchers want self-administered, secure, responsive solution for data sharing, storing and archiving; open APIs and automatic metadata generation preferred

(“Yes [storage] is really important. Basically, whenever we have been out to talk to researchers, that's the thing they have latched on to and want to talk about the most.” [10.1371/journal.pone.0114734](https://doi.org/10.1371/journal.pone.0114734))
Conclusions / policy implementation principles

- Provide platform-independent, flexible data storage
- Embed RDM training into PhD progression
- Where available, uses existing workflows:
  - Symplectic Elements: metadata management
  - Spiral (DSpace repository): public (metadata) catalogue
- Additional infrastructure:
  - use external resources
  - no long-term commitment
  - as flexible as possible
  - cost-effective
“Imperial College London is committed to promoting the **highest standards of academic research**, including excellence in research data management. This includes a robust digital curation infrastructure that supports **open data access** and protects confidential data.”

“**Principal Investigators have overall responsibility** for the effective management of research data generated within or obtained for their research, including by their research groups. The **Library** and **ICT** will provide training, guidance and services to **support PIs**.”

http://imperial.ac.uk/research-data-management
RDM service infrastructure development

1. Make a data management plan: use DMPOnline

2. Store your data management plan centrally: use InfoEd

3. Store your live data securely and safely: use Box

4. Store your final data (and/or code) for 10+ years, making it publicly available: use Zenodo

5. Tell the College where your data (and/or code) is published or stored: use Symplectic

6. Reference your funding and your data in the publications it underpins: tell your publisher

RDM Workflow, College Library Services
10.5281/zenodo.54000
Research Project

Creates data/software

Data: Box
Software: GitHub

Project ends

Data/software still needed

no
 Delete

yes

Can it be published or embargoed externally?

Internal Storage

no

External repository

yes

Can metadata be published?

Metadata, manual or automatic

Library reviews

Can metadata be published?

Elements

 Metadata, manual or automatic

Spiral
Box – Data storage, sharing and syncing

- Unlimited, backed-up data storage
- Access from anywhere, from desktop and mobile devices
- Automatic syncing of data across devices
- Share within and outside College
- Machine learning tools to help describe data
- Integration with Microsoft Office and other solutions, viewers for different file formats included

Issues
- EU storage option coming soon (sensitive data)
- No dedicated Linux client
College requires academics to archive the particular version of code developed in a project to generate or analyse data.

College-funded PyRDM project developed library to automate this process. College-recommended repository Zenodo offers GitHub integration.

College launched survey on DVS – 274 responses, 82% use Git

Decision: College to provide GitHub Enterprise to all staff

College survey on distributed version control
Software Sustainability Institute – I am a fellow
Identifiers: Imperial ORCID project

College became ORCID member in 2014:
• Raise awareness and uptake
• Issue researchers with an iD

Approach:
• Capture existing iDs (in Symplectic)
• Create new iDs on behalf of academics
• Encourage academics to link iD to Symplectic

Outcomes:
• ~75% of iDs claimed
• Academics linked 1,800 iDs to Symplectic
• Ongoing awareness raising and work with ORCID community (Imperial hosted 1st UK ORCID (HE) members meeting in 2015)

https://www.imperial.ac.uk/orcid
https://dx.doi.org/10.1629/uksg.268
Towards compliance as by-product of good workflows

Working towards:

- One workflow for data generation, publishing, reporting and curation
- Link data generation directly to storage (log into facility, data “at your desk” before you are out of the “lab”)
- Automate reporting and generating / sharing of metadata
Recommendations

• Clear governance structure, coordinated activity across the university
• Take a flexible approach, don’t wait for the perfect solution
• Use existing, in particular disciplinary solutions where possible
• Decide how to interpret funder policies in a useful way
• Engage with academic requirements to get them engaged
• Make “compliance” a by-product of good workflows
  • Aim to simplify and automate workflows
  • Academics ideally only interact with each output once
  • Systems automate sharing of data, data enhanced as it is passed on

⇒ Consider how you can add value to the research (data) process