



# Incremental

## Scoping study and implementation plan 'A pilot project for supporting research data management'

Lesley Freiman, Catharine Ward: University of Cambridge  
Sarah Jones, Laura Molloy, Kellie Snow: University of Glasgow

July 2010



University  
of Glasgow

Humanities  
Advanced Technology  
& Information Institute

Funded by:

JISC



## Preface

This report is one of the first deliverables from the *Incremental* project, which seeks to investigate and improve the research data management infrastructure at the universities of Glasgow and Cambridge and to learn lessons and develop resources of value to other institutions. Coming at the end of the project's scoping study, this report identifies the key themes and issues that emerged and proposes a set of activities to address those needs.

As its name suggests, *Incremental* deliberately adopts a stepped, pragmatic approach to supporting research data management. It recognises that solutions will vary across different departmental and institutional contexts; and that top-down, policy-driven or centralised solutions are unlikely to prove as effective as practical support delivered in a clear and timely manner where the benefits can be clearly understood and will justify any effort or resources required. The findings of the scoping study have confirmed the value of this approach and the main recommendations of this report are concerned with the development and delivery of suitable resources.

Although some differences were observed between disciplines, these seemed to be as much a feature of different organisational cultures as the nature of the research being undertaken. Our study found that there were many common issues across the groups and that the responses to these issues need not be highly technical or expensive to implement. What *is* required is that these resources employ jargon-free language and use examples of relevance to researchers and that they can be accessed easily at the point of need. There are resources already available (institutionally and externally) that can address researchers' data management needs but these are not being fully exploited. So in many cases *Incremental* will be enabling efficient and contextualised access, or tailoring resources to specific environments, rather than developing resources from scratch.

While *Incremental* will concentrate on developing, repurposing and leveraging practical resources to support researchers in their management of data, it recognises that this will be best achieved within a supportive institutional context (both in terms of policy and provision). The need for institutional support is especially evident when long-term preservation and data sharing are considered – these activities are clearly more effective and sustainable if addressed at more aggregated levels (e.g. repositories) rather than left to individual researchers or groups. So in addition to its work in developing resources, the *Incremental* project will seek to inform the development of a more comprehensive data management infrastructure at each institution. In Cambridge, this will be connected with the library's CUPID project (Cambridge University Preservation Development) and at Glasgow in conjunction with the Digital Preservation Advisory Board.

## Table of Contents

I. Executive Summary .....	4
1. Thematic findings .....	4
2. General findings .....	5
3. Implementation plan .....	6
II. Methodology .....	7
1. Requirements gathering .....	7
2. Study coverage.....	7
III. Concerns and Issues .....	10
1. Procedures for creating data .....	10
2. Data storage and access.....	11
3. Back-up .....	12
4. Preservation .....	13
5. Data sharing and re-use.....	15
IV. Existing guidance and interviewee requests .....	17
V. General findings on data management support .....	18
VI. Implementation Plan .....	19
1. Recommendations .....	19
2. Proposed tasks and activities.....	21
VII. Appendices .....	25
1. Timeline for implementation plan .....	25
2. Interview templates – Cambridge.....	26
3. Interview template – Glasgow .....	30
4. Draft costing survey.....	32

# I. Executive Summary

The Universities of Glasgow and Cambridge have completed a scoping study of key data management and preservation needs, concerns, and current practices within a wide variety of departments and research groups in both institutions.

## 1. Thematic findings

In speaking to researchers and support staff, we uncovered a wide variety of long and short-term data management concerns, which we have broken down into five data management themes:

1. Procedures for creating and organising data – Key concerns included inconsistent folder structures, versioning and naming conventions and limited/non-existent documentation. Researchers often have trouble finding and/or deciphering the files of their colleagues and students as well as their own files from previous years.
2. Data storage and access – Most departments and research groups have access to network server storage (especially at Glasgow); however, many find it insufficient or slow so some researchers choose to save data on a large variety of cheap storage media and often are not aware of the risks or benefits of each. Most researchers at Cambridge do not have remote (off-campus) access to data, while at Glasgow there is provision but many researchers are not aware of this facility.
3. Data back-up – While most networked server storage is backed-up regularly by IT staff, researchers store data on a variety of media (including personal computers, external hard drives, data sticks and e-mail) and are consequently responsible for their own back-up. Departments don't have guidelines or norms for personal back-up and researcher procedure, knowledge, and diligence varies tremendously. Many have experienced moderate to catastrophic data loss.
4. Preservation – Many researchers are concerned about platform/software obsolescence and the potential for data to be lost or destroyed. At the same time, preserving data in a repository usually entails a lot of work in terms of creating metadata and a willingness to release data publicly. Researchers are often uncertain about which formats and media are best for preserving digital data and create little documentation aside from published papers.
5. Data sharing and re-use - While many researchers are positive about sharing data in principle, they are almost universally reluctant in practice. They have invested in collecting or processing data, and using these data to publish results before anyone else is the primary way of gaining prestige in nearly all disciplines. In addition, researchers complain that data must be carefully prepared, annotated, and contextualised before they can make it public, which is all very time-consuming and funding is rarely set aside for this.

## 2. General findings

### Resources must be simple, engaging and easy to access

In speaking with researchers, we found that many were interested in guidance, simple tools, and support for data management, but this came with several caveats.

They are often unaware of existing resources and training. If they are aware of them, they often do not use these resources because they are:

- Wordy and ambiguous (teasing out answers takes too long);
- Difficult to find (or not available when needed);
- Don't feel relevant; or
- Boring.

Many researchers complained that training is often inconveniently timed and not well-tailored to their needs. Many felt that brief training, online resources, 'a really smart little leaflet' or someone to talk to face-to-face would be more helpful.

### Language matters

Many of the available data management resources include jargon and specialised language. Using clear and jargon-free language will both help us assess researcher needs and help us communicate guidance effectively (e.g. researchers don't know what 'digital curation' is and humanities researchers don't think of their manuscripts as 'data'). Additionally, using the most appropriate language to explain our goals and purposes is essential to success. Many people are suspicious of 'policies,' which sound like hollow mandates, but are receptive to 'procedures' or 'advice' which may be essentially the same thing, but convey a sense of purpose and assistance rather than requirement.

### Differences between departments

While both studies found some differences in primary needs/challenges between disciplines, there were a small number of factors that drove most of these differences:

- Resources (IT expertise and funding) within the department/research group;
- Volume of data;
- Types of data (e.g. images vs. numerical vs. documents);
- Conventions of the field (e.g. whether current data stays relevant for five years or fifty years, etc).

While disciplinary differences, *per se*, may not be the primary factor in researcher concerns, we found that researchers will only use data management resources if they feel personal and relevant. Providing some materials with discipline-specific examples, and engaging local data management 'champions' within departments may help us to achieve this.

### 3. Implementation plan

In light of researchers' wishes for simple, clear, engaging, and available guidance and support, we propose to move forward in four ways:

1. Produce simple, accessible, visual guidance on creating, storing, and managing data – This will include (1) producing a collection of webpages at each institution, pointing researchers to existing local and external resources and new resources created by the project. Examples such as the MIT pages<sup>1</sup> and University of Edinburgh advice portal<sup>2</sup> will be used as models for this; (2) producing materials including illustrated fact sheets, flow diagrams, checklists, and FAQs with solutions for common researcher concerns. Some discussions have already begun with John Cairns at Glasgow to consider an Intellectual Property Rights flow diagram along the lines of the Web2Rights work, which will form part of this.<sup>3</sup>
2. Offer practical data training with discipline-specific examples and local champions – We will work with enthusiasts within departments to embed slides and resources within existing training and inductions (i.e. training the trainer). We will also create brief online tutorials and/or screen-casts, and include case-studies from within disciplines wherever possible. Some disciplinary courses<sup>4</sup> were noted in interviews which we can look at and there are various examples of online training modules such as those produced by Cornell University.<sup>5</sup>
3. Connect researchers with support staff who offer one-to-one advice, guidance, and partnering – We will work with departments and the research office within each institution to refer researchers to existing support staff they can turn to for one-to-one advice during the proposal-writing stage of projects and beyond.
4. Work towards the development of a comprehensive data management infrastructure – We see our current work as part of an overall effort to raise awareness of the urgency of increased data management and preservation activities. In talking to departments, we have already begun to build connections towards broader infrastructure and policy within each institution and will continue to approach this goal as we move into the implementation phase of the project. This objective can be progressed further through CUPID at Cambridge and the Digital Preservation Advisory Board at Glasgow.

---

<sup>1</sup> A set of web pages that provide basic guidelines on creating and managing data. See: <http://libraries.mit.edu/guides/subjects/data-management/>

<sup>2</sup> Links and guidance in three areas: how to manage research data; data sharing and preservation; and training advice and support. See: <http://www.ed.ac.uk/schools-departments/information-services/services/research-support/data-library/research-data-mgmt>

<sup>3</sup> A flow diagram to help content creators and users define their rights in relation to IPR. Guidance and templates within the JISC IPR toolkit are pointed to at relevant points. See: [http://www.web2rights.org.uk/navigator/content/ipr/chart/IPR\\_Flowchart.pdf](http://www.web2rights.org.uk/navigator/content/ipr/chart/IPR_Flowchart.pdf)

<sup>4</sup> For example the BBSRC summer school on data management for systems biologists. See: <http://www.erasysbio.net/summerSchool>

<sup>5</sup> Digital Preservation Management tutorial, available at: <http://www.icpsr.umich.edu/dpm/dpm-eng/>

## II. Methodology

### 1. Requirements gathering

The initial months of the *Incremental* project focused efforts on requirements gathering to scope out what kind of infrastructure and support researchers needed to manage data. A study had already been undertaken at the University of Glasgow in 2009 to scope digital preservation needs,<sup>6</sup> so this approach was adapted for use in Cambridge bearing in mind the more decentralised nature of this institution.

The Glasgow approach built on lessons learned in the JISC-funded DAF projects<sup>7</sup> and involved conducting semi-structured interviews to understand how researchers create, manage and preserve digital material. The Glasgow study was broad in scope, covering research data, teaching material and administrative records. For requirements gathering at Cambridge, the focus was on research data and related records. The interview templates from both studies are available on the project websites<sup>8</sup> and in Appendices 2 and 3.

For the Glasgow study, HATII worked to cover a mix of research, teaching, administrative and technical viewpoints when selecting interviewees. Common participants included the head of department (HoD) or director, senior lecturers or heads of teaching, research group leaders/PIs of major projects, researchers, research support staff, administrators, and IT or technical staff. Cambridge took a similar approach, but focused on researchers and supporting computing staff. Common participants included project Principal Investigators ('PIs'), researchers, students and support staff.

A review of the interviewees' web page profiles was undertaken prior to the interviews in order to help interviewers pick up on specific areas during the interviews. Interviews ranged from twenty-five minutes to one hour and twenty minutes, with most taking around one hour. The interview templates were used as a guide to ensure some coverage on each of the key areas. Interviews were audio-recorded and transcribed and summarised, with the summaries returned to interviewees for comment.

### 2. Study coverage

Where possible, Cambridge project staff recruited participants from similar academic departments and research to those participating in Glasgow to allow for comparisons across disciplines as well as between the universities. Participating departments/research groups were:

---

<sup>6</sup> <http://www.gla.ac.uk/departments/hatii/research/digitalpreservationpolicystudy/>

<sup>7</sup> For information on the development project and the four pilots see: <http://data-audit.eu/>

<sup>8</sup> For the Glasgow interview template see: [http://www.gla.ac.uk/media/media\\_126658\\_en.pdf](http://www.gla.ac.uk/media/media_126658_en.pdf) and for the interview summaries used at Cambridge see: [http://www.lib.cam.ac.uk/preservation/incremental/Incremental\\_Interview\\_Summaries.pdf](http://www.lib.cam.ac.uk/preservation/incremental/Incremental_Interview_Summaries.pdf)

<b>University of Cambridge</b>	<b>University of Glasgow</b>
Archaeology	Archaeology
Chemistry	Chemistry
Engineering - Division of Mechanics, Materials & Design	Electronics and Electrical Engineering
English / Anglo-Saxon, Norse & Celtic	English Language
Public Health and Primary Care	MRC Social & Public Health Sciences Unit
Scott Polar Research Institute	Ecology & Evolutionary Biology

### Cambridge study

A total of twenty-nine semi-structured interviews were completed at Cambridge between January and May 2010, including individual and small group interviews (two to five interviewees). Prior to the first semi-structured interviews within departments, *Incremental* staff had informal discussions with one to three highly-engaged participants with each participating department in order to better understand the structures and contexts of their departments/research groups.

<b>Research Groups</b>	<b>No. of people</b>	<b>Roles*</b>
Archaeology	8	Management, research, IT
Chemistry	4	Research, IT, PhD
Engineering	11	Research, IT, PhD
English	4	Research
Public Health & Primary Care	5	Management, research, IT
Scott Polar Research Institute (SPRI)	5	Management, research

### Glasgow study

In total, twenty-seven semi-structured interviews were completed at Glasgow between March and August, 2009. These were typically on a one-to-one basis, but two were joint interviews with two members of staff and one was a group discussion with five researchers.

Additional informal discussions were held with other individuals across the University of Glasgow in light of comments made in interviews. These included meeting with staff from Enlighten (the University repository), the EDRMS project<sup>9</sup> team and a senior research fellow who advises on the University's information security strategy.

---

<sup>9</sup> Electronic Document and Records Management System Project; more information on the work of this project is available at <http://www.gla.ac.uk/services/it/projects/edrms/>

<b>Research Groups</b>	<b>No. of people</b>	<b>Roles*</b>
Archaeology	3	Management, research, admin
Chemistry	4	Management, research, teaching, IT
Electronics & Electrical Engineering (EEE)	5	Management, research, teaching, admin
English Language	6	Management, research, admin, PhDs
MRC Social & Public Health	3	Management, admin, IT
Ecology & Evolutionary Biology	7	Research (ECR), research support,

\*The roles stated denote primary responsibilities. Management interviewees were typically involved in research, and there is some overlap between researchers and IT in certain research groups.

Common themes have been drawn from the transcripts and summaries and are presented as findings in Section III of this report. All comments have been anonymised and attributed by role or department only.

### III. Concerns and Issues

#### 1. Procedures for creating data

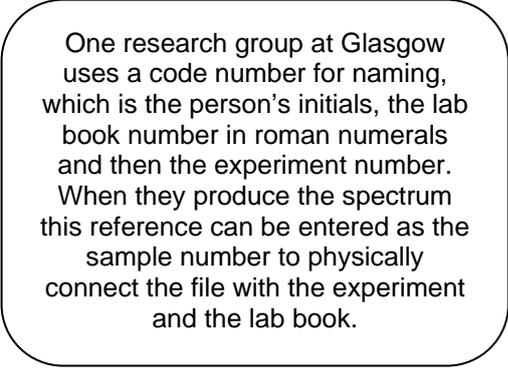
Many researchers and support staff have encountered difficulties with retrieval and re-use of data due to inconsistent file-naming, versioning and directory structures. Approaches to data creation were typically *ad hoc*, with one Glasgow researcher describing the departmental approach as a 'free-for-all'. Many interviewees acknowledged difficulties in finding and deciphering *their own* data from past years and projects.



*"the common staff network drive has always been the bane of everyone's life to find stuff on".*

Version-control issues were encountered more often when working on collaborative papers than with raw or derived datasets. Some departments, both at Glasgow and Cambridge do not have access to shared drives, which can make the situation more complicated, as there isn't a single set of files everyone can work with. A lack of standardisation in file structures also hampered retrieval; however, interviewees displayed confusion about how best to resolve this.

Few researchers or research groups document their files as they work on them, and all acknowledged that retrospective documentation is difficult and time-consuming. As a result, researchers don't document retrospectively unless they are forced to do so; for example, to share data on request. PhD students pose a particular problem in many departments as the process of leaving and handing over relevant information is not always managed. As PhD work is often part of a larger body of research on which their PIs are working, many think it would be useful to have the data in a documented form.



One research group at Glasgow uses a code number for naming, which is the person's initials, the lab book number in roman numerals and then the experiment number. When they produce the spectrum this reference can be entered as the sample number to physically connect the file with the experiment and the lab book.

Questions were raised about how best to implement procedures: should naming be defined on a lab/research group basis or should the Head of Department implement a standardised system throughout; and what level of detail should be captured for documentation and how?

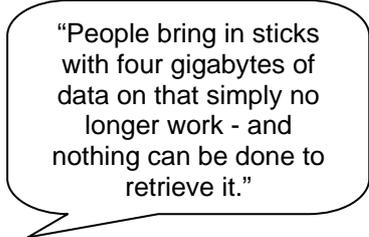
Both Cambridge and Glasgow found examples where formalised procedures for creating, structuring and documenting data have worked well. These tend to be implemented on the project or research group level. For example, there have been several initiatives underway in Chemistry at Cambridge for some time, including the rollout of a networked Electronic Lab Notebook.

Many interviewees acknowledged it would be very difficult to standardise approaches. However, they were largely enthusiastic about the prospect of clear guidance and encouragement to help researchers establish more robust methods.

## 2. Data storage and access

### Data storage

Various types of storage are used throughout each institution, ranging from networked storage (i.e. shared drives hosted on a server) to computer and laptop hard drives, external media such as memory sticks, CD/DVDs and hard disks, and online third-party providers. The degree to which researchers understand their storage options varies, but many have limited knowledge of the different storage options available to them. Researchers and technical support staff at both institutions raised concerns that many researchers do not understand the differences between types of storage, including which ones are the most resilient or secure, or how long a given storage medium is likely to last.



“People bring in sticks with four gigabytes of data on that simply no longer work - and nothing can be done to retrieve it.”

Networked storage is available in the majority of departments/research groups throughout Glasgow and Cambridge. However, researchers do not always use it. Typically, IT support encourages researchers to work and save files to the network to ensure data are automatically backed-up. Many ignore this guidance as:

- the network is felt to be too slow (primarily expressed at Glasgow);
- files are too large to work on over the network;
- there is a fear of not being able to access data when the network is down (primarily expressed at Glasgow);
- they have run out of network space and additional space is expensive.

Computing professionals at both institutions were concerned that researchers did not understand the risks of using off-the-shelf external hard drives as cheap primary storage or back-up space. There is a tendency among some research groups to write the capital costs of storage (e.g. computers or external hard drives) into grant proposals without funds for support time, for set-up, maintenance, or back-up. This was unlikely to be a problem in projects where researchers are able to anticipate that they will need a large amount of space for storage and processing (e.g. in departments where this sort of project is common, such as Public Health).

Many also commented that storage is cheap and felt that there wasn't any need to delete files. While many interviewees also mentioned that they often find it difficult (and time-consuming) to find files from important past projects, they did not make a connection between this problem and the flood of un-weeded files saved on live storage.

### Access to data

Many researchers noted they work remotely at times, but procedures for doing so vary. There are few explicit guidelines in place within departments for maintaining the security and version control of material worked on remotely. The Cambridge Institute of Public Health is an exception, having created a data storage and transfer policy (though some interviewees acknowledge that they (a) were not aware of it, or (b) do not tend to consult it).

Most people take data home on memory sticks and laptops or email material back and forth. A couple use NetStorage (Glasgow), and others keep a mirror of all files on their laptops. At Glasgow, interviewees were not always aware of existing facilities such as VPN or NetStorage. These resources for external access are not available in most Cambridge departments, though some have begun to discuss making them available if they can get the funding.

Transferring data back from the field has posed some challenges, especially when researchers are working in areas with poor internet connections. At Cambridge, interviewees in Archaeology cited this as a common issue, tending to rely on their laptops and external hard drives to bring their data home. One researcher at Glasgow has tried to find other ways to submit data, for example by text messaging, but this does not work for uploading large files.

Nearly all researchers at Glasgow and Cambridge work with people from other UK and international institutions, yet there are few clear departmental guidelines on the best methods for providing external parties with access to data.

In most cases, data are emailed or posted out on disc, or in some departments researchers are able to use an FTP or SFTP server for download. There are some groups who have addressed this problem successfully, such as the EEE at Glasgow who are currently involved in the Nano-CMOS eScience project which is using the Andrew File System to share data across partners at Manchester, Southampton, Edinburgh, York and Glasgow. This provides secure access to a single instance of a file from anywhere within the network.

### **3. Back-up**

In cases where networked storage is used, IT provides an automatic back-up service that typically copies data on a daily basis, with a secondary off-site copy and old back-up tapes kept for some months. In both Glasgow and Cambridge, however, a number of research groups do not use networked storage at all so individuals are responsible for their own back-up.

Researchers weren't always aware of best practice in these cases.

*"I just back everything up onto data sticks. I didn't even know you could back-up to servers".*

In Engineering at Glasgow, many researchers work on Macs and use TimeMachine to provide frequent incremental back-ups to external hard drives, often setup in a RAID array for reliability. Other approaches we encountered were not so robust. Some researchers copy data to memory sticks, CDs or hard drives whenever they remember (which was typically described as "not as often as I should") or periodically mirror files held on their PC to a laptop.

Many interviewees acknowledged that some researchers do not back-up regularly or at all. Some groups store hundreds of gigabytes of data so there is often no affordable way of doing back-up available to them. Insufficient back-up space is a recurring problem in some departments. However, this was seen less as a problem of lack of space and more an inability to control what people store on their hard drives. Most computing officers have witnessed 'moderate' or 'catastrophic' data losses as a result of this mix of practices and levels of resource and expertise.

*"PhD students lose material all the time. And they are exactly the people who want to be backing up. These are people who are creating data which is life and death important to them"*

The term 'back-up' was often used interchangeably with 'archiving'. Many interviewees thought that having backed-up data meant it would be secure for the long-term. To a large extent this is probably due to the limited range of options available to researchers. Indeed, continual back-up is the most common strategy adopted at both institutions to keep data accessible, as alternative provision is not made. When inactive data are kept on current storage intended for fast and reliable access to core records in active use, there is a continual risk of deletion, inadvertent change or corruption. Separate and reliable storage would be more secure and economic for preservation purposes.

#### **4. Preservation**

On the whole, preservation approaches are not very formalised. Indeed, researchers typically hadn't considered what would happen to their data in the long-term. The most common approach is to do nothing: data are kept on current storage systems and continually backed-up. Concerns were raised about the robustness of this approach, as the storage is intended to provide fast and reliable access to core data in active use, not for preservation purposes. Active and historic data are not differentiated and seldom protected, so there is a continual risk of inadvertent changes or deletion.

Other strategies in use at the two universities were to copy data onto CDs or external hard drives so a secure reference copy was held, to deposit with service providers such as Institutional

Repositories and data centres, and to rewrite code/migrate data so it remained accessible. Regenerating data was noted as a potential option in the future.

### Selection

A theme closely associated with preservation is data selection. Lots of data are being created, yet few researchers actively select data for preservation - most hope to keep everything indefinitely. Some felt that there is no need to destroy anything, as storage is now so cheap, while others voiced concerns over how one decides what to keep when future usage cannot be predicted.

At Cambridge, one researcher had kept every email he has sent or received, as he believes the information in them is important.

There was a gulf in opinion between what IT support felt it was practical to retain and the expectations of researchers. One IT manager estimated that only one tenth of what is currently held in his department should be retained and that this could be achieved if people were educated in certain ways. Both Chemistry and the Clinical School of Computing at Cambridge have introduced subscription-based computing services and have found that researchers who use these services tend to think more carefully about what data they wish to store at present and in the long-term.

“When you take a book out of the library and there are pages missing, you bring it back to the library and expect them to fix it”.

Financing preservation is a moot point. Many funding bodies expect data to be maintained in the longer-term. However, researchers seem reluctant to cost data management and preservation into grant proposals for fear that their bid will not look competitive and that money will be taken away from core research. As such, data are

typically sustained through the goodwill of the researchers involved. Several researchers expressed frustration that there is seldom leadership on data management within departments or someone clearly responsible for preservation.

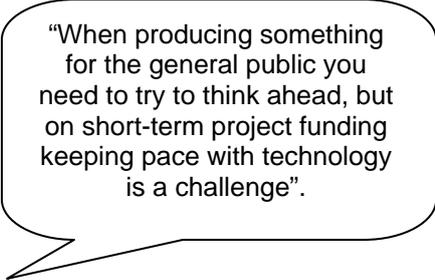
### Feasibility of preservation

Some interviewees discussed what was appropriate in terms of preservation. There was a sense at both institutions, particularly in the sciences, that the final publication is the main output that needs to be preserved. These fields move forward so quickly that there is an assumption any data more than a few years old is obsolete and what remains in research papers is sufficient. In Archaeology and English Language there was a greater desire to keep as much as possible for longer periods. Archaeological data records things that no longer exist so it's very difficult for people to justify throwing things away, while digital resources in English would often be expensive to recreate at a later date.

Discussion in English Language at Glasgow focused on what preservation means – is it a case of keeping the bits or retaining the usability? Raw data was often felt to be useless without its context and structure, as digital resources take on great value as an integrated whole. In the SCOTS project,<sup>10</sup> for example, the value comes in listening to or searching audio recordings in conjunction with the transcript with pertinent sections highlighted. The connected data inputs and tools provided for researchers to mine and analyse the corpus as a whole is what they felt needs to be preserved.

Preserving the usability of a resource is a far more challenging proposition, as it relies on maintaining interactions between operating systems, hardware and software, as well as user knowledge of these environments. The rapid rate of technological advancement has led to various issues in this regard.

Software backwards-compatibility issues have been encountered in most cases. A number of examples were provided of ‘garbled text’ when opening files from several years ago as the current software no longer supported the version used to create the data. Similarly, expensive annual software licences cannot always be afforded, so there is a risk data will be locked in a format that can no longer be read.



“When producing something for the general public you need to try to think ahead, but on short-term project funding keeping pace with technology is a challenge”.

Outdated hardware also poses challenges, particularly in the sciences. Most pieces of scientific equipment are attached to their own computer which has a much shorter lifespan. A new PC won’t always be compatible with the old hardware needed to work with the equipment. Moreover, it could require an updated version of the specialist software, which is unlikely to be compatible with earlier data files. The present solution is to sustain outdated and fragile setups for as long as possible. There was a sense in many disciplines that obsolescence is inevitable and that at some point preservation may become impractical so you have to just let go.

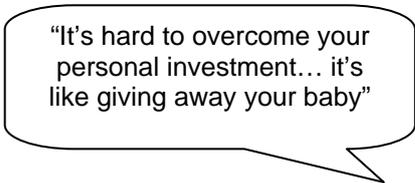
## 5. Data sharing and re-use

Most researchers acknowledged the benefits of data sharing in principle, but in practice they referenced a number of barriers to doing this. Practical aspects such as a lack of documentation were noted, as were tensions between making data open and maintaining a competitive edge, particularly when working with commercially sensitive data, as in Cambridge’s engineering design centre. Often researchers may only access data from the industrial partner’s site or will only be supplied with anonymised derived data. Researchers in English have encountered issues surrounding copyright when digitising materials from other institutions and these are further compounded when working with a number of different collaborators, all of whom have their own set of views and policies on data sharing and open access. Interviewees in Archaeology cited concerns over releasing data that might indicate the precise location of the excavation site,

---

<sup>10</sup> See: <http://www.scottishcorpus.ac.uk/>

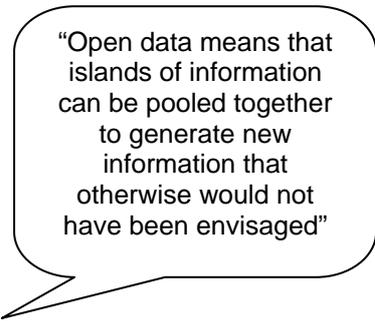
leaving it open to desecration or theft. There was an overwhelming sense in interviews that reports or publications that synthesise results are more appropriate and feasible to share than providing direct access to data.



“It’s hard to overcome your personal investment... it’s like giving away your baby”

There were a few areas where data sharing is more common: for example, on AHRC-funded projects and in biology research where DNA sequence data is deposited in NCBI GenBank. A few data sharing initiatives were also noted: the Nano-CMOS project in EEE at Glasgow is building an archive database of research results that will be accessible over the Grid; a dummy data portal is being devised by the MRC Unit in Glasgow to provide access to sensitive data; and researchers at SPRI can submit their climate/ocean model code to community models such as the Community Surface Dynamics Modelling System. Outside of these areas, data sharing is typically done on a researcher-to-researcher basis allowing control to be retained and all the necessary explanations to be made to make sure data aren’t misinterpreted.

Researchers in English Language at both institutions were particularly supportive of data sharing, noting wider benefits that ensue and hoping the tendency to ring-fence data for personal use could be overcome. Interviewees in Chemistry and SPRI at Cambridge also noted the importance of knowing what data were available in order to facilitate sharing to prevent replication and allow others to build on previous research. Members of staff in Chemistry are part of an initiative that is aiming to create an enhanced digital repository for chemical data that captures core types of chemistry data and ensures their access and preservation. The digital repository will enable scientists to make selected data available as Open Data for use by people external to the department.



“Open data means that islands of information can be pooled together to generate new information that otherwise would not have been envisaged”

Data sharing was more readily discussed by early career researchers. Some use Web 2.0 applications to facilitate collaborative working such as GoogleDocs, EditGrid and Mendeley, and institutional VLEs such as Camtools. Usage often comes down to familiarity and confidence – some noted that if they were trained they’d use collaborative environments such as wikis more.

## IV. Existing Guidance and Interviewee Requests

In speaking with researchers, we found that many were interested in guidance, simple tools, and support for data management, but that this came with several caveats.

### 1. Existing resources are difficult to find

Interviewees were often unaware of existing guidance, resources, training opportunities, and policy documents, which are scattered through internal and external websites. Many interviewees showed interest in web pages, which would point researchers to all relevant data management tools and resources.

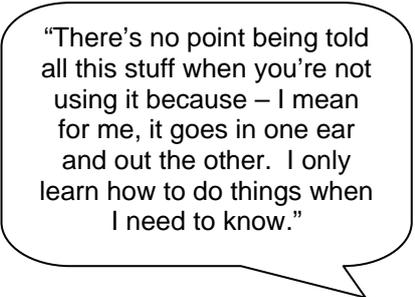
### 2. Existing resources are difficult to use

Where researchers were aware of existing resources, they usually acknowledged that they were never inclined to consult them. They found the documents and tools available (both within institutions and on other discipline-related websites) to be dense, wordy, theoretical, ambiguous, and unengaging. They want simple, practical answers for their data management problems and questions, and do not know where to find them within the available documents, or they do not generally have time to tease out the answers.



“The policy was huge and not very clear. It took a few attempts to understand, whereas you just want a quick yes/no answer.”

### 3. Training has to be more convenient and relevant



“There’s no point being told all this stuff when you’re not using it because – I mean for me, it goes in one ear and out the other. I only learn how to do things when I need to know.”

Many researchers complained that training is often inconveniently timed and not well-tailored to their needs. Researchers were unlikely to attend a training session voluntarily (e.g. there’s never any time), or to attend training sessions that weren’t tailored towards them. Many showed an interest in seeing practical (non-theoretical) data management training included in existing training provision (for example, induction procedures and mandatory research skills courses for PhD students).

### 4. Researchers don’t know who to ask for assistance

In some cases, researchers have wanted to talk through data management challenges or questions with an individual, but did not know who to talk to. For certain technical issues (e.g. advice on storage media) some researchers have the option of talking to IT personnel within their departments; however, IT and data management provision and expertise vary by department. Very few researchers were aware of existing one-to-one data management advice services within their institutions (DSpace@Cambridge and the Data Protection and Freedom of Information Office at Glasgow).

## V. General Findings on Data Management Support

1. Address simple problems and acknowledge some enhancement is better than none

The data management problems that researchers said were the most irksome and time consuming were simple, day-to-day issues, like not having a file management system or naming system, not knowing the best file format to use in the medium-term, or *ad hoc* data storage solutions. We recognise that academics are busy pursuing their research, so providing simple and practical tools and training to address a variety of the most troublesome problems can go a long way to solving some of these issues

2. Use clear language and avoid jargon

Clear, jargon-free language is essential both in assessing researcher concerns/risks and in providing services. Researchers are confused by phrases like ‘digital curation’ and most don’t know what a ‘digital repository’ is. Many people are suspicious of ‘policies,’ which sound like a hollow mandate, but are receptive to ‘procedures’ or ‘advice’ which may be essentially the same thing, but convey a sense of purpose and assistance rather than requirement. Many of the resources available assume pre-existing knowledge and vocabulary that most researchers do not have and to which most cannot relate; it is important to avoid this pitfall in any resources we create.

3. Points of intervention

Most interviewees believed that the best point to intervene with guidance, training or management tools is very early in researchers’ careers. The suggestion of data management and preservation training for PhD students and post-doctoral researchers has been particularly well-received, as this is one point where habits begin forming and young researchers do or do not learn standard practices from their more senior colleagues.

Many interviewees suggested that more senior researchers’ practices, habits, and beliefs are fully established and relatively little can be done to adjust them (one computing officer quipped that all he could do was provide services and wait for senior researchers to retire). While it may be difficult to strongly influence the practices of most senior researchers, some senior researchers will support efforts to improve practices and induction for the next generation of researchers. In addition, senior researchers often guide project procedures and hold data long-term, so they are ideal (if not easy) targets.

## VI. Implementation Plan

### 1. Recommendations

In response to the needs identified in the scoping study, we make the following four recommendations:

1. Produce simple, visual guidance on creating, storing and managing data

Many researchers asked for simple, practical advice in formats that make it easy for staff to access relevant information quickly. Various formats were suggested such as factsheets, FAQs, checklists, crib sheets, flow diagrams, bulletins, newsfeeds and email alerts.

Several concerns were raised about the ability to find relevant guidance when needed, as there's not time to trawl the website or read through large policy documents. A central place to point researchers to relevant resources was called for.

Rather than reinventing the wheel, we plan to repurpose existing guidance where possible to make it more visual and user-friendly to researchers. We will point to both institutional resources and those developed externally if appropriate. A particular case in point may be the ongoing Mellon Foundation-funded revisions of the Archaeological Data Service (ADS) Guides to good Practice<sup>11</sup>. New guidance documents such as checklists and factsheets will also be produced. Potential topics include choosing appropriate data formats, file management, negotiating consent agreements to permit preservation and re-use, and using Web 2.0 tools e.g. GoogleDocs, Dropbox and Mendeley. We also anticipate developing research data management webpages at each institution.

2. Offer practical data training with discipline specific exemplars

Researchers were keen for practical best practice guidelines and flexible modes of training such as online tutorials, video case-studies and interactive learning resources. More links to local support are also needed and this is often overlooked.

Additions will be made to existing training courses to raise awareness of services and provide more practical tips from a researcher perspective – for example, by documenting their lessons in case-studies. Basic slides and resources will be produced that could be dropped into other courses to advance our 'train the trainer' approach. We aim to embed data training early in the research lifecycle, targeting PhD students and early career researchers, as our scoping work has indicated they are often given responsibility for data management.

---

<sup>11</sup> See for example: <http://ads.ahds.ac.uk/newsletter/issue23/guides.html>

3. Connect researchers with support staff for tailored advice, guidance and partnering  
Several requests were made for tailored support. A university body or service that could be consulted to help staff adopt best practice was suggested. Many interviewees felt it would be useful to have someone come to the department and talk through specific requirements and options to help determine what is most appropriate in the given context. This was particularly called for at the grant application stage in lieu of data management plan ('DMP') requirements, at project initiation when procedures were being defined, and at wrap-up when decisions were made about the long-term.

Support of this kind is currently offered by some Research Development Officers or local technicians and through information audits; however researchers are not always aware of the available provision. A concern was raised about gaps in this support network. It was felt a listing could be provided so all members of staff have a named contact for support. A more formal network would also provide support staff with a forum in which to raise concerns and share knowledge.

Through *Incremental* we will raise awareness of existing support and can build links with research offices to point researchers in right direction. An addition to the Project Approval Form (PAF) has been suggested at Glasgow to flag data management requirements. Tailored support and partnering will be offered in key areas of need e.g. when writing DMPs.

4. Work towards the development of a comprehensive data management infrastructure  
At present there are few mechanisms available to researchers for data preservation, so they tend to keep material on the live storage systems so it can continue to be accessed. Many interviewees viewed back-up as akin to archiving. However, IT staff drew a clear delineation, commenting that keeping files on the live storage system is not sensible in the long-term as there's a continual risk things could be deleted or changed. Several called for more storage options, noting that a large pool of slow yet reliable storage to enable digital archiving would be very useful.

A data management infrastructure wouldn't, however, be based on storage alone - it would also comprise policies, best practice guidance and support staffing. The majority of people felt some form of policy or guidance was needed as current work practices are seldom coherent across groups, which can pose longevity and re-use issues. It was felt this would bring clarity so people knew what they were supposed to be doing. Local policies and procedures were seen as more achievable but it was commented that a high-level policy or statement of commitment would provide a useful overarching framework for these.

The *Incremental* project can work towards these aims by engaging researchers in discussions about data management to explore where the greatest needs are and in which circumstances preservation is relevant. Networks to share knowledge and expertise can be explored and support for the development of local policies and procedures will be offered.

## 2. Proposed tasks and activities

A draft timeline for implementation activities is available in Appendix 1

### Tasks for recommendation 1: Produce simple visual guidance

#### Create a collection of webpages to help researchers find tools and assistance

- Agree on design for webpages. Likely navigation will include question-led multiple points of entry. For example, users will be able to click on “Who are you?” and pick PhD student, PI, etc, or click on “What stage is your project in?” or “What types of data are you creating?”
- Design, trial and produce webpages. Design may vary slightly between Cambridge and Glasgow webpages, though the navigation and non-local content will be the same for both institutions. We will consult researchers (via email or workshop discussion) to ensure that the design of the webpages are user-friendly. We may make use of limited consulting assistance from a company or organisation such as CARET (Cambridge).
- Determine locations of webpages. Cambridge’s is likely to be situated within the DSpace@Cambridge web domain. Glasgow’s is likely to be situated within the Research & Enterprise web pages or other appropriate centralised location.
- Announce the webpages, workshops, and resources. This is likely to happen in a variety of ways. Cambridge and Glasgow will arrange to announce their webpages, workshops, and local support at scheduled training events. Both will also be likely to use email announcements (for Glasgow this may mean the research staff mailing list). In addition, both institutions will attempt to work with their local research office to point researchers to these resources at the point of funding application.
- Identify and select links to existing sources of information/guidance/tools. We will categorise these and will likely annotate each with the sentence to explain uses/limitations of each resource to users.

#### Create new, easy-to-use resources

- Determine what new resources would be most helpful within the scope of the project and which resources should take which forms. For example, we will need to determine what sub-topics to include for file management, file formats for

preservation, IPR, etc. Formats for these resources will include fact sheets, FAQs, checklists, and flow diagrams (possibly hypertext).

- Create the new resources, consulting with researchers at Cambridge and Glasgow informally and, where possible, through workshops.

### **Tasks for recommendation 2: Practical training with exemplars: 'Train the Trainer'**

- Carry out an appraisal of current training and guidance services at Cambridge and Glasgow. This will include desktop/web research (which has already begun) and face-to-face meetings with key actors in departmental and university-wide training and administration (e.g. speaking to post-doctoral and post-graduate committees).
- Determine which existing courses we will work to incorporate our resources into, and start making connections/arrangements with course administrators and professors. In Cambridge, this may include the Graduate and PI Development Programmes<sup>12</sup>. In Glasgow, this may include SDS courses on managing research data, and research staff conferences.
- Link training materials back to institutional policies and/or external funding body requirements wherever possible to make the case for data management and curation clearer to researchers. For example, at Glasgow we'll be referring to the GU Code of Good Practice for Research.<sup>13</sup>
- Create PowerPoint slides, video-casts/case-studies, and screen-casts. Some of these (especially slides) will incorporate discipline-specific examples and we will consult with researchers from the requirements-gathering stage of the project. In all cases, we will attempt to make these resources stand-alone and visually engaging.

### **Tasks for recommendation 3: Tailored advice and partnering**

- Raise awareness of existing support, including one-to-one support from the Research and Development Office or Data Protection and Freedom of Information Office (Glasgow) and DSpace@Cambridge (Cambridge). This is likely to take the form of informational posters, email list announcements, and working with departmental and administrative bodies to (a) ensure that they are aware of these services, and (b) get them to link to them on their websites and other resources. These materials will emphasise the concept of 'clinic' or 'surgery' services for electronic data.

### **Tasks for recommendation 4: Comprehensive data management infrastructure**

- Investigate proposed strategic changes within schools and departments to ensure that proposed services and tools will be sustainable, e.g. consult research strategy committees.

---

<sup>12</sup> For example, postgraduate transferable skills training:

<http://www.skills.cam.ac.uk/postgrads/training/>

<sup>13</sup> Code of Good Practice in Research [http://www.gla.ac.uk/media/media\\_46633\\_en.pdf](http://www.gla.ac.uk/media/media_46633_en.pdf)

- Engage researchers in discussions on topics such as data management and relevance of preservation to different researcher fields. This will include workshops for the project, as well as ongoing contact with researchers throughout the project and through resources/contacts created during the project.

### **Costs and benefits of enhanced data management activities**

- We will work through a costing survey (draft available in Appendix 4) with partners for the piloting and evaluation stages of the project. While the current scope of the project does not allow us to gain a comprehensive set of costs and benefits, this will help us get a general sense of (1) time lost or gained and (2) data lost or preserved. Neil Beagrie, of Charles Beagrie Limited<sup>14</sup>, has provided us with some very helpful feedback on the construction of the costing questionnaire and continues to offer support via JISC.
- We will incorporate cost-benefit information-gathering with topics from our costing questionnaire and from Keeping Research Data Safe (KRDS)<sup>15</sup> where possible in pre-implementation workshops and discussions. This will help improve our sense of perceived costs/benefits and increase our contextual knowledge about these issues.
- Finally, we will also make use of information that Glasgow has learned while serving as a pathfinder for the UKRDS project. This project is identifying additional high-level support costs which may be applicable for *Incremental*.

### **Evaluation activities**

- We will continue to evaluate the activities which we have proposed in this report through interviews and workshops. At all points, we will seek to identify intervention strategies that have worked well or which require further development or change of course.
- We will routinely seek the comprehensive feedback of researchers who pilot our tools through semi-structured conversations and small debriefing workshops at each institution. We will feed the information gleaned from these sessions back into the project tools and the bodies in each institution which will maintain them once the project is completed.
- Where possible, we will produce case study assessments and make use of tools like Assessing Institutional Digital Assets (AIDA).<sup>16</sup>

### **Wider dissemination of findings and outputs**

We will disseminate project outputs and findings locally within each institution in a number of ways, including:

- Through public announcements and e-mails during the implementation and evaluation phases of the project;

---

<sup>14</sup> <http://www.beagrie.com/>

<sup>15</sup> [www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf](http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf)

<sup>16</sup> <http://aida.jiscinvolve.org/wp/toolkit/>

- Through building relationships with offices which provide services to researchers (such as the Research & Enterprise Office at Glasgow and the Research Office at Cambridge). Our goal is to convince these offices to encourage researchers to speak with relevant data management support services within each institution during the funding application stage;
- Through researchers and support staff with whom we speak and work to pilot and evaluate project outputs. We have already begun cultivating relationships with these individuals and will continue to do so actively throughout the remainder of the development, implementation, and evaluation stages of the project. Associated organizations and projects at each institution will maintain and cultivate these relationships over the longer term.

We will disseminate project outputs and findings widely in a number of ways, including:

- Through publishing our findings and outputs through the JISC website;
- Through building relationships with other institutions through conferences and collaboration. We have already begun this process via JISC funding strand meetings and related conferences (for example, the Data Management Forum in Manchester on 10-11 March, 2010). In addition, we have several speaking engagements and participation in upcoming conferences planned (for example, *Incremental* will present findings for a meeting held by Oxford's Sudamih project and for a meeting of the British Library in July 2010);
- Through sharing our resources with existing data management and curation organizations within the community, such as the Digital Curation Centre (DCC), the Digital Preservation Centre (DPC), and JISC.

## VII. Appendices

### 1. Timeline for Implementation plan

Recommendation	Activity	Task	2010						2011			
			J	J	A	S	O	N	D	J	F	M
1. Produce simple visual guidance	Create web portal to help researchers find tools and assistance	Agree on design and location for data portal	x	x								
		Design and produce data portal		x	x	x						
		Publicising the portal via workshops/posters				x	x					
		Identify links to existing sources of information guidance or tools	x	x								
	Create new easy to use resources	Conduct appraisal of where resource gaps lie	x	x								
		Create new resources, e.g. check lists, flow diagrams, FAQ's consulting researchers where possible via workshops, email		x	x	x						
2. Create practical training resources with discipline specific examples	Create resources	Conduct appraisal of current training and guidance services at Cambridge and Glasgow	x	x								
		Create stand alone slides/video casts/casestudies and screen cast resources that can be dropped into other courses		x	x	x						
		Create online training materials and tutorials		x	x	x						
3. Offer tailored advice and partnering	Raise awareness of existing support staff and services	Create materials, e.g. informational posters, webpages for inclusion in data portal and training events. Forge close links with the research office so they are better informed about existing and new guidance and support in data curation			x	x	x	x				
	Offer tailored advice	Offer on-to-one support to help researchers define best approach for their context					x	x	x	x	x	
4. Work towards development of comprehensive data management infrastructure	Investigate proposed strategic changes within schools and departments	Desk based research, face to face consultation with HOD's, research strategy committees			x	x	x	x	x	x	x	
	Engage researchers in discussions about importance of data management and preservation	Organise workshops to bring academics and information/data management professionals together						x	x	x	x	

## 2. Interview templates – Cambridge

### Interview Template for Researchers

#### Scope of digital holdings

We will begin by briefly discussing major projects of the interviewee(s) with regard to electronic data i.e. what they create/use, current practices and preferences for digital material, etc.

- What electronic data 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 data do you currently create / hold? Is this growing?
- Are these files yours or do they belong to a wider group or to the institution, etc?
- Are these files replaceable/reproducible?

#### Working practices

We will discuss what happens in terms of digital data management i.e. creating, maintaining and preserving electronic data.

#### *Guidance and Responsibility*

- Are there departmental guidelines, policies or procedures you follow? (for backup, storage, sharing, documenting, etc)
- Do you know when, how and what is backed up centrally?
- Who is responsible for digital material? What role does each person play?
- Do you work differently on research projects due to funding body requirements?
- What happens in terms of legacy material i.e. files created by former staff?

#### *Individual and Group Practices*

- How do you create electronic research data? – naming conventions, filing rules...
- Where do you store files? Do you back them up or is this done centrally?
- Who can access electronic material? How is this controlled? Explain restrictions
- How do you manage digital files e.g. do you sort through and weed them?
- Does any contextual information (or data) reside solely in e-mails? How do you manage these?
- Do you ever have trouble finding or interpreting your own data from past projects?
- Do your working practices differ when working on your own and working in a group?
- Is it difficult to understand other people's systems on the shared drive?
  - Have you ever lost more than 15 minutes sorting out these version control issues/different systems? More than half a day? How often?
- Have these practices changed in recent years? If so, how/why?

#### Digital preservation issues

We will continue discussion to understand whether any issues or novel solutions have been encountered when creating and using electronic material over time and to identify room for improvement.

- Have you ever lost digital files or found it hard to find the right ones?
  - o Have you ever spent more than 15 minutes looking for or recreating information? More than half a day? How often? Who usually does the hunting/re-creation (i.e. early-career researchers, project directors, or everyone?)
- Are there version control issues when working with colleagues?
- 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 data?
  - o If you sometimes have to obtain additional storage space, who funds this?
- Are you aware of data centre services? Do you ever use data provided by other researchers to these centres?
- Do you include data management in grant applications? (What sorts of costs?)

#### Future life of electronic research data

We will move discussion on to think about future needs and preparations for electronic data i.e. can files continue to be accessed/used, do they need to be preserved, if so, for how long...

#### *Access and Data Sharing*

- Could your electronic data be reused or repurposed by others?
- Are there any sensitivity or confidentiality restrictions?
- Would other people understand your data - is it documented? (Do you share your data with colleagues inside/outside of your institution?)

#### *Preservation*

- Does all digital data (raw, derived and published) 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 in the medium/long-term?
- Have there been any changes in practice in recent years e.g. because of changes in available technology?
- How do you imagine your data management practices will develop in the future?

#### Service requirements

- Where do you currently get advice and support?
- Are you aware of current advisory services within the University, e.g. The Research Office? Are you familiar with DSpace@Cambridge (University Digital Repository)?
- What services would you like to see made available?
- What would help you create and manage your electronic files better?
- Who should be responsible for / fund digital curation and preservation?
- Would you welcome a departmental (or perhaps even University wide) policy on digital data curation and preservation? If so, what should it cover?
- What recommendations do you have for the future development of data management practices within the University? Are there services/tools that you would like to see emphasised during the implementation phase of our project?

## Interview Template for Computing Officers and IT Personnel

It is expected that interviews will take between 30 mins - 1 hour. These will be recorded, then transcribed, with the text sent back to the interviewees 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 'Incremental' project and explanation of terms will be provided. Interviews will be semi-structured to allow free-flowing discussion. The questions provided below are indicative of the topics that may be discussed, but *not all questions will be addressed in each interview*. Each interview will cover six themes:

1. what kinds of support do you offer to researchers with regard to their data?
2. to what degree are you involved in researcher working practices?
3. any data management/support issues that have been encountered
4. resources used by/efficiency from current practices
5. the future for the department's electronic records
6. requirements for support and services from the university

### Scope of Support

- What kinds of support do you offer to users?
- Are there centralised departmental IT policies? (On data management, storage, backup, security, IPR?)
- If users are given server space, how much?
  - How is this managed?
  - What do you do if users ask for more space? (How is this funded?)
- Do you have a sense of what sorts of data are on the server?
  - Does it all belong there?
- What do you see as central issues/concerns for electronic research data management in your department?

### IT and Researcher working practices

- Who controls user permissions to different data folders?
- Do you know how researchers in your department share data with collaborators outside of the university? (Is this different for sensitive vs. regular data?)
- Do researchers make you aware of their data security requirements (and other contractual data requirements) for specific projects? (Should they?)
- Do users ever involve you in their grant/contract applications to determine IT needs and costs? (In what circumstances?)
- How (if at all) have your practices with regard to researcher data changed in recent years?

### Digital Preservation

- Do users sometimes come to you because they can't access old files/formats? (Is this something that you can usually fix? How long does this take?)
- What happens to the materials on a user's account (server/e-mail) when s/he leaves the department or the university?
- Do users ever ask for help in finding files created by previous users (or their own files)?

### Future life of electronic research data

- How do you (or your users) preserve data long-term?
- Do you have a sense of how much of their data your users preserve indefinitely, and whether they tend to want access to it later?
- Are you aware of any long-term data storage and sharing centres for your users' discipline? (Do you think you should be aware of them/would you like to be?)
- How do you expect these practices for data preservation to change in coming years?

### Service requirements

- Do you look to any university offices or policies for guidelines in supporting users?
- To what degree do you think the university's centralised organisations (such as the University Library, DSpace, the Research Office or Computing Services) should be involved in researcher data management?
- We are considering providing services such as policy templates and training to help users manage data over the short and long term. Are there any services along these lines that you think might be helpful?
- Do you have any opinions on what it would take to convince users to follow new procedures, follow existing policies more faithfully, or better document their work?

### 3. Interview template – Glasgow

#### **Digital preservation scoping study 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:

7. what digital material is being created;
8. how this is being created and maintained;
9. any issues that have been encountered;
10. the future for the unit's electronic records;
11. 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

### 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 re-used 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?

#### 4. Draft costing survey

##### Survey on Costs and Benefits of Managing Research Data

Thank you very much for agreeing to participate in this survey, as part of the Incremental project. The purpose of this survey is to gain a sense of the costs and resources involved in managing digital research data.

The information provided by you in this questionnaire will be used for research purposes only. It will not be used in a manner which would allow identification of your individual responses.

##### Data management plans

**(1) In applications for funding, are you usually required to include a data management plan (i.e. a description of how you plan to create, maintain, secure, and preserve your data)?**

YES       NO       I don't usually deal with funding applications (skip to (3))

**A. If NO, do you usually...** (tick all that apply)

- Follow written departmental/group guidelines for managing data?
- Create new internal rules for each project?
- Establish procedures on an ad hoc basis as needed throughout the project?
- Other: \_\_\_\_\_  
\_\_\_\_\_

**B. If YES, what does this usually entail?**

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

**(2) Do you usually include data management, computer hardware/software, or data storage costs in the *budget* when applying for a contract or grant?**

NO (I do not usually budget for data management costs separately)

YES, including... (tick all that apply)

- Technical support/staff costs for setting up systems to manage data
- Costs for server space to store data during the project
- Costs for shared space to store and share electronic data
- Costs for sever space to store data long term

(For how many years? \_\_\_\_\_)

- Back-up facilities for project data
- Time needed for creating metadata and documentation
- Costs for preparing data to the standard required for deposit in a data archive

Other: \_\_\_\_\_  
\_\_\_\_\_

**(3) On balance, do you think that the benefits of having structured data management policies/practices justify the costs associated with creating and following them?**

- YES  
 NO

**A. Please tick all that apply:**

- Data policies make it easier to find/understand work from several years earlier
- Data policies make it easier to share data with colleagues
- Data policies help to meet funder/publisher requirements (e.g. depositing in a data repository or producing data on request)
- Structured data management practices are more likely to result in good quality data and robust research outputs
- Structured data management practices help to maximise the investment in generating research data
- Data management issues aren't causing substantial delays, problems, or resource losses with our current practices
- We don't have the time up-front to create/monitor data policies
- We have attempted to create formal procedures in the past, and it took up more time/resources than it returned in benefits

**B. Please share any additional reasons for your answer at the start of (3):**

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

## File Management

**(4) Do you ever have difficulty finding the right files or data on your own or a shared drive?**

- YES, with shared drives
- YES, with my own drives
- NO (skip to "4.F")

**If YES to either question...**

**A. What is the typical cause(s) of difficulty in finding your own files? (tick all that apply)**

- Not applicable
- Inconsistent file naming conventions
- Inconsistent file storage locations
- Dealing with too many saved files
- It is not always clear which version is newest
- Other: \_\_\_\_\_  
\_\_\_\_\_

**B. What is the typical cause(s) of difficulty in finding files on shared drives?** (tick all that apply)

- Not applicable
  - Inconsistent file naming conventions
  - Inconsistent file storage locations
  - Lack of familiarity with collaborators' parts of the work
  - It is not always clear which version is newest
  - Changes in personnel (e.g. the person who created those files left the office)
  - Other: \_\_\_\_\_
- 

**C. How long does it usually take to retrieve the desired files / data?**

- Less than 15 minutes
- A couple of hours
- More than half a day
- Other: \_\_\_\_\_

**D. Approximately how often does this happen?**

- A couple of times a week
- A couple of times a month
- A few times a year
- Rarely
- Other: \_\_\_\_\_

**E. Have you ever had to re-create data (e.g. re-do an experiment or repeat an analysis of raw data) because of lost files?**

- YES                       NO

**IF YES**, how long did it take to re-create the data? (If you had to spend money to re-create/re-collect the data – how much did this cost?): \_\_\_\_\_

---

---

**If NO to both questions...**

**F. What do you have in place to help avoid such problems?** (tick all that apply)

- Good search tools
  - Strict policies for naming
  - Strict policies for file locations
  - Regular checking, reordering, tidying up of shared drives
  - Regular checking, reordering, tidying up of personal drives
  - Strict policies for exiting staff to provide documentation for shared files
  - I tend not to work on shared drives
  - Other: \_\_\_\_\_
-

**(5) Do you use any shared systems such as GoogleDocs or shared server space to avoid version control problems when working collaboratively?**

**NO**

**YES** (Please specify): \_\_\_\_\_

\_\_\_\_\_

**(6) Have you ever been unable to open or read files because of obsolete formats?**

**YES**

**NO** (skip to Question 5)

**If YES...**

**A. Approximately how often does this happen?**

A couple of times a month

A few times a year

Rarely

Never

**B. How serious has this been?**

(Tick all that apply, and let us know how often this has happened if possible)

Some garbled text/characters, but the data/files were still usable

Major parts of the data/file were lost

The data/files were completely lost, but not deemed important

Significant data/files were completely lost

**C. Please provide a bit more context of the situation(s) if possible/applicable:**

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

**D. Have there been occasions where you have had to spend time recovering old data e.g. by converting the format or tracking down old readers?**

**YES**

**NO**

**IF YES**, please explain a bit about the process e.g. how long it took, the costs involved, whether it could have been avoided etc.: \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

**E. Have you ever had to re-collect or re-create data (e.g. re-do an experiment, re-analyse raw data) because of uninterpretable files?**

YES                       NO

**IF YES**, how long did it take to re-create the data? (If you had to spend money to re-create/re-collect the data – how much did this cost?): \_\_\_\_\_

---

---

---

**Metadata and user documentation:**

**(7) Do you usually create *metadata* (e.g. description of your datasets, what variables mean, what files refer to what piece of data, etc) or *user documentation* (e.g. documentation to assist external users using your data) for your digital research data?**

YES                       NO

**If NO...**

**A. How long do you think it would take to create metadata and documentation so that other users could understand your data?**

---

---

---

**B. Would it be realistic/feasible to do this?**

---

---

---

**If YES...**

**C. what does this metadata/documentation consist of?** (e.g. metadata with variable names, file paths for computer programs used to transform the data, etc).

---

---

---

**D. Do, you usually create metadata near the end of a project or as you collect/transform the data?**

BEGINNING/DURING                       END

**E. Do you have a sense of how long you spend on creating metadata?**

---

---

**(8) We are looking for examples of data management strategies that have worked particularly well (or badly) to distil and share lessons and best practice with other researchers.**

**Do you have any stories you're willing to share about the systems and procedures you have developed, that others might find worth investing time and effort in too?**

**THANK YOU FOR PARTICIPATING IN THE INCREMENTAL PROJECT DATA MANAGEMENT RESOURCES SURVEY!**