Research Data Management at LSHTM: Web Survey Report

**Title:** Research Data Management at LSHTM: Web Survey Report  
**Project Name:** Research Data Management Support Service (RDMSS)  
**Draft/Version:** Version 1.0
Catalogue Information

<table>
<thead>
<tr>
<th>Title:</th>
<th>Research Data Management at LSHTM: Web Survey Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator:</td>
<td>Gareth Knight</td>
</tr>
<tr>
<td>Keywords:</td>
<td>data management, web survey, research data</td>
</tr>
<tr>
<td>Description:</td>
<td>Results of a web survey performed at the London School of Hygiene and Tropical Medicine on data management practices</td>
</tr>
<tr>
<td>Publisher:</td>
<td>London School of Hygiene and Tropical Medicine</td>
</tr>
<tr>
<td>Date:</td>
<td>8th April 2013</td>
</tr>
<tr>
<td>Language:</td>
<td>English</td>
</tr>
<tr>
<td>Rights:</td>
<td>Creative Commons Attribution-NoDerivs 2.0 UK: England &amp; Wales License</td>
</tr>
</tbody>
</table>

Report Version Control

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Author</th>
<th>Change Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Draft 1</td>
<td>17/12/2012</td>
<td>Gareth Knight</td>
<td>Production of first version for presentation at the LSHTM RDM Steering Group meeting in December 2012</td>
</tr>
<tr>
<td>Draft 2</td>
<td>21/02/2013</td>
<td>Gareth Knight</td>
<td>Addition of Executive Summary and figure indicating the number of funders that support each researcher</td>
</tr>
<tr>
<td>Draft 3</td>
<td>15/03/2013</td>
<td>Gareth Knight</td>
<td>Revision based upon feedback by Anne Mills. Addition of discussion section and section on limitations of survey findings</td>
</tr>
<tr>
<td>Version 1.0</td>
<td>08/04/2013</td>
<td>Gareth Knight</td>
<td>First version finalised</td>
</tr>
</tbody>
</table>
EXECUTIVE SUMMARY ................................................................................................................. 4

1. INTRODUCTION ...................................................................................................................... 5
   1.1 Overview ............................................................................................................................ 5
   1.2 Design and implementation of the survey .......................................................................... 5
   1.3 Limitations of Survey results .............................................................................................. 5

2. WEB SURVEY RESULTS ....................................................................................................... 6
   2.1 Overview ............................................................................................................................ 6
   2.2 Q1. Research Faculty ........................................................................................................ 6
   2.3 Q2. Research Project Funder ............................................................................................ 6
   2.4 Q3. Description of Research Data ..................................................................................... 8
   2.4.1 Does it contain primary data created during your research? ...................................... 8
   2.4.2 Does it contain quantitative data? ............................................................................... 9
   2.4.3 Does it contain personally identifiable information at any stage? ............................... 9
   2.5 Q4. Commonly used software products ........................................................................... 10
   2.6 Q5. Location of data storage during the life of the project/ research activity ............... 11
   2.7 Q6. Retention period of Research Data .......................................................................... 12
   2.8 Q7. Factors that influence how Research Data is stored, managed, and/or shared ...... 13
   2.9 Q8. Security applied to Research Data .......................................................................... 15
   2.10 Q9. Approach to Data Sharing ..................................................................................... 16
   2.11 Q10: Arrangements for providing access to research data following completion ...... 17
   2.12 Q11: Data management issues and challenges encountered during the research lifecycle .............................................................................................................. 18
   2.13 Q12: Interest in data management training ...................................................................... 21
   2.14 Q13: Additional data management training ................................................................... 22
   2.15 Q14: Resources required to enable the researcher to manage their research data ..... 24

3. DISCUSSION .......................................................................................................................... 26

4. REFERENCES ......................................................................................................................... 27
Executive Summary

This report describes the findings of a web survey to determine current knowledge and application of data management practices among researchers at the London School of Hygiene and Tropical Medicine, and identify areas where further support is required. In total, 117 respondents, representing 16.25% of academic staff, completed the web survey between October 4th and November 18th, 2012. Key findings of the web survey include:

- Ninety-nine percent of LSHTM researchers that responded to the survey take steps to minimize the risk of data loss, creating multiple copies of their data for storage of different media, or copying it to School servers where it will be automatically backed up. Fifty-two percent of respondents store their data in 2-3 different locations, while an additional thirty-seven percent of respondents store their data in 4-8 locations. However, the respondent who stored their data on a local machine only is a cause for concern, suggesting a need for training in this area.

- Training and guidance is needed on retention periods for research data. Forty-five percent of respondents expressed uncertainty on the time period for which research data should be retained, while eleven percent believed it should be held for a period that was shorter than the 10 year retention period specified in the School policy.

- The majority of respondents (sixty-eight percent) recognized that external factors, such as the Data Protection Act, Freedom of Information requirements, and related obligations will influence their approach to data management and sharing. However, twenty-one percent of researchers were uncertain of the factors, suggesting that there is a need for further training in this area.

- Seventy percent of respondents indicated that they share their research data beyond the local project team during the life of a project. The majority of these researchers make their data available to other members of their research group and collaborating partners, while a small number of respondents share their data with third party data providers, funders, publishers, and others who express an interest.

- Respondents currently adopt a varied approach to data sharing: thirty-two percent of respondents store data within the department/project team and only share the data with researchers if it is requested; six percent indicate that research data will be made available through a project or other web site, and another six percent indicate that the data will be deposited with a third party data service or archive. Twenty-seven percent of respondents indicated that their research data cannot be made available in any form.

- Uncertainty on practices for data archiving were recognised as a key challenge by the majority of respondents (51%), followed by issues related to the production of data sharing agreements (30.7% of respondents), uncertainty on the documentation standards to apply (30% of respondents), and concerns regarding data security (27% of respondents).

- Training on creation and management of Data Documentation was the topic of most interest to survey respondents, followed by training on Data Storage, Developing a Data Management Plan, addressing Data Sharing obligations, and understanding funder requirements. More broadly, interest was expressed in training on data confidentiality, data capture using portable devices (PDAs, mobile phones), data mining, and other topics.
1. Introduction

1.1 Overview

This report describes the results of a web survey performed at the London School of Hygiene and Tropical Medicine to scope data management capabilities and requirements. The survey results contribute to the understanding of three factors:

1. Current knowledge and expertise related to data management among researchers at the School;
2. The challenges and issues that researchers encounter when managing and sharing their research data;
3. The resources that researchers require to help them to manage and share their research data more effectively.

Information gathered through the web survey, as well as face-to-face meetings with researchers, will be used to inform subsequent activities performed by the RDM Support Service to establish an infrastructure to assist researchers.

1.2 Design and implementation of the survey

The web survey was constructed between August – September 2012, progressing through several iterations of development and testing before its launch in early October. To scope the question list, the author drew upon the Data Asset Framework\(^1\) (DAF) and Digital Repository Audit Method Based On Risk Assessment\(^2\) (DRAMBORA) toolkit. In its final iteration, the survey was comprised of 15 questions\(^3\) consisting of a series of Boolean (yes/no), multiple choice and scaled responses. These were supplemented by comment boxes, in which the user could respond as appropriate.

1.3 Limitations of Survey results

The survey results provide a useful insight into current knowledge of data management, the challenges faced by researchers when managing and sharing research data, and the resources that researchers require to help them to perform their work more effectively. However, a number of factors affected the design of the survey methodology and limited the use of the survey results:

- **Information gathering method:** A decision was made to take a researcher-focused approach to information gathering, collecting high-level information on all research activities currently performed by the respondent. Evidence of this approach can be seen in Q3-11 that ask participants to provide a broad brush response on the type of data that they handle, software products in use, retention requirements, storage, third party requirements that apply to data. Although this approach provides information on a broad range of data, it is less accurate than the project-based approach encouraged in

---

the Data Asset Framework and makes it difficult to compare results to other DAF-based surveys.

- **Number of responses:** The Web Survey received 117 responses, representing feedback from 16.25 percent of academic staff employed at the School. Although the survey organizers were pleased with the response rate, there is an inherent danger of applying the results to the remaining 83.75 percent who did not complete the survey. The lack of response may represent a lack of engagement with RDM issues, an unwillingness to comment on a topic of which they have little knowledge, or simply that staff were too busy with other activities to complete the survey. There is a need for further investigation work to determine the views of the ‘silent majority’ who did not complete the survey.

2. **Web Survey Results**

2.1 **Overview**

The survey was published through the Bristol Online Survey tool (https://www.survey.bris.ac.uk/) on October 4th, with a closure date set for November 18th, 2012. The survey received 121 responses in total, four of which did not progress to the end of the process. The four semi-complete responses were excluded from the survey results, and the 117 remaining responses analysed.

2.2 **Q1. Research Faculty**

The purpose of the first question was to collect basic information about the School Faculty in which the respondent was located.

![Classification of respondents by Faculty](https://www.survey.bris.ac.uk/)

**Figure 1: Classification of respondents by Faculty**

The majority of respondents were located in the Faculty of Epidemiology and Population Health, with the remaining respondents being located in the Faculty of Public Health and Policy (PHP) or Faculty of Infectious and Tropical Diseases. One respondent selected the ‘Other’ radio box, indicating that they were co-located in ITD and EPH.

2.3 **Q2. Research Project Funder**

The second question sought to establish the funding sources that currently support research activities performed by the respondent. The responses were useful in identifying the combination of funding received by researchers and, in conjunction with other questions, respondent’s awareness of data management and sharing requirements.
Half of respondents were currently receiving funding from a single body (Figure 2); nearly a quarter from two funding sources; and the remainder from more sources. A small number of respondents (4%) indicated that they did not receive external funding, while an equal number indicated that they received funding from various Non-Governmental Organisations (NGOs) and/or charital organisations, but did not provide an exact number.

![Figure 2: No. of funders that currently support researchers](image)

The funding bodies selected by respondents from the multiple choice checkboxes are outlined in Figure 3.
Figure 3: Research Funders identified by respondents

The Bill and Melinda Gates Foundation funded the largest number of researchers in the survey (Figure 3), followed by the Wellcome Trust, and Medical Research Council, Department of Health UK, Department for International Development (DfID), and a number of other agencies.

Sixty-nine respondents identified other agencies that funded their research activities (Figure 4), including the NHS National Institute for Health Research (NIHR), European & Developing Countries Clinical Trials Partnership (EDCTP), and various unstated European Union Funding.

Figure 4: Other funding agencies that support respondent’s research activities

2.4 Q3. Description of Research Data

The third question sought to gather information on the type of research data that the respondent handled in their research activities. Respondents were asked to provide a broad overview of research data that they managed in all of their research projects. The question was separated into three components to determine the source (primary/secondary), content type (qualitative, quantitative), and the presence of personally identifiable information.

2.4.1 Does it contain primary data created during your research?

The purpose of the first question was to determine the source of the digital information used by the respondent in their research.
The majority of respondents indicated that their research involves the use of primary data only, while a small number indicated that their research utilizes secondary information only, or contained a combination of primary and secondary information.

2.4.2 Does it contain quantitative data?
The purpose of the second question was to determine the type of information handled by the respondent.

The majority of respondents indicated that their research data contained quantitative data, while a smaller number indicated that their data contained a combination of quantitative and qualitative information; or qualitative data only.

2.4.3 Does it contain personally identifiable information at any stage?
The purpose of the third question was to establish whether the respondent handled personally identifiable information in their research.
Figure 7: Data containing personally identifiable information

The majority of respondents indicated that they handled personally identifiable information at some stage in the research lifecycle, while a smaller indicated that they handled data that contains personally identifiable information in some projects, or did not handle personally identifiable information at all.

2.5 Q4. Commonly used software products

The purpose of the fourth question was to identify the software products that are most commonly used by respondents to perform research activities. Respondents were asked to list three software products that they most frequently used to create and manipulate their research data.

Figure 8: Commonly used software products

As shown in Figure 8, the most commonly used software tool is StataCorp’s STATA, followed by Microsoft Excel and Microsoft Access. There are also a number of lesser used software
applications that should be recognised and supported by the RDM Service in guidance and advice.

2.6 Q5. Location of data storage during the life of the project/ research activity

The fifth question sought to determine the storage facilities that respondents used to hold their research data during the period of active development and use.

Figure 9: Storage systems commonly used by respondents

The survey responses identified a large number of systems in use for storing research data during the primary stage of its creation, analysis and use.

An analysis of the number of storage options selected by each respondent (Figure 10) demonstrated that researchers were aware of the options available for data storage, holding their data in a number of different locations (however, it does not equate that the same data is held in these locations). More than half of respondents store their data in 2-3 of the named locations, and a quarter store their data in four or more locations. A small number of respondents indicated that they used one storage option only. Further analysis of these responses found that the majority of respondents had selected network storage system/servers maintained by collaborating institutions, network servers dedicated to the project at School, or Content/data management system operated by a project partner/collaborator. However, there was one respondent who indicated that they stored their data on their local disk drive only.
Figure 10: Distribution of respondent research data on various systems

Thirteen respondents indicated that they stored data in a number of additional locations, as shown in Table 1

<table>
<thead>
<tr>
<th>Location</th>
<th>No. of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Third party storage providers (Royal College of Surgeons, University of Bristol, Clinical Practice Research Datalink, and a 4th unstated provider)</td>
<td>4</td>
</tr>
<tr>
<td>Department Shared drive</td>
<td>2</td>
</tr>
<tr>
<td>LSHTM Secure Server</td>
<td>2</td>
</tr>
<tr>
<td>LSHTM SFX Server</td>
<td>1</td>
</tr>
<tr>
<td>Research data filed in hard copy at partner institution</td>
<td>1</td>
</tr>
<tr>
<td>Secure PC in a locked room</td>
<td>1</td>
</tr>
<tr>
<td>Project servers</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1: Additional storage locations

2.7 Q6. Retention period of Research Data

The sixth question sought to establish the number of years for which respondents were required to retain their research data after the end of the project.
Figure 11: Understanding of the retention period for research data

The survey results found that a large number of respondents were uncertain on the time period that the research data should be retained for, while other respondents specified a retention period between 6-20 years, or indefinitely.

Eleven respondents provided a description of their retention requirement in the ‘Other’ field:

- Five respondents indicated that the retention requirements varied between different projects. One respondent elaborated on this, indicating that ‘The Lancet’ required them to store data for 5 years\(^4\), whereas clinical trial requirements were required to be stored for 20 years.
- One respondent indicated that the data would be made available in the public domain so would, theoretically, be retained indefinitely.
- One respondent indicated that their research data would be retained indefinitely, but that personally identifiable information would be removed after 1 year.
- One respondent indicated that the retention period was dependent upon the lifetime of the consortium in which they were participating and plans for a follow-up project.
- Two respondents indicated that they were not required by their funder to retain their data after the end of the project, and a third indicated that their data would be transferred to the sponsor.

2.8 Q7. Factors that influence how Research Data is stored, managed, and/or shared

The seventh question sought to identify the factors recognised by the respondent that influence how Research Data is stored, managed, and/or shared. The first component of the question was a multiple choice that asked respondents to indicate if there were legal, regulatory, or confidentiality issues that influence data management and sharing.

\(^4\) This is incorrect - The Lancet specifies that data should be retained for a ten year period (http://www.thelancet.com/lancet-information-for-authors/after-publication)
Respondent who had indicated that their approach to data management and sharing was influenced by external factors were asked to identify the relevant legal, regulatory, or confidentiality issue(s) that affected their research data\(^5\).

As a follow-on question, respondents were asked to identify additional obligations with which they must comply. Fifteen respondents completed the field, in total.

---

\(^5\) Several options were provided (Data Protection Act, Freedom of Information, Funder requirements, NHS ethics/NRES, Environmental Information Regulations, School requirements, and Institutional policy of a third party), of which they could select one or more.
Table 2: Additional factors that influence the approach adopted to manage and share research data

A final respondent indicated that they found it difficult to store data on the LSHTM network when overseas.

2.9 Q8. Security applied to Research Data

The eighth question sought to determine whether respondents applied measures to secure their Research Data and, if so, the type that they applied. The first component was a multiple choice question that sought to identify if the respondent applied security arrangement to their Research Data.

Respondents that answered in the affirmative were asked to indicate one or more security measures that were applied.
Nine respondents provided further detail in the free text ‘other’ field:

- “agreement of database”
- “locked cabinets to close paper based data”
- “not for excel; need to log in to access data managed by partner”
- “Registered access to online resources”
- “Restricted access to network folder”
- “Schools secure server”
- “secure drive on the school network”
- “stored in H drive, so password protected”
- “we have not yet received signed the data sharing agreement on our most sensitive project, but all the above will be used”

2.10 Q9. Approach to Data Sharing

The purpose of question 9 was to determine the approach that respondents took to sharing their research data during the life of the project. Although researchers may not formally make their data available for use by others during the development phase, it was considered possible that they would share their research with third parties on an informal basis.

First, respondents were asked to indicate whether they shared their research data beyond the project team during the life of the project/research activity.
Respondents that answered in the affirmative were asked to identify one or more parties from a pre-defined list with whom they would share the data (Figure 17).

![Figure 17: Indication of third parties with which researchers may share their data during the life of a research project](image)

Finally, respondents were asked to indicate other stakeholders with whom they shared the research data. Twelve respondents provided further information. However, in several cases, it appeared that they were referring to post-project activities, rather than data sharing during the active life of the project:

- “only selected anonymised files are sent as necessary”
- “Anyone who expresses an interest if authorised by Principal Investigator”
- “anyone who submits request if approved”
- “available to researchers on request following publication”
- “DSMB” (Data and Safety Monitoring Board)
- “only in summary form”
- “Other bona fide researchers who want to use our data”
- “Selected others who express an interest”
- “So far a number of MSc students have used data for their projects”
- “students at the school and at collaborating institutions”
- “Transcribers receive a password-protected copy of each audio file (qualitative interviews) and are bound to a confidentiality agreement.”
- “We host our own data sharing system”

### 2.11 Q10: Arrangements for providing access to research data following completion

The purpose of question 10 was to determine the arrangements, if any, that had been made to make data available, in full or in part (e.g. an anonymised subset) for access following completion of the work.
Thirteen respondents provided additional information in the free text ‘other’ field:

- Four respondents indicated that decisions on data sharing were made on a project-by-project basis, potentially encapsulating all of the provided options;

- Three respondents indicated that data would be made available on request. There was some variation in the formality of the process – one respondent indicated that data could be accessed by submitting a Data Access Request to the Data Controller (whom may not be a member of the project team), whereas other respondents adopted ad-hoc approaches to data sharing: “If someone asked me, or my boss, we’d dig out the data for them”.

- Two respondents indicated that data sharing was planned, but they were uncertain on the approach that would be adopted and would make arrangements at a later date;

- One respondent indicated that the data would be made available through the GenBank NIH genetic sequencing database (http://www.ncbi.nlm.nih.gov/genbank/) or similar services, and a second indicated their data was made available through LSHTM’s FreeBird system.

- One respondent indicated that data was owned by a third party and, as a result, they were not responsible for making it available.

- Two respondents indicated that data had been made publicly available. However, they did not indicate where it was held.

2.12 Q11: Data management issues and challenges encountered during the research lifecycle

The eleventh question sought to identify the issues and challenges encountered by researchers when creating, managing, and/or sharing research data. The respondent was provided with ten options, of which they could select one or more. Other issues could be documented in a free text field.
Figure 19: Data management challenges and issues encountered during the research process

The most common issue specified was an uncertainty on data archiving practices (Figure 19), followed by issues when preparing data sharing agreements, uncertainty on documentation standards to apply, and security concerns.

Twenty-five respondents utilized the other field to raise other challenges and concerns. These were classified into six categories:

1. **Data management practices**
   a. “No guidance as to specific expectations in my area”

2. **Data development**:
   b. “Challenges in establishing common shared databases across partners”
   c. “data entry protocols and database creation/management”
   d. “Difficulty in finding the most up-to-date versions of the databases I’m supposed to use”

3. **Data storage**
   e. “Physical storage limitations”

4. **Data sharing during project lifetime**:  
   f. “Finding a safe way to share data between team members in different institutions”
   g. “Multi centre remote data entry/access”
h. “Secure file transfer from field site to LSHTM unavailable”

5. Data archiving or sharing following project completion:
   i. “not sure where to store raw data (questionairres)”
   j. “time for archiving”
   k. “Getting data ready to be shared requires a huge amount of time and resources that is not adequately budgeted for. Part of the problem lies in motivation - if I’m not required to make data public why spend the time preparing the data and documentation”
   l. “lack of time to prepare datasets”

6. Project specific issues:
   m. “Reaching agreement with national bodies where projects are sited”
   n. “Time spend in helping investigators to analyse our data”
   o. “providing technical support to research groups in Africa”
   p. “Instability of NVivo”

The remaining nine responses indicated that the respondent did not encounter any data management challenges or issues during their research.

As a follow-on question, respondents were asked to provide further detail on the issues and challenges that were encountered. The responses were subsequently classified into six categories:

1. Challenges related to availability of expertise:
   a. “Lack of expertise within Dept on data management”
   b. “i work overseas where the main challenge is in employing and maintaining suitably trained and capable staff for designing and managing research databases - projects also don’t provide sufficient and realistic funds to do this”
   c. “working overseas with staff who have very limited technical capacity for data management--would really help to have some central lshtm facilitation on this”
   d. “Too many people have managed data over the years and documented their work inadequately”

2. Challenges related to grant application development:
   e. “The lack of official school documents for grant applications, People mis-understanding what is an what is not possible, Sharing data afterwards is new issue, funding unsure.”

3. Challenges related to the data management infrastructure:
   f. [Confusion regarding] “development of our own repository software (LAMP based)”
   g. “We have had problems with simultaneous use of MS Access stored on the shared drive”
   h. “We store most data at the Royal College of Surgeons after investing in appropriate hardware / software and institutional security policies”

4. Challenges related to data sharing during project lifetime:
   i. “Local field staff have yahoo or hotmail email, internet is very poor, LSHTM secure file transfer system does not work for transferring the files across so they are then sent unsecurely over these other email systems.”
5. **Challenges related to data archiving practices:**
   
j. The data are deposited in what is essentially a summary form. Unclear what is supposed to happen to the raw data files - they occasionally need to be checked but there is no archive for these so I just keep them (usually on CD).

6. **Challenges related to data sharing:**
   
k. “We have data from a RCT that should be made open access but we don’t know where to put it. Also, the metadata are probably not currently sufficient to allow others to use the data and we are interested in guidance on what is required.”
   
l. “I keep the originals and all versions that are shared, but have not created a version for general access as the data files are complicated”
   
m. “Finding and operating software which is compliant with applicable standards.”

2.13 Q12: Interest in data management training

The purpose of the twelfth question was to gauge interest in data management training on topics related to different stages of the research data lifecycle. Respondents were provided with a list of nine topics and asked to rate their interest in training on the topic on a likert scale. The responses for each course were subsequently mapped to compare the level of interest in each training course, shown in Figure 20 - Figure 23.

---

Figure 20: No. of respondents with a strong interest in the training course

Figure 21: No. of respondents with a moderate interest in the training courses

---

7 Results indicating “No opinion” have been omitted
Figure 22: No. of respondents with a slight interest in the training courses

Figure 23: No. of respondents with no interest in the training courses

By combining the number of responses that indicated a strong or moderate interest in the training courses, it may be determined that respondents have most interest in a training course on data documentation (89), data storage (84), developing a data management plan (80), data sharing (78), followed by understanding funder requirements for RDM (71), metadata creation (57), ethics and consent issues related to research data (56), copyright and IPR issues related to research data (55) and data citation (54).

2.14 Q13: Additional data management training
As a follow-on question, respondents were asked to identify other training courses that they would find useful. In total, twenty respondents indicated training that they would find useful. These have been organized into seven categories:
<table>
<thead>
<tr>
<th>Category No.</th>
<th>ID</th>
<th>Summary description</th>
<th>Quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Broad Skills</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>Capacity building</td>
<td>“Capacity building for partner institutions”</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>Good Clinical Practice</td>
<td>“How can LSHTM be GCP compliant”</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>Qualitative data</td>
<td>“Training specific to qualitative data”</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Information Governance</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>Information Governance</td>
<td>“Information governance legislation”</td>
</tr>
<tr>
<td>E</td>
<td></td>
<td>Confidentiality training</td>
<td>“...school needs to provide compulsory data management and confidentiality training for all staff. Offering different levels so staff know when they should be doing what, not just one course for all. The majority of staff here do not even know there is a confidential server.”</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>Confidentiality training</td>
<td>“Proper confidentiality training and data management training provided with input from researchers, not just theory”</td>
</tr>
<tr>
<td>G</td>
<td></td>
<td>Regulations</td>
<td>[unstated] “Regulations”</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>Data Capture</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td></td>
<td>Web-based data entry</td>
<td>“Web based data entry systems”</td>
</tr>
<tr>
<td>I</td>
<td></td>
<td>Portable data capture</td>
<td>“use of PDAs and form scanning systems for field data collection”</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>Data analysis and manipulation</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td></td>
<td>SQL DB design</td>
<td>“Help in creating SQL databases”</td>
</tr>
<tr>
<td>K</td>
<td></td>
<td>SQL DB design</td>
<td>“MySql courses”</td>
</tr>
<tr>
<td>L</td>
<td></td>
<td>SQL DB design</td>
<td>“I'm desperate to learn how to efficiently enter data, check quality and learn about 'data cleaning'. Prior to entry though, I think there's a lot to learn about how to create a data base and how to design the questionnaire to facilitate data entry and coding.”</td>
</tr>
<tr>
<td>M</td>
<td></td>
<td>SQL DB design</td>
<td>“Setting up databases which are GCP compliant. Setting up databases full stop. &quot;Managing consistency checks for data accuracy. Coding of variables for ease of data sharing. Labelling variables appropriately. How to have data collapsed or by individual etc.”</td>
</tr>
<tr>
<td>N</td>
<td></td>
<td>STATA</td>
<td>“Data management in STATA training”</td>
</tr>
<tr>
<td>O</td>
<td></td>
<td>Data mining</td>
<td>“Data mining”</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>Metadata creation</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td></td>
<td>Data Documentation Initiative</td>
<td>“Standardisation of data resource - e.g. DDI”</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>Data archiving</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td></td>
<td>Best practice for data archiving</td>
<td>“Best methods for archiving data.”</td>
</tr>
<tr>
<td>R</td>
<td></td>
<td>Image storage</td>
<td>“How to store images”</td>
</tr>
<tr>
<td>S</td>
<td></td>
<td>Best practice for data archiving</td>
<td>“Options for in-house or third party archiving of data; what data to archive”</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>Data sharing</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
<td>Data sharing during project</td>
<td>“Sharing data from LSHTM network with collaborators”</td>
</tr>
</tbody>
</table>

Table 3: Additional data management training required
2.15 Q14: Resources required to enable the researcher to manage their research data

Finally, respondents were asked to state the resources that they believe the School should provide to enable them to more effectively and easily manage their research data. Many of the responses replicated information provided for previous questions. However, there are a number of additional requirements expressed.

<table>
<thead>
<tr>
<th>Category No.</th>
<th>ID</th>
<th>Summary description</th>
<th>Quotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>Capacity Building</td>
<td>“Support for capacity building at partner institutions”</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Data archive</td>
<td>“Collection of questionnaires and validated scales to allow students and new researchers to produce high quality questionnaires”</td>
</tr>
<tr>
<td>2</td>
<td>C</td>
<td>Central RDM Service</td>
<td>“In my LSHTM group there are no specialist data managers - - nor in the dept to my knowledge - - and it’s a skill that people have to source themselves for each project. I’d be useful to have a central resource for this even if just for advice.”</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
<td>Single resource</td>
<td>“Clear guidance on the entire process - a one-stop resource that researchers can go to (ie. documentation but also staff support).”</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>Data storage options</td>
<td>“Clearly explain the options for backing up research data in School servers. There are several options at the moment, eg iFolder, saving into my own drive on the school server etc but am not sure how much storage capacity is available to me.”</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>Discussion forum</td>
<td>“Forum for discussion of techniques and best practices”</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>Funder requirements</td>
<td>“Guidance on data requirements of common funders.”</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>Consortium/funder agreements</td>
<td>“Guidance on agreements (mainly DTA) but both home grown and external ones.”</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td>Regular clinic sessions</td>
<td>“As well as formal training, it would be helpful to have an advisory service e.g. a once a week clinic where you can go with specific questions / problems”</td>
</tr>
<tr>
<td></td>
<td>J</td>
<td>Advice on request when planning applications</td>
<td>“Go to advice, advice as a new project is being planned and initiated”</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>Advice on request when developing data management &amp; sharing plan</td>
<td>“Support in designing a data management and access plan for qualitative and quantitative data that meets requirements of Gates Foundation and LSHTM’s ambitions for data access.”</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>Advice on database development</td>
<td>“…Technical support in database design-setting up databases for research projects overseas Advice on use of PDAs, scanning systems, etc for data collection”</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>Advice on metadata creation and retrospective cataloguing</td>
<td>“… what metadata is necessary when making data open access”</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Support group</td>
<td>“… Maybe a data users group, instead of the current set up”</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>--------------------------</td>
<td>------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>Assistance in developing databases</td>
<td>“Access to programme-help (rather than IT help) eg Access experts who can be booked for a day or two to help with a specific task”</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Education and training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>Database development</td>
<td>“RDBMS training - advanced techniques inc. spatial queries”</td>
<td></td>
</tr>
<tr>
<td>Q</td>
<td>Database development and quality controls</td>
<td>“A course in creating a database for data entry from a questionnaire, how to check double entry and reconcile errors quickly and effectively, how to summarize and cross check data, quality control in the field and after entry and appropriate coding for different types of variables (multiple response, skips, ‘other’ and missing data).”</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>Training overseas staff</td>
<td>“training or apprenticeships for overseas research staff”</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Support staff that may be allocated to perform data management activities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>Database developers</td>
<td>“a pool of data entry clerk”</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>Database developers</td>
<td>“Are there staff at the School who have the expertise to build sophisticated SQL databases for managing large datasets accessible to multiple research partners? I’m developing a project that would require this and have only found people outside the School who I can include on the grant as collaborators to do this critical piece of work. Staff at the School who could support research projects in this way would be an enormous asset.”</td>
<td></td>
</tr>
<tr>
<td>U</td>
<td>Metadata experts able to assist with creation of MD for new and existing data</td>
<td>“...are there any expert staff within the school who can help with retrospective creation of metadata and/or packaging data for open access.”</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Data policies and templates</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>IG / system security policies for secure service</td>
<td>A secure data server and clear IT system security policies / information governance structures</td>
<td></td>
</tr>
<tr>
<td>W</td>
<td>School-wide Data Sharing policy</td>
<td>“...A data sharing policy of the school would also help. This applies equally to those who wish to use the data of other research projects within the school. There should be a policy on this to facilitate research. Much of the data collected by the school lies idle after the completion of a project but there is great potential for using the data for additional research questions”</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Enhancements to existing technical systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>Confidential data server</td>
<td>“The confidential server needs to be speeded up, at present it is unusable for large datasets (about 10 times slower than my local drive”</td>
<td></td>
</tr>
<tr>
<td>Y</td>
<td>Secure file transfer</td>
<td>“Secure file transfer that works from field sites in Africa. VPNs from collaborators computers in Africa to here so that LSHTM staff can access locally stored files which are not backed up in the same way, nor are they secure in the same way i.e. security to the building is not the same.”</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>SQL templates for database development</td>
<td>“LSHTM should provide a common SQL template for all databases so that research staff become familiar with the</td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Resources required to enable researchers to manage their research data

3. Discussion

The survey results provide a useful insight into current knowledge of data management, the challenges faced by researchers when managing and sharing research data, and the resources that researchers require to help them to perform their work more effectively.

LSHTM Researchers demonstrated a good understanding of the practicalities of the research process itself, providing a clear indication of the type of data with which they worked (Figure 5, Figure 6 & Figure 7), software tools that they utilized within their research (Figure 8) and the location of data storage (Figure 9). However, the survey found significant variation in data management knowledge and expertise. Differences in understanding of data management issues emerge when specific requirements and practices of research data management were taken into account. Although the majority of respondents (77 percent in Figure 12) expressed confidence on the factors that impacted their approach to data management, almost half of respondents indicated they did not know how long data should be stored for, and another eleven percent chose a retention period that was shorter than that established by School (Figure 11). Further problems are likely to occur following completion of the research project: many researchers recognised a need for their data to be curated and preserved in the long-term, but were uncertain of the practices that should be applied. Although there is no evidence that this has resulted in data being deleted before the allocated time/stored for a longer period than is necessary or managed/shared in an inappropriate manner, it identifies a need for training in these areas to ensure that knowledge of retention requirements and curation practice is spread among all research staff, rather than focused upon specific individuals in a project.
4. References

