Experiences of running an institutional data repository

Stuart Lewis
Deputy Director, Library & University Collections
@stuartlewis

Pauline Ward
Data Library Assistant
@PaulineDataWard

THE UNIVERSITY of EDINBURGH
Experiences of running an institutional data repository

1. The context – The University of Edinburgh
2. The theory – Reasons for Research Data Management
3. The policy – Local and national policies
4. The programme – An overview of the Research Data Service
5. The practice – The DataShare repository
The context
The University of Edinburgh

• The context to our work:
  • Edinburgh is the capital city of Scotland
  • The University of Edinburgh is Scotland’s largest university
  • A large thriving research-led University: 35,258 students, 13,272 staff
  • Breadth of research disciplines across three colleges:
    • Humanities and Social Science
    • Science and Engineering
    • Medicine and Veterinary Medicine
The University of Edinburgh

• Information Services Group at the University of Edinburgh - CIO
  • Library & University Collections
  • User Services
  • IT Infrastructure
  • IT Applications
  • Learning Teaching and Web

• Digital Curation Centre and EDINA
The Theory
What is research data?

• Observational: real time, unique and irreplaceable
• Experimental: reproducible, may be expensive
• Simulation: modelled data; model & metadata may be more important than output data
• Derived or compiled: combining 'raw' data, reproducible, may be expensive
• Reference or canonical: collection of peer reviewed datasets, published and curated

Research Information Network. “Stewardship of digital research data - principles and guidelines

• Not just ‘scientific’ data!
Why manage research data?

• Benefits to different groups...
Seeking win + win + win + win + win......

Where do I safely keep my data from my fieldwork, as I travel home?

How do I best keep years worth of research data secure and accessible for when I and others need to re-use it?

How can our research collaborations share data, and make them available once complete?

How do we ensure compliance to funders’ requirement for several years of open access to data?

How do we ensure we have access to our research data after some of the team have left?

Professor Jeff Haywood
Vice Principal for Digital Education
(Ex CIO)
Why manage research data?

• Benefits to different groups...

• If for no other reason, access to **my own** data!
  
  A true story....
LOST/STOLEN BAG

Have you seen this bag?
It was stolen from here on Thursday 22nd
October around midday
It has half of my university dissertation research inside!!

If found please call on
Reward if found. Thanks 😊
The Policy
Research Data Management Policies

• Growing policy support for Research Data Management
  • University of Edinburgh Policy – May 2011
  • UK RCUK Research Councils
  • EC Horizon 2020 pilot
  • Many other research funders
Research Data Management Policies

• University of Edinburgh Policy
  • First institutional RDM policy in the UK
    • Approved by Senatus Academicus in May 2011
    • http://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy

• An aspirational policy
  • “It is acknowledged that this is an aspirational policy, and that implementation will take some years.”

• Assigns responsibilities
Research Councils UK (RCUK) Principles

1. Data made freely and openly available as soon as possible
2. Data Management Plans required, preservation of long-term data is required
3. Appropriate metadata should be made openly available
4. Acknowledgement of constraints on data release
5. Dataset users should acknowledge sources
6. Limited period of privileged access
7. It is appropriate to use public funds to support preservation and management of data
Horizon 2020

All project proposals submitted to "Research and Innovation actions" as well as "Innovation actions" include a section on research data management which is evaluated under the criterion 'Impact'. Where relevant, applicants must provide a short, general outline of their policy for data management, including the following issues:

• What types of data will the project generate/collect?
• What standards will be used?
• How will this data be exploited and/or shared/made accessible for verification and re-use? If data cannot be made available, explain why.
• How will this data be curated and preserved?
The Programme

CONTENTS.

Cyanometer..........................475
Photometer..........................476
Endiometer..........................477
Instruments for Terrestrial Magnetism...id.

SECTION II.

Operations
Measurements by the eye..................id.
Estimation of distances by actual measure........480
Standard measures of length.................481
To walk in a straight line................id.
To measure distances by pacing and time......482
Of measuring distances by sound.............483
To measure a straight line................id.
To raise a perpendicular from any point on a given line 484
From a given point out of the line, to let fall a perpendicular to it.........485
To raise a perpendicular at the end of a line........id.
To draw a line parallel to a given line........487
To make an angle on the ground equal to a given angle...............id.
To measure an angle of a building, &c., having no proper instrument........id.
To ascertain the length of a line accessible only at its two extremities....488
To ascertain the breadth of a river, a marsh, &c...................489
To measure the distance of two inaccessible objects from each other........491
To take the plan of a wood, a marsh, a lake, a crater, or other hollow.....492
To take the plan of any crooked line, as of a river, a neck of land, &c. 492
To obtain a meridian line................id.
To ascertain the height of a building when the base is accessible........493
To measure the time........................492
To ascertain the time of day or night........493
To make a sand glass, or a clock..........501
To know the hour of the day or night........501

CONTENTS.

To kindle a fire..........................512
To ascertain the breadth, depth, and slope of the bed of a river........513
To ascertain the velocity of a river........516
To find the quantity of water discharged by a river at a given point in a given time.....518
To ascertain the quantity of solid matter held in suspension by running water 521
To observe the temperature of the air..................523
To observe the temperature of springs........524
To take the temperature of rivers, of lakes, of the sea, and of wells........525
To take the temperature of rain..................526
To take the temperature of the soil..............527
To take the temperature of gases rising from Volcanoes..................528
To observe the atmospheric pressure........529
To determine the transparency of water........530
To observe the colour of water..................531
To ascertain the quality of water..............532
To take the specific gravity of water............532
To bring up water from considerable depths........id.
To estimate the evaporation from standing or running water..........id.
To observe the quantity of rain, snow, or hail..................533
To observe the quantity of dew which falls........id.
To observe the force of the wind..................534
To observe Solar radiation..................id.
To observe Terrestrial radiation................id.
To observe the electrical state of the atmosphere..................535
To observe Terrestrial magnetism................id.
To collect the gases from Volcanoes, Springs, &c. ..........................536
To observe the direction of the undulatory motion of earthquakes........id.
To measure the heights of Mountains...........id.

Method of taking the dip and strike of strata........544
To take the length, breadth, circumference, and surface of lakes........546
To copy inscriptions.....................547

To MAP A COUNTRY

..........................548
The University of Edinburgh’s approach

• Research Data Management services for everybody

We must provide core infrastructure for all researchers to support the transition to Open Science
University of Edinburgh RDM Programme

• Planning a Research Data Management Service:

• Research Data Management Programme
  • Delivered by Information Services
  • Supported by central funding
  • Phase 1: August 2012 to May 2015
  • Phase 2: June 2015 to July 2016
  • RDM Roadmap:
    http://www.ed.ac.uk/information-services/about/strategy-planning/rdm-roadmap
Research Data Management Services

Data Management Support

- Data Management Planning
- Active Data Infrastructure
- Data Stewardship
Data Management Planning

- DMPOnline National tool to create Data Management Plans
  - [https://dmponline.dcc.ac.uk/](https://dmponline.dcc.ac.uk/)

<table>
<thead>
<tr>
<th>Sections</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Capture</td>
<td>- Describe the types of data the research will generate</td>
</tr>
<tr>
<td></td>
<td>- How will the data be documented?</td>
</tr>
<tr>
<td></td>
<td>- How much data will be generated?</td>
</tr>
<tr>
<td>Data Management</td>
<td>- Where will the data be stored?</td>
</tr>
<tr>
<td></td>
<td>- How will the data be backed-up?</td>
</tr>
<tr>
<td></td>
<td>- Who will be responsible for this data management?</td>
</tr>
<tr>
<td>Integrity</td>
<td>- How will you quality assure your data?</td>
</tr>
<tr>
<td>Confidentiality and IPR</td>
<td>- How will you manage any ethical issues?</td>
</tr>
<tr>
<td></td>
<td>- How will you manage copyright and Intellectual Property Rights (IPR) issues?</td>
</tr>
<tr>
<td>Retention &amp; Preservation</td>
<td>- How long should the data be retained after the end of the research?</td>
</tr>
<tr>
<td></td>
<td>- Which data should be retained after the research?</td>
</tr>
<tr>
<td></td>
<td>- Identify the repository to which the data will be offered for deposit</td>
</tr>
<tr>
<td>Sharing &amp; Publication</td>
<td>- Specify and justify which data will have value to others and should be shared</td>
</tr>
<tr>
<td></td>
<td>- Indicate how data will be shared</td>
</tr>
<tr>
<td></td>
<td>- Are any restrictions on data sharing required?</td>
</tr>
</tbody>
</table>
Active Data Infrastructure

• **DataStore**
  - 0.5 TB per person (PGRs upwards)
  - Personal allocation
  - 5TB per ‘project’
  - Extra can be purchased by grants @ £200 per TB per year

• **DataSync**
  - OwnCloud
    - Open Source DropBox-like web / sharing / sync system
    - Imminent launch

• Trusted Research Environment / Data Safe Haven
Data Stewardship

• **PURE**
  - Current Research Information System (CRIS)
  - Records research outputs (papers, grants, equipment, awards, datasets, etc)
  - Allows datasets to be described, and linked to if shared online

• Person A, was awarded Grant B, which funded Equipment C, which created data D, which generated paper E
Data Stewardship

• DataVault
  • Long term archival storage.
  • Move data from DataStore.
  • In development
    • With University of Manchester
    • Sponsored by Jisc
    • http://datavaultplatform.org/
Data Stewardship

• DataShare
  • Online open data repository
  • Uses the DSpace open source repository platform
  • Creates DOIs for datasets
  • Data Seal of Approval (DSA)
  • http://datashare.is.ed.ac.uk/
Data Management Support

• Awareness raising sessions
• Training courses
• On-demand support
• MANTRA online course
  • http://datalib.edina.ac.uk/mantra/
  • http://datalib.edina.ac.uk/mantra/libtraining.html
What is Edinburgh DataShare?

Edinburgh DataShare is a digital repository of research data produced at the University of Edinburgh, hosted by Information Services. Edinburgh University researchers who have produced research data associated with an existing or forthcoming publication, or which has potential use for other researchers, are invited to upload their dataset for sharing and safekeeping. A persistent identifier and suggested citation will be provided.

Deposit Your Data

Deposit data
DataShare

• DataShare is an institutional repository which allows University of Edinburgh researchers to share their data online.

• It uses the DSpace open source repository system.
DataShare History

• Created in 2009 as part of a project to assess whether traditional institutional repository software could handle data.

• Is now the catch-all repository for data which either does not have an appropriate disciplinary repository or is too big or complex for the disciplinary repository.
DataShare Growth

No. Items Accessioned

Year

No. Items Accessioned:
- 2010: 3
- 2011: 5
- 2012: 44
- 2013: 24
- 2014: 110
- 2015: 1076
- 2016: 111
Configuration of DSpace for DataShare

• The following configuration is applied to our DSpace instance
  • Dublin Core metadata fields
    • https://www.wiki.ed.ac.uk/display/datashare/metadata-2.1
  • Look and feel / branding
Extra Features in DataShare cf DSpace

• The following features are additions / customisations made by our developer
  • HTML5 upload, allowing files up to 9 GB to be uploaded via web deposit interface, and multiple file upload via drag’n’drop
  • Download All button
  • DOI assignment
  • Coming soon: FTP server to facilitate download of datasets above 20 GB
Challenges

• Big files
  • Upload and download

• Catch-22: Depositors want the DOI before depositing data

• Culture change
  • Make data sharing the norm
    • Researchers are busy, want the process to be very quick
    • Require minimum documentation, encourage detailed documentation

• Not using ORCID

• Not detecting DOI citation / not capturing Altmetrics
Experiences of running an institutional data repository

stuart.lewis@ed.ac.uk       pauline.ward@ed.ac.uk

Děkuji mnohokrát!

THE UNIVERSITY of EDINBURGH