Implementation Guide

October 2009

Acknowledgement
This guide is based on the lessons learned through DAF pilot projects and early exemplars. We're very grateful to those groups for sharing their experiences with us to help refine the methodology and assist future users. They were:

- Cuna Ekmekcioglu and Robin Rice, University of Edinburgh
- Neil Jerrome and Jonathan Breeze, Imperial College London
- Stephen Grace and Gareth Knight, King's College London
- Panayiota Polydoratou and Martin Moyle, University College London
- Harry Gibbs and Teresa McGowan, University of Southampton
- Luis Martinez-Uribe, University of Oxford
- Alex Ball, University of Bath (part of DAF development team)
- Sam Searle, Monash University

We're also indebted to the JISC, which has supported this research.

JISC
University of Glasgow | Humanities Advanced Technology & Information Institute
DCC

CONTENTS

BACKGROUND

What is DAF?

The Data Asset Framework is a set of methods to:

- find out what data assets are being created and held within institutions;
- explore how those data are stored, managed, shared and reused;
- identify any risks e.g. misuse, data loss or irretrievability;
- learn about researchers' attitudes towards data creation and sharing;
- suggest ways to improve ongoing data management.

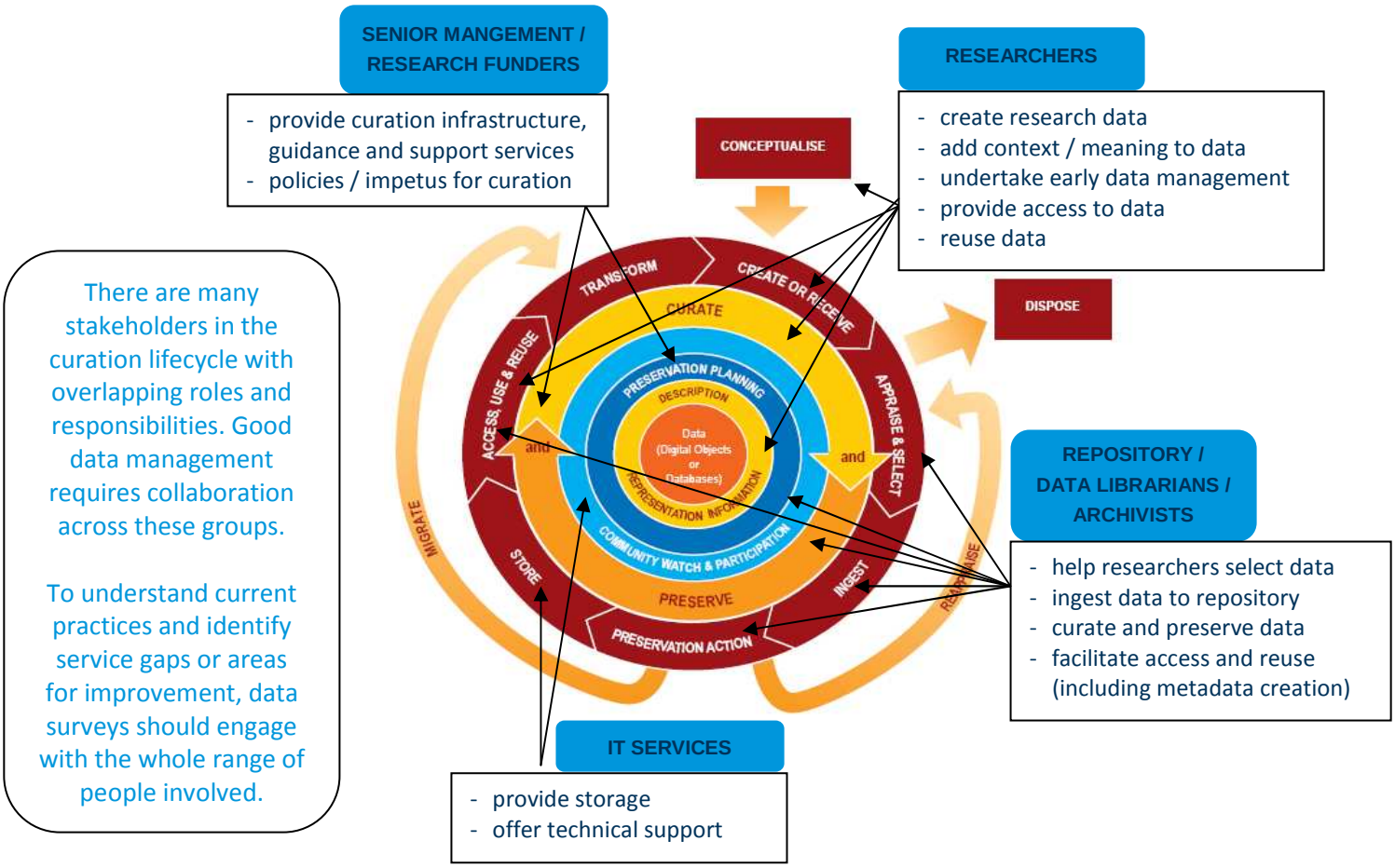
Originally called the Data *Audit* Framework, the tool is being renamed in light of user feedback. Some pilots found the term audit could be off-putting to researchers and misrepresented the survey process, which focuses more on uncovering researchers' data needs and concerns than auditing assets.

Who is DAF for?

DAF was created for Higher Education Institutions to help them take stock of data holdings and ensure appropriate data management practices were in place. It is a useful tool to engage researchers in data curation and to scope their data management requirements. It can also be applied in non-HEI contexts to investigate or build on existing approaches to information management.

The DAF methodology is written for information professionals. It is envisaged the person undertaking a survey would have either a qualification in library, archive or information management, or significant experience working with data. Such skills are needed to understand the information lifecycle and identify risks in existing research workflows and data management practices.

The DAF survey process should involve a variety of stakeholders, for example senior managers, University services such as IT support or repositories, and most importantly researchers.



WHY USE DAF?

What are the benefits?

Stakeholder	Role	Benefits
Information professionals	Undertaking DAF surveys	- Opportunity to engage with users to build relationships: data interviews are a good outreach opportunity - Find out what kinds of support researchers require - Identify content that could be ingested into the Institutional Repository / research outputs archive - Identify opportunities for data sharing - Raise awareness of good practice to ensure information and data are well-managed within the institution
Researchers	Participating in surveys by providing information on data and working practices	- Provide a forum to raise concerns and needs - Opportunity to find out about existing data support - Help shape future services to ensure the support required to curate data effectively is available - Increased awareness of good data management practice to help with future grant proposals - Enable compliance with publishers' requirements e.g. keeping data accessible for validation purposes - Learn how to select and manage data for the long-term so you can find and understand what you need
Senior managers / strategists	Providing high-level support or the impetus for improving data curation	- Undertake service gap analysis to identify duplication of effort and ensure optimum use of existing resources - Acknowledge and mitigate risks associated with data loss, irretrievability or mismanagement - Assist information strategy development - Identify data assets that enhance research status and ensure their value is realised through access and reuse - Increase awareness of data to promote collaboration and interdisciplinary research - To meet expectations such as the OECD principles on public access to funded research

Encouraging participation

Ensuring researcher participation is crucial for success as they create the data and take many decisions affecting long-term curation and reuse. It helps to make the benefits of taking part clear. Some methods used by pilot studies to encourage participation were:

- Using local advocates such as IT technicians or research support, to encourage others to take part
- Obtaining senior management endorsement
- Motivations such as prizes for completing questionnaires
- Attending staff meetings to explain the survey aims and understand researcher expectations / needs
- Determining approach and coverage with researchers – a researcher-led approach was felt to be more sustainable

TIP
Framing data interviews in terms of information asked for by Research Councils in data management plans will provide a tangible benefit to participating researchers – they know what to write on the next grant proposal.

HOW TO USE DAF?

DAF methodology

The DAF model suggests an incremental, four-step approach to undertaking data surveys. These stages can be applied flexibly to suit the specific context and needs. Depending on the survey aims you may wish to focus efforts more in one area than another, or conduct the main stages in a different order.

- Stage 1 is for planning, defining the purpose and scope of the survey and conducting preliminary research.
- Stage 2 is about identifying what data assets exist and classifying them to determine where to focus efforts for more in-depth analysis.
- Stage 3 is where the information life cycle is considered to understand researchers' workflows and identify weaknesses in data creation and curation practices.
- Stage 4 pulls together the information collected and provides recommendations for improving data management.

TIP

An internal researcher e.g. a PhD student, may be well-placed to undertake the survey. They often have access to the data, have a disciplinary understanding and know the researchers.

Practical implementation

The DAF methodology is designed to be flexible so you can customise the approach. The pilot studies did this in a number of ways, for example:

- *Tailoring the metadata collected*
  Additional details from the extended metadata set, such as data size, retention period and desired retrieval speeds, may be important, for example, if you're planning to develop a repository. Similarly, subject-specific metadata may be called for. Defining the purpose of the survey early on is crucial as this will help to define what information you need to collect to fulfil your aims.

- *Focusing more on the assessment stage rather than creating inventories*
  Many HEIs are at the early stages of developing infrastructure for data curation, so scoping requirements through interviews with researchers may be more appropriate at first than creating inventories to take stock of data holdings. Identifying data was found to be more valuable in cases where repositories were in place or planned, as the information could be used to prompt deposit. Users may wish to invert stages 2 & 3 to complete data interviews as a requirements gathering / scoping exercise, before going on to identify data.

- *Rethinking classification*
  Some pilots encountered challenges when classifying data as the process is based on value judgements. The criteria, however, do not need to be expressed in terms of value. The classification could be more directly linked to the survey aim (e.g. basing it on the potential for deposit in a repository building exercise) or reflect other approaches such as risk-analysis.

The four stage descriptions that follow focus on practical implementation lessons noted by the pilot studies. More detailed descriptions of each survey stage with activities that may be relevant are available in the DAF methodology.

PLANNING THE SURVEY: STAGE 1

Planning

Preparing as much as possible in advance helps to make sure the data survey runs smoothly. One key aspect to cover is when the survey should take place – arrange a convenient time for the survey so it's easy for people to participate. Pilot studies found that participation was affected by annual leave, exam board meetings, fieldwork and other major commitments.

Elapsed time between meetings, questionnaires and interviews can be significant. Wherever possible conduct background research in advance and set up appointments to speak with researchers early on in the process to ensure the survey runs smoothly. Collecting information is very time consuming, particularly in interviews, so having a clear aim and tight focus is crucial.

Defining aims

Being clear about the aims of the data survey from the outset helps to define the scope. It also means you can provide clarity for researchers about what will be achieved and the benefits of taking part. There were many aims behind pilot project data surveys, including:

- Scoping researchers' requirements to inform the development of new systems
- Performing service gap analysis to see where services should be developed / brought together
- Capacity planning exercises to inform future storage needs
- Responding to identified issues e.g. improving archiving workflow

The information you decide to collect and the approach you adopt to do so will vary according to the survey's underlying aims. It may be worth looking at the customisations on page 4 to consider how you will tweak the approach to meet your context and researcher needs.

TIP
An initial meeting with researchers or the Head of Department is a useful way to find out what they want and to set the aims and scope of the survey to meet this.

Setting the scope

With so much data being created and used by researchers, the pilot projects found it crucial to scope surveys tightly to ensure it was feasible to meet survey aims. Some approaches used that may be of help are:

- Limiting the time period being covered e.g. only data from the last three years
- Excluding certain types of data e.g. forensic archaeology data due to sensitivities
- Focusing on certain research groups or staff e.g. full-time academics not fellows
- Selecting examples of each type of data or project
- Working with projects at different stages of the lifecycle
- Snowball sampling e.g. interviewing research group leaders then others as directed

You may wish to review the scope mid-survey in light of how the information exercise is progressing or new needs that arise, so it is useful to adopt a flexible approach.

INFORMATION COLLECTING EXERCISE – STAGES 2 & 3

From the aim and scope defined at the planning stage you should have a clear idea about the kind of information you want to collect. The next phase of the survey is to undertake this work. This is covered by stages 2 and 3. These stages overlap significantly so they'll be discussed collectively here. Some pilots found it useful to run them concurrently.

Collecting information

There are various ways to collect information in data surveys. The pilot studies found a combination of approaches worked best. Questionnaires were found to be the most useful means of collecting basic contributions from a wide range of stakeholders, while interviews were useful for more detailed, qualitative information on data management and user needs.

TIP
Questionnaires are a useful way to identify researchers willing to participate in more detailed interviews.

Desk-based research	
✓ Good to collate background information ✓ Research articles often provided details of how the data were created	× Remote access to data may not be granted × Hard to understand local filing / naming systems
Questionnaires or wiki for researchers to fill in	
✓ Good for collecting basic overview ✓ Allows wide participation ✓ Wiki approach lets researchers adapt survey to add fields relevant to them	× Response rate can be low due to survey fatigue – best if pushed by internal advocate × Need to make sure software meets your needs – Bristol Online Surveys found to work best
Interviews	
✓ Provide high quality information ✓ Can develop questions to tease out points ✓ Help gauge awareness of data issues	× Requires significant input from researchers × Can be hard to schedule × Very time consuming – can be useful to have two surveyors: one to interview, one to note-take

The DAF online tool is a place where information collected on data assets can be stored and shared. It mirrors the four stage approach and provides survey forms for completion. These could be filled in by the survey organisers or participating researchers – multiple logins with different levels of access can be assigned. Basic analysis tools are provided, and import and export facilities are planned.

Creating an inventory of data

Stage 2 is intended as a quick mapping exercise to get a feel for the volume and types of data being held. Pilots found it was not feasible to be comprehensive, so scoping decisions are key. Suggested metadata fields focus on identifying data (name and description), recording the location of data and/or information about them (reference), and noting who - if anyone - is responsible for managing the data (asset manager). The other aspect covered here is a classification to restrict the amount of data being considered in greater detail at the next stage. Some pilots found it useful to collect additional information early on, as seen in the example questionnaires at the end of the guide.

Before completing the inventory you'll need to define what you mean by *data assets*. Will this include software, non-digital items such as lab notebooks that are integral for interpretation, or third-party data for which you do not have curatorial control? The definition has varied in pilot surveys depending on what is important for the discipline being surveyed. Most have viewed 'data' as encompassing:

> *numerical data, statistics, output from experimental equipment, survey results, interview transcripts, databases, images or audiovisual files, amongst other things*

The level of granularity to adopt also needs to be defined. Are assets recorded as single files, as datasets, or in terms of collections or projects? The answer will vary in each case depending on the scope set to ensure that meeting the survey aims is feasible. Pilots did not always encounter well-documented, homogenous datasets - often there were just ad hoc collections of data and resources used to support particular research, which could be difficult to interpret or group.

Ensuring the sustainability of the inventory was raised as a concern in some pilot studies. The inventory could be used to start active data management, for example as a prompt to deposit in repositories. Embedding the inventory in the work of the department so it becomes a local data tracking and management tool was suggested. Using wikis and encouraging a researcher-led survey, as was done with climate change researchers at Monash University, could be a useful way to achieve this.

Classifying data

The DAF methodology suggests data is classified in order to restrict the scope when moving from the wide, shallow inventory to the more detailed assessment of data management. In practice this was not always feasible or necessary; pilot studies found inventories were often representative samples rather than comprehensive registers, so the scope did not need to be restricted.

Some criteria for classifying data suggested by pilots that may be of use were:

- National Science Foundation data categories[1]
- Risk-assessment e.g. data most at risk of loss or cases with penalties for misuse
- Institutional responsibility e.g. not third party data where HEI does not have curatorial control
- Potential return e.g. if data could be ingested into repository for data sharing

Data interviews

Interviews were found to be the best way to elicit information on how data are being managed as there was an opportunity to build rapport with the researcher and gauge their understanding of potential issues. Simply asking pertinent questions was found to be useful to raise awareness of good curation practice, as it made people reflect on their approach to data management more critically.

TIP

The data lifecycle model was found to be a useful way to frame interviews, as activities were familiar to researchers. It helped to introduce the range of stakeholders too. (see p1)

Interviews were sometimes used to gather all the information for stages 2 and 3 at once. One approach used was:

1. going through the interviewee's personal drives (and, where appropriate, shared drives) to determine which collections of data constituted data assets;
2. recording names, descriptions, statements of responsibility and locations;
3. discussing the importance of the asset in terms of current and future research;
4. recording additional information about file formats, software requirements, derived reports/ papers, dates of creation and update, etc.;
5. discussing how the interviewee managed the data.

Example interview frameworks used by the pilot studies are available at the end of the guide.

[1] NSF, *Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century*, Appendix D. *Digital Data Collections by Categories*. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540_11.pdf

STAGE 4 / NEXT STEPS

Reporting

The final stage is to collate findings and report back with recommendations on how data management practices could be improved. The findings from the pilot implementations were broadly aligned and echoed the conclusions of other case studies conducted recently, such as those done by the UKRDS. As the data landscape was found to be common across institutions and research areas, the issues and recommendations below may help inform your survey. More discussion on findings from early DAF audits can be found in an IJDC paper.[2]

Common data issues faced include:

- Poor naming and filing systems so retrieval is a challenge;
- Lack of storage space meaning researchers revert to using external hard drives and laptops;
- No active transfer of data on staff retirement / departure meaning legacy material is lost, mismanaged or remains on the server unused;
- Limited data archiving facilities, so researchers often have to maintain their research outputs;
- Growing requirements e.g. from publishers and RCs that researcher feel ill-equipped to meet.

Recommendations made in pilot studies included:

- Guidance on creating data and metadata/documentation to enable retrieval and reuse;
- Training and advisory support to help researchers adopt best practice through the lifecycle;
- Assistance with composing data management plans and carrying out suggested actions;
- Implementing data policies that clarify roles and responsibilities;
- Support on selecting data for the long-term so only that which is needed is kept;
- Additional storage when capacity is insufficient or to support different needs i.e. active data store and offline storage for archiving.

Where to go next?

Further information on using DAF is offered through the DCC. We can provide guidance on how the methodology has been implemented in different contexts and run training courses for those wishing to conduct data surveys. If you would like more information, get in touch with us at: info@dcc.ac.uk

Example questionnaires and interview frameworks from DAF exemplars are provided in the pages that follow. If you would like to repurpose these please acknowledge the source institutions as noted.

Additional information on the lessons learned by the DAF exemplars, for which these questionnaires and interviews were developed, is available in their final project reports available on the DAF website.

[2] Ball, Alexander, Ekmekcioglu, Cuna & Jones, Sarah, 'The Data Audit Framework: a first step in the data management challenge' in the *International Journal of Digital Curation*, Vol 3, No.2, 2008, available at: http://www.ijdc.net/index.php/ijdc/article/viewFile/91/62

PRACTICAL EXAMPLES

Edinburgh Data Audit Implementation: Online questionnaire

Section I. Personal details

1. Full name
2. Academic role
3. Research group or research active area

Section II. Details of your research and research data

Please describe here your most recent research project and provide information on the data generated or used in this research project by answering the questions below.

4. Project details
5. Description of the data
6. Ownership: who owns the data?
7. Characteristics of the data (select all that apply)
   a. Observational
   b. Experimental
   c. Reference
   d. Derived
   e. Simulated
   f. Not Applicable
8. Data types (select all that apply)
   a. Data automatically generated from or by computer programs
   b. Data collected from sensors or instruments (including questionnaires)
   c. Images, scans or x-rays
   d. Websites
   e. MS Word files
   f. Excel sheets
   g. SPSS files
   h. Digital audio files
   i. Digital video files
   j. Fieldwork data
   k. Laboratory notes
   l. Photo collection
   m. Video tapes
   n. Audio tapes
   o. Slides
   p. Text corpus
   q. Documents or reports
   r. Patient records
   s. Other *(please specify)*:
9. Size of the data
10. Importance of the data
    a. Vital
    b. Important
    c. Vital

11. Retention period
a. Only over the project period
b. Up to 5 years
c. Up to 10 years
d. More than 10 years
e. Don't know

12. How frequently do you update your data over the project period?
a. Never
b. Daily
c. Weekly
d. Monthly
e. Annually
f. Don't know

13. Is your data backed up regularly?
a. Yes
b. No
c. Don't know

If it is, where is it backed up? (Select all that apply)
a. School server
b. Storage area network
c. DVDs
d. CDs
e. USB/Flash drives
f. External hard drives
g. Tapes
h. Third party (including commercial data storage)
i. Don't know
j. Other *(please specify)*:

14. Do you currently have a formal Research Data Management Plan in place in your school/centre?
a. Yes
b. No
c. Don't know

15. Who is currently responsible for managing the data? (select all that apply)
a. Research project manager
b. Designated person on project
c. External project partners
d. IT staff within your school, centre or research institute
e. Research assistant
f. Yourself
g. National data centre or data archive
h. Nobody
i. Don't know
j. Other *(please specify)*:

Imperial College London DAF Survey Questionnaire

Welcome Note
The purpose of this survey is to build a better understanding of research data held in your Department, to inform strategic planning for data management at Imperial College and to help inform the wider UK data management community.

This survey consists of 15 questions split across two pages and should take no more than 20 minutes to complete.

To start the survey, click the continue button below.

Section/Question	Mandatory Y/N	Available Responses
About You		
1. Please enter your full name	N	Free text
2. Please confirm your research role	Y	• Senior Researcher • Principal Investigator • Research Assistant • Research Technician • Research Support • Research Student • Other (please specify)

General Data Management For the purpose of this section you should consider the term 'electronic research data' to include all data associated with your projects - this may include numerical data produced by computational experiments, output from experimental equipment, images or audio created from experimental data or data gathered as part of the project or even data collected from surveys relating to the project.		
3. Which of the following categories best describes the electronic data created in your field of research?	Y (Multi Answer)	• Experimental • Simulated • Observational • Derived • Reference • Other (please specify)
4. In what way is your electronic research data important to your Research Group or Department?	N	Free text
5. Please estimate how much electronic research data you currently hold/maintain?	Y	• < 1 GB • 1 - 50 GB • 50 - 100 GB • 100- 500 GB • 500 GB - 1 TB • 1 - 50 TB's • 50 - 100 TB's • > 100 TB's • Don't know

6. Who is responsible for managing your electronic research data?	Y (Multi answer)	• Yourself • Research Project Manager • Research Assistant • Research Technician • PhD Student • Other Designated person in Research Group • Departmental IT Officer • Central ICT • Local Data Centre • National data centre / data archive • International data centre / data archive • Don't know • No one • Other (please specify)
7. Please confirm where your electronic research data is primarily stored?	Y (Multi answer)	• Hard disk drive of instrument/sensor which generates data • Hard disk drive of PC • External hard drive • Local server • ICT server • Third party • CD/DVD • USB/Flash drive • Other (please specify)
8. Is your data backed up regularly?	Y	• Yes • No • Don't know
a. If yes, how frequently is it backed up?	Y -subject to 8	• Daily • Weekly • Monthly • Ad hoc • Don't know • Other (please specify)
b. What data tends to be backed up?	Y - subject to 8	• Everything • Data critical to project • Data required for publication • Don't know
c. Where is it backed up?	Y (multi answer) - subject to 8	• CD / DVD • USB/Flash Drive • External Hard Drive • Tape • Local server • ICT server • Third party • Other (please specify)
9. Do you currently have a data management plan for your research data (for example, data preservation policy, record management policy, data disposal strategy)?	Y	• Yes • No

a. If yes, what was the main driver for developing your strategy?	Y (multi answer) - subject to 9	• Research requirement to access/analyse/annotate others' data • Requirement of project funder • Size of project team (i.e. multiple data creators) • Volume of data associated with project • Complexity of data associated with project (e.g. multiple formats) • Absence of College data management policy • Other (please specify)
b. If no, please confirm why	Y (multi answer) - subject to 9	• Not required / appropriate to field of research or research group • Not required by project funder • Time and effort required • Lack of training / expertise within research group • Lack of local support / guidance (e.g. Central Library, ICT) • Absence of College data management policy • Don't know • Other (please specify)
10. Do you currently allow others to access your research data?	Y	• Yes • No
a. If yes, who to?	Y (multi answer) – subject to 10	• Students / Colleagues in Department • Students / Colleagues within Imperial • Research Group • Other Institutions • As supporting evidence to publication • General public • Other (please specify)
b. If no, what access issues are of concern to you?	Y (multi answer) – subject to 10	• Confidentiality /IPR • Commercial value of data • Possible misinterpretation of data • Time/effort required • Other (please specify)
11. Have you ever been asked to make your electronic research data openly available outside of a publication (e.g. required by project funder)?	Y	• Yes • No
a. If yes, please supply high level details	Y– subject to 11	Free text

Your Data Assets
In this section we would like you to provide details of electronic research data you consider critical to your own work or that of your Research Group/Department.

For example, if a Research Council were to ask you to safeguard your data for future re-use or if you were to leave College, what data should be preserved? Alternatively, please provide examples of data which you consider critical to your own work or that of your Research Group/Department. Including datasets and information systems that:

- are still being created or added to;
- are used on frequent basis in the course of your work;
- underpin scientific replication e.g. revalidation;
- play a pivotal role in ongoing research;
- or are being used to provide services to external clients and partners

12. Please provide the following high level information for each data asset

Question	Mandatory Y/N	Available Responses
a) Brief Description	N	Free text
b) Principle Data Type	N	• Raw data generated by program • Raw data from instrument • Images, scans or x-rays • Digital audio • Digital video • Database (e.g. MySQL, Oracle) • Text document (e.g. Word, PDF) • Spreadsheet (e.g. Excel) • Other proprietary format • Software • Lab notes • Patient data • Other
c) Effort Associated with Creation of data	N	• Hours • Days • Weeks • Months • Years • Other
d) Planned Retention Period	N	• < 1 year • 1 - 2 years • 2 - 5 years • 5 - 10 years • 10 - 20 years • 20 - 100 years • 100+ years • Indefinitely • Don't know • Other

e) Frequency of Use	N	• Daily • Weekly • Monthly • Yearly • Reference Only • Other
f) Estimated 'final' Size of Data	N	• < 1 GB • 1 - 50 GB • 50 - 100 GB • 100- 500 GB • 500 GB - 1 TB • Multiple TB's

13. Are any of your 'critical' research data assets stored in a proprietary format?	N	• Yes • No
a. If yes, please confirm format	Y – subject to 13	Free text
b. Please also confirm what sort of services you would like to see offered by College to guarantee access to this data in the future?	Y – subject to 13	Free text
14. Please confirm if you would be willing to participate in a short follow-up interview to this survey?	Y	• Yes • No
a. If yes - and you have not already done so in question 1 - please supply your name so that we can contact you.	Y – subject to 14	Free text
15. Do you have any comments regarding this survey?	N	Free text

Final Page

Thank you for completing this survey, your contribution is very much appreciated.

If you have any questions relating to this survey or if you would like to contribute to the formation of a research data management strategy, please click the 'Contact Us' button at the bottom of this screen.

University of Southampton Questionnaire

You may re-use or adapt this documentation for research or private study with acknowledgement to **McGowan, T. & Gibbs, T. A. (2009) Southampton Data Survey: Our Experiences & Lessons Learned [unpublished]. University of Southampton: UK.**

Welcome to the Research Data Survey questionnaire

Thank you for participating in this survey which aims to find out about research data held by staff in the School of Social Sciences and improve our understanding of the data management processes you employ.

For the purpose of this study 'research data' is data that you currently hold that has been collected and/or used in the course of your research at the University of Southampton. Research data can be primary data collected by you or your research group or secondary data provided by a third party. It may be quantitative or qualitative e.g. survey results, interview transcripts, databases compiled from documentary sources, images or audiovisual files.

Data that you 'currently hold' is all the research data that you currently store anywhere. For example, in your 'My Documents' folder, on the shared 'R' drive, a PC or laptop, on portable media such as CDs or memory sticks, or on paper.

It would help us greatly if you respond to this questionnaire even if you do not currently hold any research data (you will only be required to answer 2 questions).

The questionnaire is a maximum of 25 questions and should take no more than 10 minutes to complete.

Thank you for your time.

Participant consent form

Please read the following statements carefully before agreeing to take part in this study;

I have read and understood the participant information sheet (attached to the email in which you received this link).
I understand that;

- All results from this study will be anonymous. Information extracted from this questionnaire and any subsequent interview will not, under any circumstances, contain names or identifying characteristics of participants.
- I am free to withdraw from this study at any time without penalty.
- I am free to decline to answer particular questions.
- Whether I participate or not there will be no effect on my progress in employment in any way.

I consent to take part in this study on the terms described above; ○ Yes ○ No

[IF NO, GO TO END]

1. Do you currently hold any research data?

- ○ Yes
- ○ No **[GO TO END]**

2. Thinking about the primary data you hold, what type of data is it? *[Please select all that apply]*

- ☐ I don't hold any primary data **[GO TO QUESTION 4]**
- ☐ Cross sectional survey data
- ☐ Longitudinal survey data
- ☐ Interview/focus group transcripts
- ☐ Database compiled from documentary sources
- ☐ Image files
- ☐ Audio files
- ☐ Audio-visual files
- ☐ Other

If other, please specify

3. Who funded the collection of the primary data you hold? *[Please select all that apply]*

- ☐ ESRC
- ☐ EU-EDULINK
- ☐ Leverhulme Trust
- ☐ Nuffield Foundation
- ☐ UK Government department
- ☐ Wellcome Trust
- ☐ Other

If other, please specify

4. Thinking about the secondary data you hold, who collected this data? *[Please select all that apply]*

- ☐ I don't hold any secondary data **[GO TO QUESTION 6]**
- ☐ Datastream (Thomson Reuters)
- ☐ Eurostat
- ☐ International Labour Organization (ILO)
- ☐ Measure DHS
- ☐ Organisation for Economic Co-operation and Development (OECD)
- ☐ Office for National Statistics (ONS)
- ☐ US Census Bureau
- ☐ World Bank
- ☐ World Health Organization (WHO)
- ☐ Other

If other, please specify

5. What type of secondary data is it? *[Please select all that apply]*

- ☐ Cross sectional survey data
- ☐ Longitudinal survey data
- ☐ Interview/focus group transcripts
- ☐ Database compiled from documentary sources
- ☐ Image files
- ☐ Audio files
- ☐ Audio-visual files
- ☐ Macro-economic time series data
- ☐ Stock market data
- ☐ Company level data
- ☐ Other

If other, please specify

6. *The remaining questions relate to all the data you currently hold, both primary and secondary;*

When using or creating this data, did you collaborate with anyone else?

- ○ Yes
- ○ No **[GO TO QUESTION 9]**

7. How did you share data when you were collaborating? *[Please select all that apply]*

- ☐ By emailing files to colleagues
- ☐ Using a shared storage facility
- ☐ Using portable storage such as CDs, DVDs, memory sticks etc
- ☐ Other

If other, please specify

8. Did you encounter any practical problems when you were collaborating? *[Please select all that apply]*

- ☐ No
- ☐ Finding suitable shared storage space
- ☐ Lack of file naming conventions made it difficult to identify files
- ☐ Lack of version control caused confusion
- ☐ Legal issues arising from international transfer of data
- ☐ Problems establishing ownership of data
- ☐ Other

If other, please specify

9. Where do you store your data (excluding back up copies)? *[Please select all that apply]*

- ☐ On paper
- ☐ My Documents
- ☐ Shared drive (R-drive)
- ☐ Hard drive of office PC
- ☐ Hard drive of laptop PC
- ☐ Memory stick/USB/Flash drive
- ☐ CD/DVD
- ☐ External hard drive
- ☐ Other

If other, please specify

10. Have you ever experienced any problems storing your research data due to the size of the files?

- ○ Yes
- ○ No **[GO TO QUESTION 12]**

what problems

11. How did you overcome these storage problems? *[Please select all that apply]*

- ☐ Requested additional storage space from iSolutions
- ☐ Purchased an external hard drive
- ☐ Saved to portable media
- ☐ Other

If other, please specify

12. Is the data that you currently hold backed up anywhere?

- ○ Yes, all of it is
- ○ Yes, some of it is
- ○ No, none of it is **[GO TO QUESTION 14]**

13. Where do you back up your data?

- ☐ On paper
- ☐ My Documents
- ☐ Shared drive (R drive)
- ☐ Hard drive of office PC
- ☐ Hard drive of laptop PC
- ☐ Memory stick/USB/Flash drive
- ☐ CD/DVD
- ☐ External hard drive
- ☐ Other

If other, please specify;

14. Do you deposit your data with a data service, such as the UK Data Archive?

- ○ Yes, all of it **[GO TO QUESTION 20]**
- ○ Yes, some of it
- ○ No, none of it

If yes, please tell us which service/s you deposit with;

15. Do you think that any of your data needs to be preserved by the University for your own use or that of others?

- ○ Yes
- ○ No **[GO TO QUESTION 17]**

16. If you would like someone from the University Library to contact you about preserving your data please enter your name below;

17. Thinking about your data that is <u>not</u> deposited with a data service, could any of this data be re-used by others?

- ○ Yes, all of it could be re-used **[GO TO QUESTION 20]**
- ○ Yes, some of it could be re-used
- ○ No, none of it could be re-used

18. Thinking about your data that can't be re-used or shared, please tell us why *[Please select all that apply]*;

- ☐ Confidentiality or data protection issues
- ☐ Licence agreements prohibit sharing
- ☐ The data is not fully documented
- ☐ The data is in a format that is no longer widely readable **[IF SELECTED GO TO QUESTION 19, OTHERWISE GO TO QUESTION 20]**
- ☐ Other

If other, please specify;

19. Please provide brief details of the data you have that is no longer widely readable (e.g. what software/hardware the data is on, its age etc);

20. Would you like to receive any additional support with managing your data? *[Please select all that apply]*

- ☐ Training
- ☐ Written guidance
- ☐ Help with writing data management plans for research bids
- ☐ Additional personal storage
- ☐ Additional shared storage
- ☐ None
- ☐ Other

If other, please specify

21. Which Division do you work in?

- ○ Economics
- ○ Gerontology
- ○ Politics and International Relations
- ○ Sociology and Social Policy
- ○ Social Statistics
- ○ Social Work Studies

22. Would you be prepared to participate in a follow up interview to explore data management issues in more depth (max. 1 hr)?

- ○ Yes
- ○ No

If yes, please provide your name and email address so that we can contact you;

23. If you would like to expand on any of your above answers or make further comment, please do so here;

University of Southampton Generic Interview Schedule

Introduction

INTRODUCE

My name is Teresa McGowan and this is **Harry Gibbs.** Harry is the School of Social Sciences **librarian** and I am a **research assistant** here in the School.

RESEARCH

We are working together on a **project funded by** the Joint Information Systems Committee (JISC). **JISC has developed a framework methodology** aimed at helping institutions find out what research data they hold, where it's located and who is responsible for it. **We are using** an adaptation of that framework today to test its usefulness and to help the School of Social Sciences find out more about data management and what can be done to aid staff in the use and management of their data.

THANKS

Thank you for agreeing to take part in this interview. Based on the information that we receive we will produce **two reports,** one for JISC simply discussing how we used and modified their framework, and a second for the University which we hope will be used to improve data management in the school.

RECORDING AND CONFIDENTIALITY

We would like to **record our discussion** as it is so difficult to write down everything that is said, and we don't want to miss anything. What you say in this interview will be **anonymous** – your names will not be recorded on the transcripts and only **me, Harry and one transcriber** will have access to the recording and notes. **No reports or publications** that are produced will identify you in any way.

WANT TO KNOW

Thank-you for taking part in the questionnaire, **the purpose** of this interview is to find out more about **the data you hold** that has been collected or used in the course of your research at this University and **your experience of managing** this data. **There are no right or wrong answers, we are just interested in what you have done and how you did it.**

We want this to be more like a **discussion** than a question and answer session. We have a list of **x** things we are interested in but it is important to us that you tell us about what is important to you. If there is anything I ask that **you don't understand** please tell us and we can explain further. If there is **anything you want to ask us** you can do that too. *(If they ask questions that anticipate later discussions, ask if it's OK to leave it until later)*

Do we have your permission to proceed?
(Record some thoughts about participants, body language etc)

Discussion Guide

1 Could you please tell us a bit about your area of research?

2 We can see from the questionnaire that you hold xxx data, please could you give us some more details about the xxx data that you compiled from documents?

Data Holdings		
Name of Interviewee	**Primary Data**	**Secondary Data**
[interviewee's name]	[list of primary data types held by individual, as reported in questionnaire]	[list of secondary data types held by individual, as reported in questionnaire]

Tick	Metadata Heading	Notes
	ID	A unique identification assigned by the research team
	Author	Person, group or organisation responsible for the intellectual content of the dataset
	Owner(s)	Current legal owner(s) of the dataset
	Source	The source(s) of the information found in the data asset
	Purpose	Reason why the asset was created, intended user communities or source of funding / original project title
	Title	Official name of the data asset, with additional or alternative titles or acronyms if they exist
	Description	A description of the information contained in the data asset and its spatial, temporal or subject coverage
	Subject	Data topics and keywords describing the subject matter of the data
	Geographical coverage	The countries, regions, cities etc covered in the data
	Time period covered	The date (or date range) covered by the data
	Date of collection	The date (or date range) on which the data was collected

		(for social surveys this will often be the same as the time period covered)
	Sample size & description	The number of individuals surveyed and characteristics
	Current location	Path or www address where the data can be found
	Format	Physical formats of dataset, including file format information
	Size	Size of the data in Mb/Gb
	Restrictions	Access restrictions placed on user of secondary data or restrictions owner would place on reuse of primary data
	Documentation available	Documentation that is available (e.g. user manuals, code books), including references to its location
	Retention period	Planned retention period for the data & ideal retention period

3 As I am sure you are aware, increasingly funding bodies want researchers to include a data management or data sharing plan in the funding application. Have you ever experienced this?

OR

TO THOSE WHO ASKED FOR HELP WRITING DATA MANAGEMENT PLANS:

3 Can you tell us about your experience of data management or data sharing plans?

Prompts		Tick
YES	**Which funder?**	
	How did you find this experience?	
	How did it influence your actual data management/data sharing?	
NO	**What do you think about in terms of data at the bid writing stage?**	
	How far did planning influence your actual practice? (Did it go to plan?)	

4 TO COLLABORATORS:

In the questionnaire you mentioned you had some problems collaborating, could you tell us some more about this?
OR
Please can you tell us about one experience of collaborating?

Problems collaborating	
[name]	[list of problems reported in questionnaire]

	Prompts	Tick
Who, what where?	Who with?	
	Where were they geographically?	
	How many of you were sharing?	
Sharing methods;	What did you do?	
	What methods did you use to share?	
How did you deal with;	version control	
	file naming conventions	
	legal issues transferring data	
	How was ownership decided?	
Confidential data?	Have you ever shared confidential data?	
	How did you do it?	

OR

4 TO NON COLLABORATORS:

How do you deal with the day-to-day management of data?

Prompts	Tick
Version control	
File naming system	

5 BACK UP

You told us you use xxx methods to store and back up data, can you tell us why you chose these methods?

Storage location		
	Main	**Back up**
[name]	[list of main storage methods reported in questionnaire]	[list of back up methods reported in questionnaire]

	Prompts	**Tick**
What affects your choices?	**Anticipated lifespan**	
	Importance	
	Confidentiality	
	Physical space	
Have you ever experienced any problems with;	**Data loss**	
	Old formats	
	File size	

AND

TO THOSE WHO DO NOT BACK UP ALL DATA

You said you don't back up all your data, can you explain to us why you don't?

6 **You mentioned that you had data storage problems and [list of problems reported in questionnaire], can you tell us a bit more about what happened?**

	Tick
Was the unforeseen expense a problem for the project?	
Did the problem affect the project due to lost time?	

7 **In the questionnaire you said you did not have anything that should be preserved by the university, do you have anything that you think should be preserved by yourself or anyone else?**

	Prompts	Tick
YES Could you tell us about it?	How could it be best preserved?	
	Why should it be preserved?	
	Who should preserve it?	
	For how long?	
NO Why doesn't it need preserving?	Already preserved? (BY UKDA)	
	Data not reusable [why?]	
	Time	
	Money	

OR

7 **You mentioned in the questionnaire that you have something that you think the university should preserve for the future, could you tell us about it now?**

Prompts	Tick
How could it be best preserved?	
Why should it be preserved?	
Who should preserve it?	
For how long?	

8 **What support is available to you to help you manage your data?**

OR

8 **In the questionnaire you said you would like some additional support in carrying out data management, what support is available to you now?**

Prompt	Tick
How sufficient is it?	
How would you like data management support to look in an ideal world?	

9 ***(Summarise what's been said, then;)*** **Is there anything you can think of I haven't asked or anything you wanted to say that has not been covered?**

INVITE HARRY TO ASK ANY QUESTIONS ON ANYTHING SHE WOULD LIKE TO FOLLOW UP/CLARIFY

Closure

Thank-you for allowing us to talk to you today, it has been very interesting to listen to your views. We will email you to let you know when the results are available.

Thanks again for coming today, we are very grateful for your help.

University of Oxford Interview Framework

The following framework is based on the interview frameworks developed for the IBVRE[3] and eIUS[4] projects with some changes to adjust it to the aim and objectives of this scoping study.

Introduction

Give brief introduction to the Scoping Digital Repositories Services for Research Data Management including overall aim and objectives. Provide an overview of the questions that will follow and remind the interviewee about the nature of the semi-structured interview, the intention of taking notes, record the interview (with permission) and to publish findings.

Interview

1. Could you briefly explain your area of research and the types of research questions, with examples, that you try to answer?

2. I am interested in learning more about the research tasks that involve some form of data management that you carry out in order to help you move forward with your research agenda. I'm interested in doing this by going through one of your research projects in the context of a generic "research life-cycle", from funding application, data collection/processing, all the way to publishing, in order to understand to what extent the following elements fit in your average working day.

a. The funding application – increasingly funding agencies require data management and data sharing plans as part of the funding application.
    - When applying for funding how do you decide that new data will need to be collected and how do you go about providing a plan for this?

With this question I want to learn more about how researchers think about data at this stage, why they decide that data needs to be collected, how they ensure that this data has not been created already and how they go about making data management plans.

[3] The integrative biology virtual research environment project :http://www.vre.ox.ac.uk/ibvre/
[4] The e-Infrastructure use cases and service usage models http://www.eius.ac.uk/

b. Data collection –

- Could you please explain what sorts of data (primary, secondary, experimental, simulation) you collect and provide details about the process of collection?

In this part my aim is to engage in conversation to find out about data collection methods, types of data produced, the instruments and software used to do this and whether the data could be helpful to others. I will also ask about secondary data to find out where and how are found and accessed. Finally I will explore why the collection of data happens in the way described (is it a discipline or departmental common practice?)

c. Processing of data –

- Once the data have been collected could you describe how they get processed i.e. how they get annotated, where are they stored, what security measures are taken to preserve confidentiality or integrity, etc?

Here I want to make sure that I understand how annotation/storage/back-up/manipulation/analysis/collaboration happens. Again, I will explore why the processing of data happens in the way described (is it discipline or departmental common practice?)

d. Publishing – the publication of the research outputs is the end of this generic "research life-cycle", what happens with the data after this i.e. they get published or deposited somewhere, you need to destroy the data, etc?

In this part of the life cycle I want to find out whether deposit in an archive occurs and if not I will attempt to find out the reasons that stop researchers doing so (data needs to be destroyed, does not want to share initially or at all, no place to deposit, etc) and where will the data be stored.

3. How are researchers supported either at local or institutional level for carrying out all the management of data required?

With this question I will attempt to figure out how support for data management across the generic life cycle occurs (researchers help each other at local level, departmental guidelines, etc).

4. What are your challenges and worries when managing research data and what services would help you do this work more effectively?

With this question I will attempt to get a top 3 requirements for services that would be most useful to researchers.

5. Is there anything else that you would like to add?

De-Brief

6. How do you think the interview went?

7. What are the benefits you believe you get from participating?

8. Could you suggest anyone you know that could participate in these interviews?

University of Glasgow digital preservation study[5]: interview template

It is expected interviews will take between 30 mins - 1 hour. Ideally these would be recorded then transcribed, with the text sent back for approval. An overview of the topics to be discussed will be circulated in advance to allow the interviewee to prepare ideas.

At the start of the interview, details of the preservation study and explanation of terms will be provided. Scoping interviews will be semi-structured to allow free-flowing discussion. The questions provided below are indicative of the topics that may be discussed. Each interview will cover five themes:

1. what digital material is being created;
2. how this is being created and maintained;
3. any issues that have been encountered;
4. the future for the unit's electronic records;
5. requirements for support and services.

Scope of digital holdings

A general discussion will begin by asking interviewees to describe their day-to-day work with regard to electronic records i.e. what they create and use, their attitude towards digital material, how central electronic records are to their work…

- What electronic records do you create?
- What type (docs, emails, databases) and formats are these files?
- What software do you use? Is that general to the department?
- How much digital material do you currently create / hold? Is this growing?
- Are these files yours or do they belong to a wider group or to the institution?
- Who owns the IPR of the electronic records you create?
- How crucial are these files? – could you continue work if they were lost?

Working practices

Discuss what happens in terms of digital curation i.e. creating, maintaining and preserving electronic records. Are there set procedures? What role does each person play…

Individual

- How do you create electronic records? – naming conventions, filing rules…
- Where you store files? Do you back them up or is this done centrally?
- How do you manage digital files e.g. do you sort through and weed them?
- What happens in terms of email? Do you save or print certain messages?
- Do you work differently on research projects due to funding body requirements?

Departmental

- Are there departmental guidelines, policies or procedures you follow?
- Who is responsible for digital material? What role does each person play?
- What happens in terms of legacy material i.e. files created by former staff?
- Do you know when, how and what is backed up centrally?
- Who can access electronic material? How is this controlled? Explain restrictions

[5] See: http://www.gla.ac.uk/departments/hatii/research/digitalpreservationpolicystudy

Digital preservation issues

Continue discussion to ascertain whether any issues have been encountered when creating and using electronic material to identify areas where practices could improve

- Have you ever lost digital files or found it hard to find the right ones?
- Are there version control issues when working with colleagues?
- Is it difficult to understand other people's systems on the shared drive?
- Have you struggled to use older files? e.g. obsolete format, outdated disk…
- Do you have enough storage space? If not, where do you keep material?

Future life of electronic records

Discuss what happens in the future i.e. how can these files continue to be accessed and used (if appropriate), do they need to be preserved, if so, for how long…

Access

- Could your electronic material be reused or repurposed by others?
- Are there any sensitivity or confidentiality restrictions?
- Would other people understand your material - is it documented?

Preservation

- Does all digital material or just a subset need to be preserved in the long-term?
- Who would know what to keep and for how long? Who makes the decision?
- Is there a place where your digital material can be preserved?

Service requirements

Ask where the interviewee currently gets advice and support and what else s/he would like to see provided by the University. Key thing is to gauge desire for preservation policy, suggested coverage and any supplementary support needed to implement it.

- Have you used the records management service, archive or Enlighten? Are you aware of what these services can offer?
- Where do you currently get advice and support?
- What would help you create and manage your electronic files better?
- Who should be responsible for / fund digital preservation?
- Would you welcome a University wide policy on digital preservation? If so, what should it cover?