ANALYSIS OF RESULTS FROM THE RESEARCH DATA PRESERVATION SURVEY – FULL REPORT

CONTENTS

1. Background ........................................................................................................................................... 1
2. Introduction ............................................................................................................................................. 1
3. Summary ............................................................................................................................................... 2
4. General Observations .......................................................................................................................... 3
5. PhD-specific observations .................................................................................................................... 8
6. Non-PhD specific observations ............................................................................................................ 16
7. Research experience-specific observations ......................................................................................... 19
8. Data Preservation observations ........................................................................................................... 27
9. Support and training observations ....................................................................................................... 28
10. Other information management services .......................................................................................... 29
11. Conclusions and recommendations .................................................................................................... 30

1. BACKGROUND

1.1. The Digital Communication Enhancement (DICE) project has been funded by JISC to produce training materials to support digital research data preservation at LSE, and to make these materials available to the UK HE community. In order to focus the materials on the needs of the researchers, an online survey was conducted.

2. INTRODUCTION

2.1. A survey was conducted with the aim of discovering the awareness of and current practice relating to research data preservation at LSE. We targeted PhD research students and research-active staff by inviting these groups to respond to an online survey that was active from 14 to 25 March 2012.
3. **SUMMARY**

3.1. This questionnaire has shown the general lack of current awareness amongst LSE researchers of digital data preservation. It has also shown that there are cultural challenges to address as well as more technical training that is needed if researchers are to send their research data and materials into the future with confidence.

3.2. Researchers are using a wide range of technologies to create research data in a variety of formats, so as well as the general principles of digital preservation, we will use case studies of the major data types in the materials that DICE produces.

3.3. The survey shows that researchers have a preference for informal support and autodidactic means of learning. Although no one method is suited to everyone, we will target the methods of training and support that showed the most popularity, and use re-purposable materials wherever possible.

3.4. There is a presumption that data preservation and public availability of that data go hand-in-hand, however we identified a need for data preservation without public availability. There seems to be an acceptance by the researchers that this restricted preservation will take place under the control of the researchers themselves and on their own systems. Nevertheless, training and support should be available to these researchers even if the School does not provide the repository for the data.

3.5. In general, research data are not well documented: researchers using others’ data have reported difficulty understanding it, or even understanding their own data after a few years! The documentation of data is important when preserving it so this will be included in the DICE outputs.

3.6. Several issues emerged from the survey that are beyond the scope of the DICE project but which the School may want to consider. Specifically:

   3.6.1. the experience profile of researchers at LSE shows a large number with less than 9 years’ experience but then tapering rapidly to give a low number of more experienced researchers. The School may want to consider whether this balance is correct and whether its succession-planning for research continuity is adequate;

   3.6.2. three-quarters of researchers recorded “self-funding” as a means of funding their research. The School may wish to investigate this statistic further;

   3.6.3. 80% of LSE researchers do no data management planning. In an age of dependence on digital data this is a situation that needs addressing. A bid has been submitted to JISC for funding to extend the work of DICE, but alternative plans should be made if the bid is unsuccessful;

   3.6.4. researchers are using cloud-based services as part of the tool set available to them for document maintenance, sharing and backup. LSE’s policies and support for these services seem somewhat “behind the curve” compared with user behaviour and likely future demand. The School may want to re-examine its approach in this area;

   3.6.5. some researchers, particularly those in the 6-12 year experience band, make use of departmental servers. The School might want to examine its costs and policy regarding the use of departmental servers versus the central network storage.
4. GENERAL OBSERVATIONS

4.1. There were 167 responses to the survey, 93 (56%) from PhD students and 74 from staff.

4.2. All Academic Departments and most of the larger Centres were represented, making the responses a reasonable cross-section of the community.

4.3. Respondents were asked how much research experience they had. Answers are in 3-years bands except the first (less than 12 months) and last (24 years or more).

4.4. We can see that nearly 80% of researchers at LSE have less than 9 years' experience. This may have implications for the School that are beyond the scope of this project (e.g. does the School have sufficient numbers of very experienced researchers? Is succession planning for the most senior research posts adequate?), but for the DICE project there are implications for our approach to and expectations of researchers to whom we offer support that will be explored as we progress through this report.

4.5. We can see that nearly 80% of researchers at LSE have less than 9 years' experience. This may have implications for the School that are beyond the scope of this project (e.g. does the School have sufficient numbers of very experienced researchers? Is succession planning for the most senior research posts adequate?), but for the DICE project there are implications for our approach to and expectations of researchers to whom we offer support that will be explored as we progress through this report.
4.6. 65% of LSE researchers spend more than 50% of their time conducting research. Bearing in mind that nearly half of respondents say that they are at least partly self-funded and therefore, we can assume, need to earn money, it is not surprising that full-time researchers are not as common as might be expected but it does indicate a high level of commitment to research that the respondents must have.

4.7. 65% of LSE researchers spend more than 50% of their time conducting research. Bearing in mind that nearly half of respondents say that they are at least partly self-funded and therefore, we can assume, need to earn money, it is not surprising that full-time researchers are not as common as might be expected but it does indicate a high level of commitment to research that the respondents must have.
4.9. The sources of funding for research are diverse, with the ESRC being the major Research Council sponsor (bear in mind that the survey measured respondent numbers, not monetary value). Another significant source of funding is “Other”, which covers the EU and country governments, charities and endowments. LSE scholarships are major sources of funding, but largest of all (in number if not monetary value) is self-funding, with 44% of respondents contributing partly or wholly to the cost of their research.

4.10. Almost 90% of respondents don’t know about or don’t use data management planning (DMP) (Q8). This is not a surprise: since data preservation is just one topic within DMP, there wouldn’t be much point in running the DICE project if we had found otherwise. It does, though, provide useful evidence to justify a wider project to implement and support DMP as a technique for researchers to adopt.

4.11. LSE does not provide cloud-based services and does not encourage their use. Nevertheless, researchers are clearly adopting Internet-based services for some of the functions they need, whether for data creation, organisation or backup, as the responses to Q9, Q10 & Q13 show:
Q9. Where do you store the working version of the digital material and data you use or create in your research?

<table>
<thead>
<tr>
<th>Location</th>
<th>Number of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>On the hard drive of my computer</td>
<td>119</td>
</tr>
<tr>
<td>On a departmental server</td>
<td>24</td>
</tr>
<tr>
<td>On the LSE network (e.g. the H: drive)</td>
<td>72</td>
</tr>
<tr>
<td>On a separate memory device (e.g. external hard drive or memory stick)</td>
<td>100</td>
</tr>
<tr>
<td>In an Internet-based file management service (e.g. Google Docs, Dropbox)</td>
<td>72</td>
</tr>
</tbody>
</table>

Q10. How do you organise the digital material and data you use or create in your research?

<table>
<thead>
<tr>
<th>Organisational Method</th>
<th>Number of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>I keep it in a single folder</td>
<td>19</td>
</tr>
<tr>
<td>I use a folder structure</td>
<td>141</td>
</tr>
<tr>
<td>I use file naming conventions</td>
<td>64</td>
</tr>
<tr>
<td>I use software that supports version control</td>
<td>14</td>
</tr>
<tr>
<td>I use a database to organise my files</td>
<td>41</td>
</tr>
<tr>
<td>I keep my files in an Internet-based file management service</td>
<td>46</td>
</tr>
</tbody>
</table>
4.15. That researchers are “voting with their feet” on Internet storage is significant for DICE in catering for support in this area.

4.16. Responses to Q9 show a small but significant use of departmental servers for storage for research data. Looking at the graph analysed by years of experience show the majority of these users are in the 6 to 12 years bracket, while the less experienced and more experienced are using alternatives. Although outside the scope of this project, the School might want to examine its policy on the use of departmental servers versus the central network storage.
5. PHD-SPECIFIC OBSERVATIONS

5.1. Responses to Q6 show that PhD students are creating a richer set of research materials than the staff researchers, with the categories “Audio”, “Still image”, “Video” and “GIS, mapping, CAD and 3D models” being more popular with PhD students than staff. But we also see a distinct reluctance to make this data publicly available (Q14):

5.2. And when asked about the barriers to public availability a worryingly large proportion could not see why it would be necessary (Q17):
5.4. There were, though, clearly other reasons for not making data public. There will be times when the concern is perfectly valid but other times when some training could help overcome the barriers. It would be interesting to know the balance between those who have a natural reluctance or even fear of making their data available, and those who have valid and insurmountable justification for it. The author’s intuition on this is that there is a significant number of PhD students who need encouraging into a more open frame of mind and will be happy to make data public if they can be convinced that it is to their advantage. This means that the training materials produced by DICE should include an element of culture-change as well as practical steps to make data publicly available.

5.5. There is a noticeable difference between PhD and non-PhD researchers in the storage of the working versions of their data (Q9). This question allowed multiple responses, and the 387 answers exceeds the 167 respondents by a factor of 2.3, indicating that it is common for researchers to keep documents in different places:
5.7.

5.8. PhD students seem to have a preference for more mobile technologies such as external drives and Internet-based systems. We could also speculate that their use of computer hard drives may well be more laptop than desktop-based. There is certainly a reluctance to use LSE servers and storage.

5.9. When asked about the reuse of data, quite a large number noted the difficulty of understanding the material or data they were re-using:
5.10. But despite this experience, there is a high proportion that do not document their own data:

5.11. This will clearly have to be addressed in the training materials from the DICE project, as there is little point in preserving data that cannot be understood by subsequent users.

5.12. 87 of the 93 PhD respondents (i.e. 94%) either don’t know about or don’t use data management plans. Since research data preservation is part of data management planning, this confirms that
Awareness and planning for preservation is extremely limited and DICE has much work to do to assist the trainers and the students to improve.

### Q8. Does the research you are currently working on have a data management plan? PhD only

<table>
<thead>
<tr>
<th>Options</th>
<th>Number of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don't know what a data management plan is</td>
<td>43</td>
</tr>
<tr>
<td>I don't know if my project has a data management plan</td>
<td>8</td>
</tr>
<tr>
<td>My project does not have a data management plan</td>
<td>36</td>
</tr>
<tr>
<td>A data management plan was written as part of the bid but has not been updated</td>
<td>0</td>
</tr>
<tr>
<td>My project has a data management plan that covers the duration of the project</td>
<td>0</td>
</tr>
<tr>
<td>My project has a data management plan that extends beyond completion of the research</td>
<td>1</td>
</tr>
</tbody>
</table>

### 5.15

Answers to Q10 shows that there is recognition by most users of the need to organise files, though the sophistication of the solutions varies widely. The question allowed multiple responses and the 325 answers is 3.5 times the number of PhD respondents, indicating that users implement a range of techniques. Responses to “I use a database to organise my files (e.g. Lightroom or Picasa for images, EndNote or Zotero for references)” was surprisingly low given the expected use of references in research. It would be interesting to enquire further into the reasons for this. By contrast, the use of Internet-based file management services is quite high. The choice of organisational technique will almost certainly be decided on for reasons of immediate use. The DICE training materials should include consideration of longer-term factors including the needs of data preservation.
5.17. Responses to Q13 indicate a mixed approach to backing up data. There is still a worrying number of people not making backups, and these are mostly PhD students. This shouldn't be too difficult to correct provided they don’t insist on learning the hard way about disc crashes and data loss! It is interesting to see that there is a significant number of students using Internet backup services, and the same number maintaining more than 1 backup copy. The DICE project will need to include material on back-up practice and how this relates to data preservation in the long term.
We can see a noticeable difference between PhD and non-PhD researchers when questioned about the deterrents to making data publicly available (Q17). The PhD students see legal, ethical and reputational issues as more of a barrier than the non-PhDs. Whatever the reasons for this, the DICE project should include materials that help researchers quantify the risks and opportunities around public availability so that informed decisions can be made.
Q17. What are the main factors that could deter or prevent you from making your material and data publicly available?

![Bar chart showing factors that prevent public availability of research materials]

- Contractual factors
- Commercial sensitivity
- The data cannot be anonymised
- Intellectual property rights issues, including...
- Ethical issues
- Reputational considerations
- The cost of preserving material and data
- Technological issues
- There is no reason to make it available
- I don’t know how to make it available

Number of respondents:

- Non-PhD only
- PhD only
6. NON-PHD SPECIFIC OBSERVATIONS

6.1. The non-PhD respondents differ in several respects to the PhD students. We have already noted that PhD students are using more diverse technologies in the research data (Q6). Since the non-PhD respondents generally have more research experience, their responses to Q7 concerning problems encountered during the reuse of data should carry more weight:

---

Q7. If you have ever re-used your own material created in previous research, or used digital material or data created by others, have you encountered any of the following issues? Non-PhD respondents only

<table>
<thead>
<tr>
<th>Issue</th>
<th>Number of Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not re-use mine or others’ material or data</td>
<td>19</td>
</tr>
<tr>
<td>Obsolete file format (e.g., floppy disc)</td>
<td>20</td>
</tr>
<tr>
<td>Obsolete medium (e.g., obsolete computing system)</td>
<td>21</td>
</tr>
<tr>
<td>Incompatible computer code (e.g., obsolete operating system)</td>
<td>15</td>
</tr>
<tr>
<td>Intellectual property rights (e.g., copyright or licensing)</td>
<td>13</td>
</tr>
<tr>
<td>Security/Confidentiality/privacy</td>
<td>13</td>
</tr>
<tr>
<td>That we not encountered any issues when re-using material/data</td>
<td>21</td>
</tr>
</tbody>
</table>

---

6.2.

6.3. 21 of 115 (18%) had not encountered any issues re-using data, 30% had had technical difficulties due to obsolescence, and 41% had had “soft” issues (understanding the data, intellectual property, or security/confidentiality/privacy). Only 11% did not reuse data. The implication for DICE is that we must cover actions that can mitigate these issues when our own data is preserved for others to use.
6.4. It is interesting to see that whether you are a PhD student or not makes little difference to whether a researcher documents data. Nor does it depend on experience, with those not documenting or describing their data being distributed evenly through the experience categories:

6.5.

6.6.

6.7. DICE will need to include advice on the documenting of material and data, targeting all categories of researcher irrespective of their experience.

6.8. We can see that non-PhD researchers are more decided and more willing to make their data publicly available. This may well be because they are more confident about doing so and more aware of the advantages to them if they do:
6.9. As previously noted, PhD students perceive legal, ethical and reputational issues as a greater barrier than non-PhD respondents (Q17), so DICE should investigate whether those more willing to make their data publicly available could provide useful case study material to encourage the PhD students.
7. RESEARCH EXPERIENCE-SPECIFIC OBSERVATIONS

7.1. The data show a tendency for newer researchers to use a greater range of technologies in their research, especially the use of video as data, and online services for document creation, sharing and backup (Q6, Q9, Q11, Q13).

7.2. 

7.3.
Q11. How do you share your research material and data?

- Use a shared folder on the local network
- Use a shared folder on the University intranet
- Use a Web-based or online service
- Use email to exchange documents
- Use an FTP/FTP server
- Use a Web-based repository
- I do not need to share material or data

24 years or more
at least 21 years but less than 24
at least 18 years but less than 21
at least 15 years but less than 18
at least 12 years but less than 15
at least 9 years but less than 12
at least 6 years but less than 9
at least 3 years but less than 6
at least 1 year but less than 3
less than 1 year

7.6.

Q11. How do you share your research material and data? proportioned to 100%

- Use a shared folder on the local network
- Use a shared folder on the University intranet
- Use a Web-based or online service
- Use email to exchange documents
- Use an FTP/FTP server
- Use a Web-based repository
- I do not need to share material or data

24 years or more
at least 21 years but less than 24
at least 18 years but less than 21
at least 15 years but less than 18
at least 12 years but less than 15
at least 9 years but less than 12
at least 6 years but less than 9
at least 3 years but less than 6
at least 1 year but less than 3
less than 1 year

7.7.
7.10. Researchers in the 6-9 year group tend to document their data more than the less experienced and, somewhat surprisingly, the more experienced researchers (Q12).
7.13. It is the researchers with less than 12 years experience that do not backup their data, though
the overall number that are at risk from data loss in this regard is small (Q13, above).
7.14. The more experienced the researcher, the more likely they are to be willing to make their data publicly available (Q14).

7.15.

7.16.

7.17. Researchers with less than 9 years’ experience are more likely to perceive legal, ethical and reputational issues as a greater barrier to public availability (Q17).
Q17. Deterring factors to public availability?

7.18.

Q17. Deterring factors to public availability? proportioned to 100%

7.19.
7.20. Researchers with 9 or more years’ experience are more likely to have used a data archiving service, though the number is still very small:

7.21.
8. DATA PRESERVATION OBSERVATIONS

8.1. Researchers still produce large amounts of text during their work, but increasingly use diverse media, such as images, sound and video, too. Computer programs are also a significant product of research at LSE. The materials produced by DICE will need to cover the preservation of all types of media.

8.2. Although organisational techniques vary, many researchers use the operating system’s built-in folder structure and file naming ability to organise their research material. Building on this to give advice about effective organisation for preservation purposes should be part of the DICE output.

8.3. Only a small number of researchers document their research data and materials, so encouraging the routine documentation of work will need to be a priority for preservation and reuse purposes.

8.4. Researchers perceive many barriers to making their data publicly available (Q17). Some of these may just be a natural caution in the face of something that has not been tried or where techniques to overcome the barriers are unknown to the researchers. The DICE materials will need to address these concerns as well as the perception that there is no reason to make the data/material available.

8.5.

Q17. Deterring factors to public availability?

![Chart showing deterring factors to public availability](chart.png)
9. SUPPORT AND TRAINING OBSERVATIONS

9.1. Question 21 on the survey was about respondents’ preferred methods of support and training delivery. Respondents with less than 9 years research experience tended to respond more positively to this set of questions, which suggests that the more experienced researchers will be harder to train and support in matters related to research data preservation. But since only 20% of LSE researchers fall into this category we could decide not to target this group specifically.

9.2. Respondents were asked to rate the likelihood of them using a range of services. Their answers have been reduced to a balanced popularity score; the closer the score is to +10 the more likely it is to be used, 0 is neutral, and the closer to -10 the less likely it is to be used. This popularity score simply shows a median-like value and does not indicate the spread of responses; although the “not sure” responses were not included in the popularity calculation these do give an indication of the confidence we can have in the popularity score - the lower the %, the more reliable the popularity score:

<table>
<thead>
<tr>
<th>Mode of delivery</th>
<th>Balanced popularity (+10 to -10)</th>
<th>Not sure responses (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email/Telephone support</td>
<td>3.6</td>
<td>8%</td>
</tr>
<tr>
<td>Helpdesk style face-to-face support</td>
<td>1.8</td>
<td>8%</td>
</tr>
<tr>
<td>One-to-one training</td>
<td>-0.4</td>
<td>5%</td>
</tr>
<tr>
<td>1-hour group training</td>
<td>1.2</td>
<td>6%</td>
</tr>
<tr>
<td>Structured training course</td>
<td>-4.0</td>
<td>5%</td>
</tr>
<tr>
<td>Half-day training course</td>
<td>-1.9</td>
<td>4%</td>
</tr>
<tr>
<td>One-day training course</td>
<td>-5.1</td>
<td>5%</td>
</tr>
<tr>
<td>Discussion with a data preservation specialist at start of research</td>
<td>2.7</td>
<td>7%</td>
</tr>
<tr>
<td>Discussion with a data preservation specialist at end of research</td>
<td>3.0</td>
<td>6%</td>
</tr>
<tr>
<td>Web-based training course</td>
<td>-1.2</td>
<td>8%</td>
</tr>
<tr>
<td>Web-based FAQs</td>
<td>4.8</td>
<td>5%</td>
</tr>
<tr>
<td>Moodle-based training resources</td>
<td>0.0</td>
<td>11%</td>
</tr>
<tr>
<td>Include in Information Literacy course (PhD respondents only)</td>
<td>-1.8</td>
<td>6%</td>
</tr>
</tbody>
</table>

9.3. We can see that there is an aversion to most kinds of formal training including, interestingly, Web-based training; the only exception to this is a 1-hour training course, which has a slight positive score. The most popular modes of delivery are informal and ad hoc training targeted at the researcher and their individual case.
9.4. Rather surprisingly, the most popular of all the delivery mechanisms is a Web-based FAQ approach. This is followed by email/telephone support, then discussions with a specialist.

10. OTHER INFORMATION MANAGEMENT SERVICES

10.1. A meeting was held with LSE Records Management staff where the objectives of DICE were outlined. The outcome was that DICE would keep them informed of progress but that no active involvement was needed and no RM factors need to be included in the training materials.

10.2. An email exchange took place with the Research Office, who could see no need for their involvement in the project.
11. CONCLUSIONS AND RECOMMENDATIONS

11.1. This survey has shown that most LSE researchers have little awareness of the need for preservation of their research data, let alone the techniques and facilities to do so. This is no surprise as it is only recently that the desire to preserve research data, as opposed to publications, has arisen in most of LSE’s disciplines.

11.2. The drive to preserve data has come principally from funding bodies such as the Research Councils, but this survey shows that only a small number of LSE researchers are funded from these sources. It is unlikely that the diverse governments, charities, endowments and even LSE scholarships, which provide a great deal of our researchers’ funding, will introduce conditions that require them to preserve their data or make it publicly available in the near future, so any drive to raise awareness will need to be based on data preservation being for the common good or to the researcher’s own advantage. Nevertheless, a link between the Research Councils’ requirement for data planning, of which data preservation is one part, would be sensible to get maximum value from the training materials and to future-proof the material as much as possible.

11.3. Although many researchers plan to preserve their data beyond the life of the research project, most do not expect to make it publicly available but rather keep it restricted to some degree. In some cases this will be the researcher’s “default setting” and it should be possible to encourage a greater openness particularly with the PhD students, though it will take years for such a culture change to spread widely. In other cases the researcher will have valid justification for restricting the data, such as on ethical or legal grounds, in which case preservation without public availability will be needed. Such preservation will almost certainly be undertaken by the researchers themselves without reliance on institutional or national repositories. This points us in the direction of personal archiving1 as a potential source of techniques that we could harness for DICE, obviously adapting them for use in academic research.

11.4. Few researchers document their data to aid navigation or understanding of that data by themselves or others in the future. This is despite a number of researchers reporting problems when using other peoples’ undocumented data. A significant habit that needs to be encouraged is appropriate documentation, and material on this must be included as part of DICE’s output. Since there are some researchers who do document their data, particularly in the 6-9-year experience group, it may be possible to develop some case study material.

11.5. Researchers are using and creating diverse data types during their research. Text-based material, including emails, forms the bulk, but tabular material, charts and graphs are very popular, and computer code & scripts are also significant. These media will need to be prioritised, perhaps by means of case studies, in the DICE project outputs.

11.6. Some researchers are using Cloud-based services to develop, maintain and backup their data. DICE should include material to support these services, including advice on the risks involved.

11.7. Most, though not all, researchers are organising and backing up their files to some extent. Since both practices have a bearing on data preservation, advice on these topics will be included in the DICE outputs.

---

1 Personal archiving has developed largely in response to the increased interest in family history and the recognition of the potential value of “that stuff in the attic”. Some archives, libraries and museums run courses for the general public in this topic.
11.8. There is no method of training or support that suits everyone: we will need to aim for diversity of delivery using the most popular methods identified in the survey. These are:

11.8.1. Web-based FAQs
11.8.2. Email & telephone support
11.8.3. Discussion with a data preservation specialist
11.8.4. 1-hour group training session

11.9. Some support materials would also be useful in Moodle; these could be linked to the 1-hour training course but able to function separately too. Although DICE is specifically charged with producing materials for the MY592 Information Literacy course, and will do so, the evidence is that this will not be the most effective method of supporting data preservation. However, where possible the training materials will be capable of being re-packaged so that, for example, Moodle material can support the 1-hour training course and can be re-packaged for MY592. It is also clear that someone will need to maintain the list of FAQs as well as provide email support and specialist discussion. Learning material will be needed that should be able to facilitate the training of this person/team by providing more depth and detail than the training aimed at researchers. Moodle modules would be ideal for this, and could also be made available to researchers for personal study.

11.10. No strong synergies were found between research data preservation and administrative data preservation (Records Management). The training materials produced by the project will be freely available to re-purpose for Records Management, but they will not be specifically tailored for this.