Big data: the potential role of research data management and research data registries

Joy Davidson, Sarah Jones and Laura Molloy
Digital Curation Centre (DCC)

www.dcc.ac.uk
The DCC Mission

“Helping to build capacity, capability and skills in data management and curation across the UK’s higher education research community”

Phase 3 Business Plan

www.dcc.ac.uk
RDM policy drivers for universities

Declaration data are a public good and should be openly available.

Common principles on data policy
www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

EPSRC has the following clear expectations of organisations in receipt of EPSRC research funding:

i. Research organisations will promote internal awareness of these principles and expectations and ensure that their researchers and research students have a general awareness of the regulatory environment and of the available...
The ‘big data’ challenge for universities

Large datasets

Often curated by large international projects, national facilities or disciplinary initiatives.

Some large data, but mostly a long-tail of small, distributed datasets with diverse needs.
Services to help unis

Jisc Research Data Registry and Discovery Service

A service to ensure that research data held within UK data centres and universities can be found, understood and reused

A web-based tool to help researchers write Data Management Plans
What is a data management plan?

A brief plan to define:

• how the data will be created?
• how it will be documented?
• who will access it?
• where it will be stored?
• who will back it up?
• whether (and how) it will be shared & preserved?

DMPs are often submitted as part of grant applications, but are useful whenever researchers are creating data.
Why create Data Management Plans?

• Plans help researchers to manage their data
  – Make informed decisions
  – Anticipate and avoid problems

• Many funders ask for DMPs in grant proposals
  www.dcc.ac.uk/resources/data-management-plans/funders-requirements

• DMPs are often required by universities too
  “All new research proposals must include research data management plans or protocols that explicitly address data capture, management, integrity, confidentiality, retention, sharing and publication.”
DMPPonline

- Free to use
- Code is open source
- Helps researchers write DMPs
- Provides funder questions & guidance
- Provides help from universities
- Examples and suggested answers
- ...

https://dmponline.dcc.ac.uk
Registration

Sign up with your email address, organisation and password

Select ‘other organisation’ if yours is not listed
Creating a plan

Select funder (if any)
Select organisation for additional questions and guidance
Select other sources of guidance
Plan details: summary

This plan is based on:

| Funder | Economic and Social Research Council |

The ESRC requires that all applicants seeking ESRC funding include a statement on data sharing in the relevant section of the Je-S application form. If data sharing is not possible, the applicant must present a strong argument to justify their case.

<table>
<thead>
<tr>
<th>Sections</th>
<th>Questions</th>
</tr>
</thead>
</table>
| Existing data                                 | - An explanation of the existing data sources that will be used by the research project (with references)  
- An analysis of the gaps identified between the currently available and required data for the research |
| Information on the data that will be produced | - Methodologies for data collection  
- Data volume and data type, e.g. qualitative or quantitative data  
- Data quality, formats, standards documentation and metadata |
| Planned quality assurance and back-up procedures (security/storage) | - Quality Assurance  
- Back-Up |
| Management and archiving of collected data    | - Plans for management and archiving of collected data |
| Overcoming data sharing difficulties          | - Expected difficulties in data sharing, along with causes and possible measures to overcome these difficulties. |
| Consent, confidentiality, anonymization and other ethical considerations | - Explicit mention of consent, confidentiality, anonymisation and other ethical considerations |
| Copyright and intellectual property ownership of the data | - Copyright and intellectual property ownership of the data |
| Responsibilities for data management and curation | - Responsibilities for data management and curation within research teams at all participating institutions |

Summary of the sections and questions in your DMP
Answering questions

My project (DCC Template)

Data Collection (2 questions, 0 answered)

Documentation and Metadata (1 question, 1 answered)

What documentation and metadata will accompany the data?

Metadata will be tagged in XML using the Data Documentation Initiative (DDI) format. The codebook will contain information on study design, sampling methodology, fieldwork, variable-level detail, and all information necessary for a secondary analyst to use the data accurately and effectively.

DCC Guidance

Questions to consider:
- What information is needed for the data to be read and interpreted in the future?
- How will you capture/create this documentation and metadata?
- What metadata standards will you use and why?

Guidance:
Describe the types of documentation that will accompany the data to help secondary users to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed.

Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.

Notes who has answered the question and when

Progress bar updates how many questions remain

Answered less than a minute ago by Sarah Jones
Sharing plans

Withdrawal of services for young people

You can share your plan to allow others to read or edit it. Please insert the email address of the person you wish to share it with. You can leave a note to explain why you are sharing the plan, or what you wish them to look at.

Collaborators

<table>
<thead>
<tr>
<th>User name</th>
<th>Permissions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarah Jones</td>
<td>Owner</td>
</tr>
<tr>
<td>Laura Molloy</td>
<td>Read only</td>
</tr>
</tbody>
</table>

Add collaborator

Email
Permissions:
Read only

Add collaborator

Allow colleagues to read-only, read-write, or become co-owners
Co-writing DMPs

Sections are locked for editing when they’re being worked on by colleagues
### Exporting DMPs

**Withdrawal of services for young people**

**ESRC Data Management Questions**

#### Existing data

<table>
<thead>
<tr>
<th>Questions</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>- An explanation of the existing data sources that will be used by the research project (with references)</td>
<td>The ESOS archive has been systematically searched using a series of search terms related to and derivative of ‘public service withdrawal’, ‘impacts’, and/or ‘children’ and ‘young people’. Our overall assessment is that there are no datasets that will adequately address the aims of this project. The following datasets are tangentially related to this project.</td>
</tr>
<tr>
<td>- An analysis of the gaps identified between the currently available and required data for the research</td>
<td>Given the contemporary nature of the proposed project, we know of no datasets that cover users’ (especially young people’s) views and experiences of austerity measures and service withdrawal. The proposed project will therefore capture new and unprecedented data, for which there is an evident demand among national and regional stakeholders, decision-makers and service-providers (see ‘Pathways to Impact’ attachment). Moreover, whilst several datasets incorporate longitudinal data, none includes data gleaned from oral history and multigenerational family interview methods central to the proposed project. The proposed project therefore represents an extension to the methods and data quality of the tangentially-related projects listed in section 1. Finally, the novel use of ‘impact’ activities to generate data (e.g. workshops, mapping software) exceeds the scope of all extant and even tangentially-related datasets.</td>
</tr>
</tbody>
</table>

#### Information on the data that will be produced

<table>
<thead>
<tr>
<th>Questions</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Data volume and data type, e.g. qualitative or quantitative data</td>
<td>The project shall generate new quantitative data (Bristol Online Survey outputs, SPSS data and outputs), qualitative data (digital audio files, audio transcripts, digital photographic and video data, workshop outputs, NVivo files), and mapping data (TIFF files). Metadata, in the form of pdfs and Excel spreadsheets, shall be used to facilitate the management and archiving of these data. Data shall be stored in password-protected folders on the host institution’s secure servers. Data transfer between the PI and CoI shall take place via face-to-face meetings.</td>
</tr>
<tr>
<td></td>
<td>Quantitative data shall be generated from an anonymous online survey (target 10,000 responses). The survey will be administered via Bristol Online Surveys (BOS) software: a secure, quality-assured, widely-used online survey tool. Data will be exported to SPSS for analysis.</td>
</tr>
</tbody>
</table>
Unis can customise the tool....

• Add templates
• Add custom guidance
• Provide example or suggested answers
• Monitor usage in their organisation
• Offer foreign language versions
• ...

www.dcc.ac.uk/news/customising-dmponline-admin-interface-launches
Ways libraries could use DMPonline

• To identify requirements for support
  – volume of data being created
  – what needs to be preserved and for how long

• Monitor awareness or uptake of services
  – how many DMPs say they will use repository?

• For outreach – advocate and improve practice
More information

Screencast on how to use DMPonline

http://www.screenr.com/PJHN

Customising DMPonline

www.dcc.ac.uk/news/customising-dmponline-admin-interface-launches

GitHub

Get the code, amend it, run a local instance, flag issues, request features...

https://github.com/DigitalCurationCentre/DMPonline_v4
Research data repositories and registries

Image Credit: Hitachi Data Systems
Archiving – external data centres

Research funders’ data centres...

Structured databases

Registries of international data centres

Disciplinary & community initiatives
Institutional data repositories

Not intended to replace national, subject or other established data repositories

Acknowledging hybrid environment

http://datashare.is.ed.ac.uk

www.dspace.cam.ac.uk

https://databank.ora.ox.ac.uk

Research Data at Essex and DataPool at Southampton
Data catalogues

- Oxford is developing its DataFinder tool [http://blogs.it.ox.ac.uk/damaro](http://blogs.it.ox.ac.uk/damaro)

- Research Data @ Essex has developed a profile based on DataCite, Inspire and DDI standards [www.data-archive.ac.uk/media/395364/rde_march2013_repositoryoutputs.pdf](http://www.data-archive.ac.uk/media/395364/rde_march2013_repositoryoutputs.pdf)

- C4D is developing a research data extension to the CERIF standard - [http://cerif4datasets.wordpress.com](http://cerif4datasets.wordpress.com)

- CKAN is being explored by various projects [http://ckan.org](http://ckan.org)
Research Data Registry and Discovery (RDRDS) pilot

What’s in the Research Data Registry and Discovery Service

Collections (49)
Research datasets or collections of research materials.

Parties (36)
Researchers or research organisations that create or maintain research datasets or collections.

Activities (0)
Projects or programs that create research datasets or collections.

Services (0)
Services that support the creation or use of research datasets or collections.

Spotlight on research data

http://rdrds.cloudapp.net/

Who contributes to the Research Data Registry and Discovery Service?

Research organisations from around the UK contribute information to Research Data Australia.
See All
Where does RDRDS fit in?
RDRDS metadata

- Description/Abstract
- Dataset identifier
- Subject
- URL of landing page
- Creator (+ID)
- Release date
- Rights information
- Spatial coverage
- Temporal coverage
- Publisher
Jisc RDRDS Pilot Partners

Data centres:
- UK Data Archive
- NERC Data Catalogue Service
  - BADC
  - BODC
  - EIDC
  - NEODC
  - NGDC
  - PDC
  - UKSSDC
  - ADS

Universities:
- Edinburgh
- Glasgow
- Hull
- Lincoln
- Leeds
- Oxford
- Oxford Brookes
- St Andrews
- Southampton
RDRDS Phase 2

- Define a set of clear use cases and workflows.
- Compare different possible platforms for the service and assess their suitability.
- Establish a working instance of the system, involving all UK data centres and university data repositories.
- Establish a simple workflow for adding more data sources to the service, adapting to changes in existing data sources, and avoiding duplication.
- Test the system for usability.
- Produce recommendations for quality and standardisation of metadata records.
- Evaluate the costs and benefits of the system.
Short-term opportunities for ongoing research and development
Harvesting metadata from CRIS to institutional data catalogues

OpenAIRE Guidelines for

1 Literature Repository managers (version 3.0)
2 Data Archive Managers
3 CRIS Managers

http://guidelines.openaire.eu
Include equipment data into data management plans and institutional catalogue records

equipment.data
Enabling access to UK HE research equipment

Search for equipment: e.g. Microscope

It's good to share!
Welcome to equipment.data, your one stop shop for accessing UK wide HE research equipment. Use this searchable database to locate a piece of equipment or facility within your own or at another institution. equipment.data could be your starting point for some valuable collaborations.

http://equipment.data.ac.uk/
Publish DMPs

http://www.metajnl.com/
Use of data registries to gather usage statistics: institutions

http://www.ref.ac.uk/
Use of data registries to gather usage statistics: funders

http://gtr.rcuk.ac.uk/

https://www.researchfish.com/
Use of data registries to gather usage statistics: researchers

DISTINGUISH YOURSELF IN THREE EASY STEPS

ORCID provides a persistent digital identifier that distinguishes you from every other researcher and, through integration in key research workflows such as manuscript and grant submission, supports automated linkages between you and your professional activities ensuring that your work is recognized. Find out more.

1. REGISTER
   Get your unique ORCID identifier Register now! Registration takes 30 seconds.

2. ADD YOUR INFO
   Enhance your ORCID record with your professional information and link to your other identifiers (such as Scopus or ResearcherID or LinkedIn).

3. USE YOUR ORCID ID
   Include your ORCID identifier on your Webpage, when you submit publications, apply for grants, and in any research workflow to ensure you get credit for your work.

http://orcid.org/
Longer-term challenges
Providing tools to analyse and manipulate the data

http://www.ceda.ac.uk/services/
Common agreement on attribution stacking

https://www.force11.org/datacitation
Thanks – any questions?

DCC guidance, tools and case studies:

www.dcc.ac.uk/resources

Follow us on twitter:

@digitalcuration and #ukdcc