The Source to Output Repository Project: Social Science

Guy Burton
Information Services
LSE Library

August 2006

Table of Contents

3 EXECUTIVE SUMMARY

4 I. IDENTITIES
- *4 1. Questionnaire Phase: March-April 2006*
- *6 2. Interview Phase: May-June 2006*

8 II. PROJECT AIMS
- *8 1. Social science respondents' attitudes towards source-to-output repositories*
- *9 2. Perceived value of source-to-output repositories by respondents' roles*
- *11 3. Perceived missing functionality in source and output repositories*
  - *11 3.1 Questionnaire respondents' attitudes towards functionality in source repositories*
  - *11 3.2 Questionnaire respondents' attitudes towards functionality in output repositories*
  - *12 3.3 Interviewees' attitudes towards functionality in source and output repositories*

13 III. SOURCE DATA
- *13 1. Producing data*
- *15 2. Accessing data*

17 IV. SOURCE REPOSITORIES
- *17 1. Respondents' choice and frequency of depositing data in repositories*
- *20 2. General perceptions regarding source-to-output links by repository usage*
- *21 3. General perceptions regarding output-to-source links by repository usage*
- *23 4. Relationships between kinds and formats of data produced by respondents and frequency of submitting data to repositories*

25 V. METADATA
- 25 1. Types, stages and responsibility of metadata assignation
- 27 2. Relationships between metadata assignation and frequency of depositing data

28 VI. DATA ACCESS AND SHARING
- *28 1. The accessibility and sharing of primary research data*
- *33 2. Measures to control access to data by frequency of depositing data in repositories*

35 VII. OUTPUT REPOSITORIES
- *35 1. Respondents' use of output repositories*
- *39 2. Relationships between respondents' use of output repositories and depositing in source repositories*
- *43 3. Relationships between respondents' use of output repositories and their search techniques*

44 VIII. SUPPORT
- *44 1. Respondents' perceptions of guidance and support*
- *45 2. Respondents' perceptions of assistance for output repository use and assistance by librarian/knowledge management support*
- *46 3. Relationships between guidance/support and metadata assignation*

48 IX. CONCLUSIONS
- *48 1. Current ways of working*
- *48 2. Producing and consuming at the same time*
- *49 3. Same community, different needs*

50 X. FIVE SCENARIOS

Executive Summary

- There was widespread support across the social science research community regarding the aims of the StORe Project
- Nearly half of social science respondents claimed that both source-to-output and out-put-to-source repositories would offer a 'significant advantage to my work'; a third in both cases claimed it would be 'useful but not of major significance'
- Postgraduate students were generally more enthusiastic about source-to-output and output-to-source repositories than academic staff
- Academic staff suggested the StORe Project aims would benefit postgraduate students more than themselves; they anticipated StORe as a training tool
- Researchers reported limitations in access, completeness of data, navigability of data and concern with deposited research in repositories as their main concerns
- Researchers generally supported sharing of research in principle, but also admitted a sense of ownership of their own research data and work (producer-consumer distinctions)
- Researchers work in a relatively independent and 'organic' fashion; they use different search techniques and consult as widely as possible to find the information and material they want
- Researchers tend to produce and store their data in an idiosyncratic fashion, assigning their own metadata to their work
- Few researchers deposit data in source repositories
- Researchers that deposit data in source repositories do not show substantial differences to the wider research community in terms of metadata assignation or controlling access to their work
- Researchers appear to share their work in a personal manner, based on direct requests and based on merit

I. Identities

The data collection for the STORE project occurred in two phases: first, between March and April 2006 when responses were sought to an online questionnaire; second, from May to June, when interviews took place with researchers in either face-to-face situations or by telephone.

1. Questionnaire Phase: March-April 2006

In the questionnaire phase, potential social science respondents were sought by sending out over 500 invitations to heads and managers of university departments and selected non-academic research institutions across the UK. These contacts were identified through online search of email addresses and sent a standard message detailing the aims and objectives of the STORE project and an invitation both to participate in the questionnaire and request its wider dissemination among their colleagues.

The contacts list was put together, using data provided by the LSE Library's Information Services that provided a rough guide to the types of subject areas likely to participate in the questionnaire. Despite not having precise figures, the LSE Library has loosely maintained records between January 2002 and November 2005 that indicate which of the School's departments have requested particular datasets. In descending order these were: Economics (76), Geography and Environment (59), Social Policy (52), Government/politics (47), Economic History and Accounting and Finance (43), Statistics (30), Centre for Economic Performance (23), Sociology and European Institute (22), Methodology (15), International Relations (14), Media and Communication (13) and STICERD (economics-related) (10). The resulting list of contacts can be observed in Table 1.

Table 1: Contacts by Subject Area

Academic	Non-Academic	Subject Area	Total
X		Anthropology	27
X		Economics, Business	128
X		European Studies	9
X		Geography	84
X		Politics (and IR)	114
X		Sociology	64
X		Gender	6
	X	Party Political	2
	X	Think Tanks	23
	X	Business representatives	3
	X	Government	67
X		LSE PhD student group list (203 members)	1
X		LSE E-Newsletter Briefing (1400 staff)	1
			529

In March and April two emails were sent out to these contacts. The first included the invitation to participate in the questionnaire; the second was sent out in the last week of the online poll, reminding contacts of the project and the encouraging both them and their colleagues to take part before it closed. In both cases the 529 total was not achieved, with 503 and 501 contacts being successfully reached in March and April respectively.

From the total of 377 respondents that responded to the questionnaire across all disciplines, 61 (16.2%) were subsequently identified as belonging to the social sciences. The questionnaire allowed respondents to identify their subject area as they wished and is shown in Table 2 in descending order. While the rough data provided by the LSE suggested that economics was the most common background for individuals requesting datasets, the largest number of questionnaire respondents identified themselves as belonging to sociology, followed by economics, business and management, anthropology and politics. Both sociology and anthropology contacts were proportionally fewer than both the economics and politics ones in the email invitations sent out, but this does not provide an explanation as to why those associated with the former two subject areas were more inclined to respond than the latter.

Table 2: Social Science Respondents by Subject Area

Sociology	10
Economics	9
Business and Management	7
Anthropology	5
Politics	4
Development	3
Geography	3
History	3
Women's Studies	3
Psychology	2
Accounting and Finance	1
Criminology	1
Economic history	1
Education	1
Hospitality, Leisure & Sport	1
Information Studies	1
Organization Studies	1
Media and Communications	1
Music	1
Photography	1
Public Administration and Organizational Theory	1
Public health	1

The questionnaire asked respondents to identify themselves by institution. The LSE topped the number of respondents with 10, followed by Oxford University and UCL on 4 each. Universities constituted the overwhelming bulk of responses, with 56 against 6 from non-academic research institutions. The majority of responses were either from London-based (24) or English (29) institutions, although there was some limited participation from the other home countries: there were four responses from Scotland, two from Northern Ireland and one each from Wales and Eire.

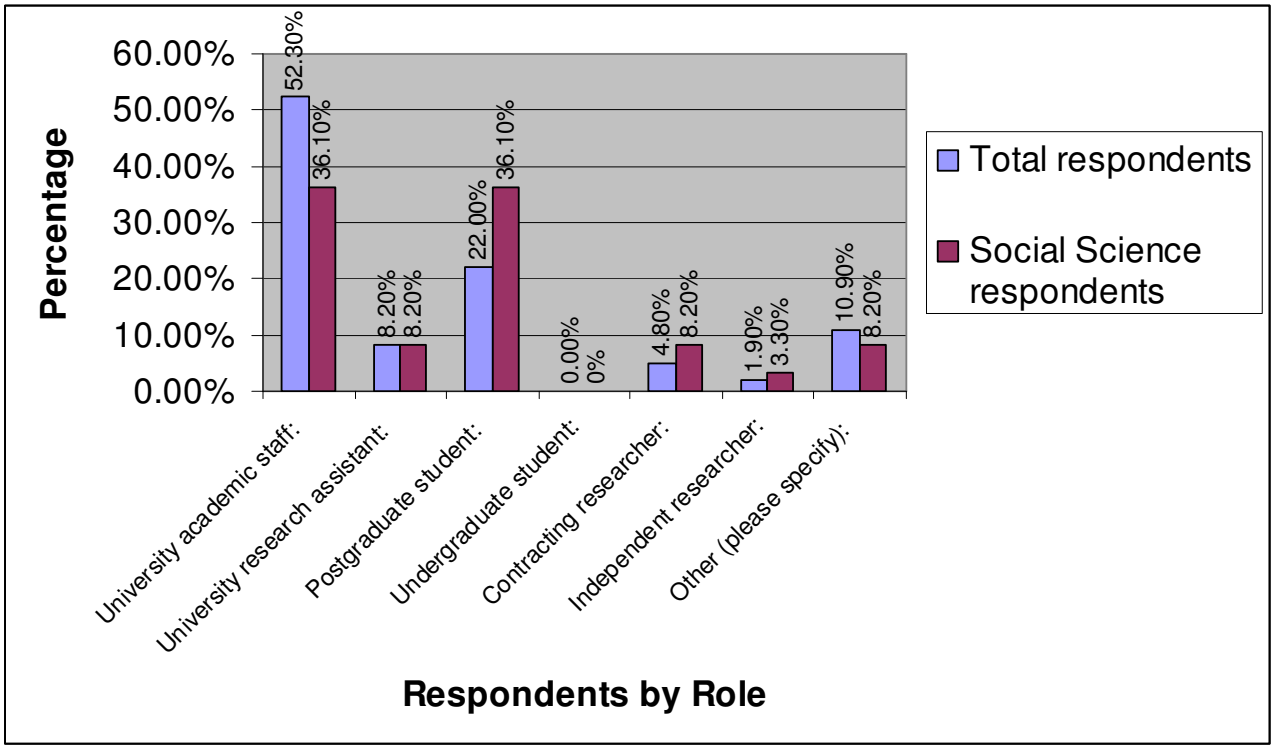
Table 3: Social Science Respondents by Institution

LSE	10
Oxford University	4
UCL	4
Exeter	3
Glamorgan	3
Essex	2
Goldsmiths	2
Heriot-Watt	2
Liverpool	2
Nottingham	2
ODI	2
Queen's University Belfast	2
York	2
Birkbeck	1
Birmingham	1
Bristol University	1
Cardiff Business School	1
Durham	1
Institute of Public Health, Ireland	1
Kent University, Canterbury	1
Leeds	1
Leeds Met	1
Loughborough	1
New Policy Institute	1
Plymouth	1
Royal College of Music	1

Royal Holloway	1
St Andrews	1
Stirling	1
Social Market Foundation	1
UEA	1
Wales Office, London	1
Warwick	1
Wolverhampton	1

Table 4 presents the respondents in the social sciences and the questionnaire as a whole, broken down by role. This shows that most social science respondents were either university academic staff or postgraduate students. Compared to the rest of the questionnaire respondents, the social sciences were underrepresented among university academic staff and overrepresented by postgraduate students, contracting and independent researchers.

Table 4: Respondents by Role



2. Interview Phase: May-June 2006

The second phase of the STORE project began upon closure of the questionnaire. Individuals were sought to participate in a series of interviews, either face-to-face or by telephone, between May and June 2006. There were two approaches taken to identify such people. First, questionnaire respondents were invited to say whether they were willing to take part in the interview phase and providing contact details. In this way, 20 potential interviewees were identified and contacted by email twice. Half of this number, 10, agreed to participate and were interviewed. Second, an email was sent to all department heads and managers at the LSE, asking them to circulate the request for interviewees. In this way, six individuals were identified and agreed to participate, all of whom had not taken part in the questionnaire phase. This provided an alternative perspective on the STORE project, by ensuring that a range of responses from those who had previous knowledge of the project and its questions and those who had none.

Table 5 gives a breakdown of the interviewee respondents who participated in the interview phase in chronological order. The figures loosely reflect the proportions in the questionnaire as whole. University academic staff make up half of the interviewees against 52.3% in the questionnaire; postgraduate students account for 25% of the interviewees (22% in the questionnaire); 12.5% of

interviewees were university research assistants (8.2% in the questionnaire); one contracting researcher and one 'other' each account for 6.3% of the interviewee sample against 4.8% and 10.9% in the questionnaire respectively.

All but one of the interviews involved those involved in higher education; the one not engaged in university-related work represented a research institution. Geographically, the LSE provided seven of the participants, the interviews included other institutions in England and three from Scotland (2) and Northern Ireland (1). As a result, this wide variety of interviewee experiences ensured considerable variation, including geographic, type of institution and role.

While anonymous, each interviewee is given an identity in the form of a letter which is used throughout the report.

Table 5: Interviewees by Institution, Role and Subject Area

Identity	Institution	Role	Subject Area	Questionnaire respondent	Type of interview	Interview Date
A	LSE	University Academic Staff	Anthropology	No	Face-to-Face	9 May
B	LSE	University Academic Staff	Economic History	No	Face-to-Face	9 May
C	St Andrews University	Postgraduate Student	Anthropology	Yes	Telephone	10 May
D	LSE	University Academic Staff	Social Psychology	No	Face-to-Face	11 May
E	Kent University	University Academic Staff	Sociology	Yes	Telephone	15 May
F	LSE	Contracting Researcher	Research Unit	No	Face-to-Face	16 May
G	LSE	University Research Assistant	Sociology	Yes	Face-to-Face	18 May
H	Birmingham University	Postgraduate Student	Public Administration	Yes	Telephone	22 May
I	Durham University	University Academic Staff	Sport Psychology	Yes	Telephone	23 May
J	Queen's University, Belfast	University Research Assistant	Social Survey	Yes	Telephone	23 May
K	Leeds University	Postgraduate Student	Development	Yes	Telephone	23 May
L	Heriot-Watt University	University Academic Staff	Business and Management	Yes	Telephone	24 May
M	ODI	Other	Development	Yes	Telephone	25 May
N	LSE	University Academic Staff	Social Policy	No	Face-to-Face	25 May
O	LSE	Postgraduate Student	Media and Communications	No	Face-to-Face	25 May
P	Essex	University Academic Staff	Economics	Yes	Telephone	6 June

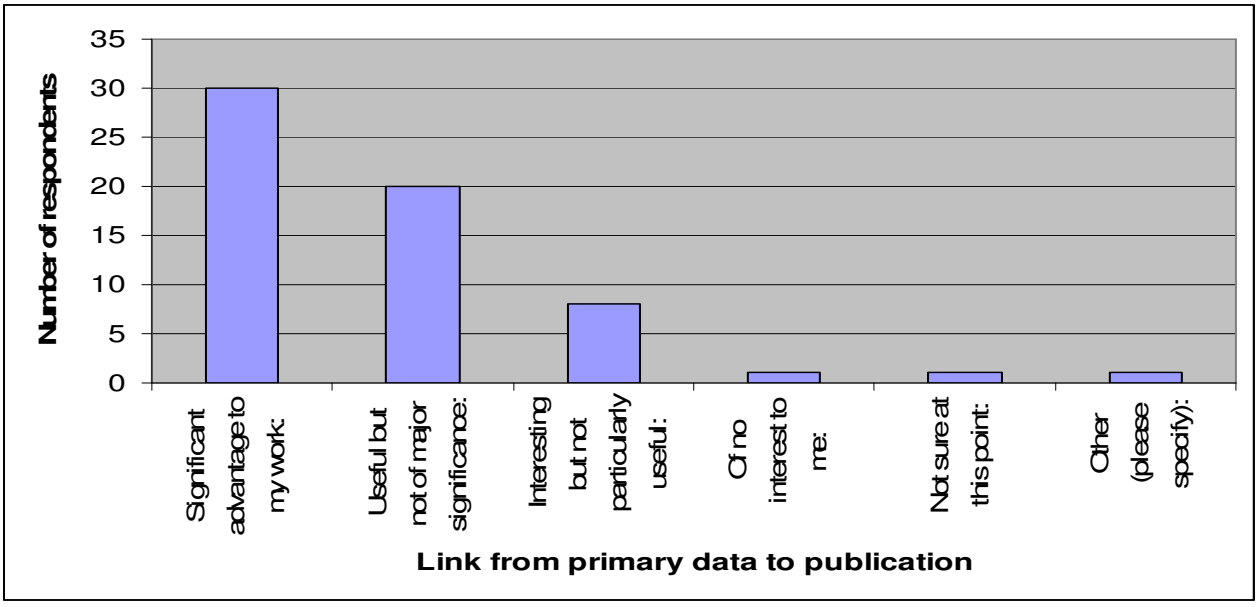
II. Project Aims

The concept of source-to-output and output-to-source repositories were treated generally favourably among social science respondents. Two questions were asked in the survey on this subject, the first concerning the ability to link from source repositories to publications developed from such data, the second regarding the ability to link directly from an online publication to the primary source data it was based on.

1. Social science respondents' attitudes towards source-to-output repositories

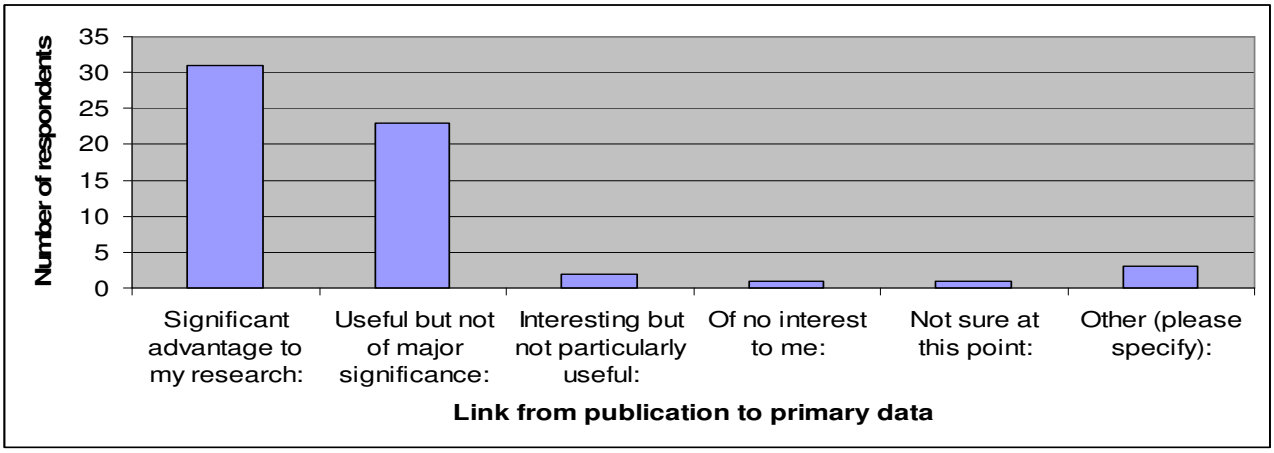
According to the results regarding repository links to publications, figure 1 shows that nearly half of all social science respondents – 30 of 61 – agreed with the statement that it would give a 'significant advantage to my work'. However, an almost equal number had a favourable but lukewarm response, with 28 stating that it would be 'useful but not of major significance' or 'interesting but not particularly useful'. Nevertheless, despite this response, only two it would be 'of no interest' or were 'not sure at this point'.

Figure 1: Respondents' attitudes towards source-to-output repositories



When asked if it would be advantageous to go directly from an online publication to the source data it was based on, social science respondents appeared particularly favourable, as presented in figure 2. Just over half – 31 – said that they thought it could give a 'significant advantage' to their work.

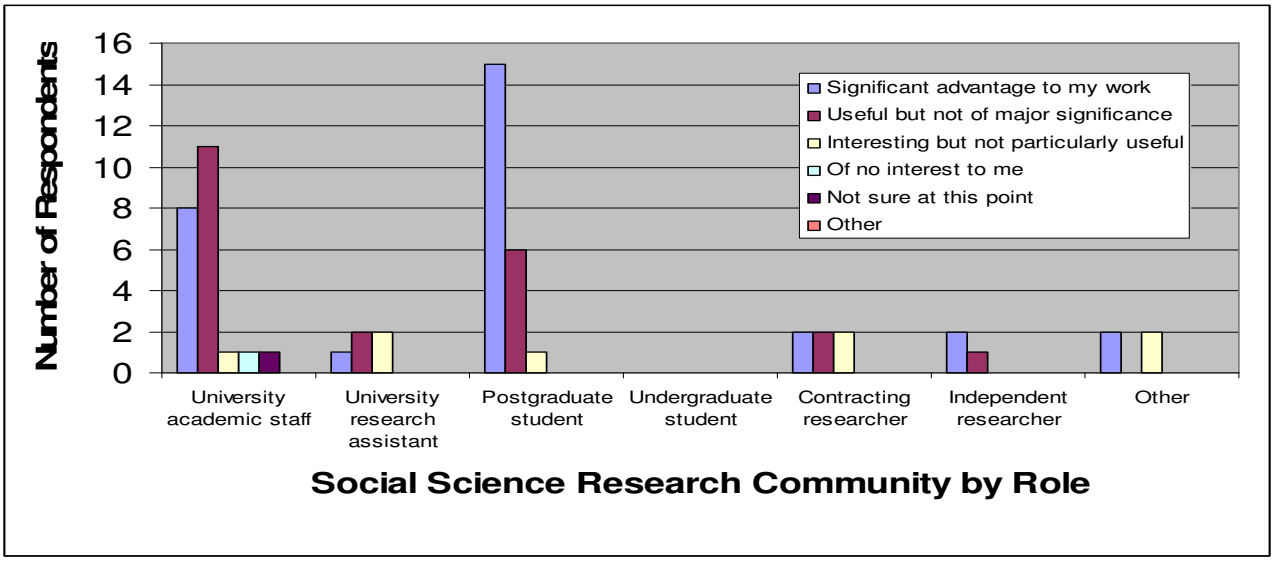
Figure 2: Respondents' attitudes towards output-to-source repositories



2. Perceived value of source-to-output repositories by respondents' roles

When broken down by role, the usefulness of a repository linking source data to publications varies slightly from the capacity to go from an online publication back to the data. With regard to links from repository data to publications, figure 3 shows that of the two main types of respondents, proportionally more postgraduate students (15 out of 22) agreed that such links would give a 'significant advantage to my work' compared to university academic staff (8 out of 22). University academic staff were more inclined to see such links as 'useful' or 'interesting' (12 out of 22) than postgraduate students (7 out of 22). Of the other roles represented in the survey, the numbers who responded were much smaller, but reflect a balance between the more enthusiastic and milder levels of support.

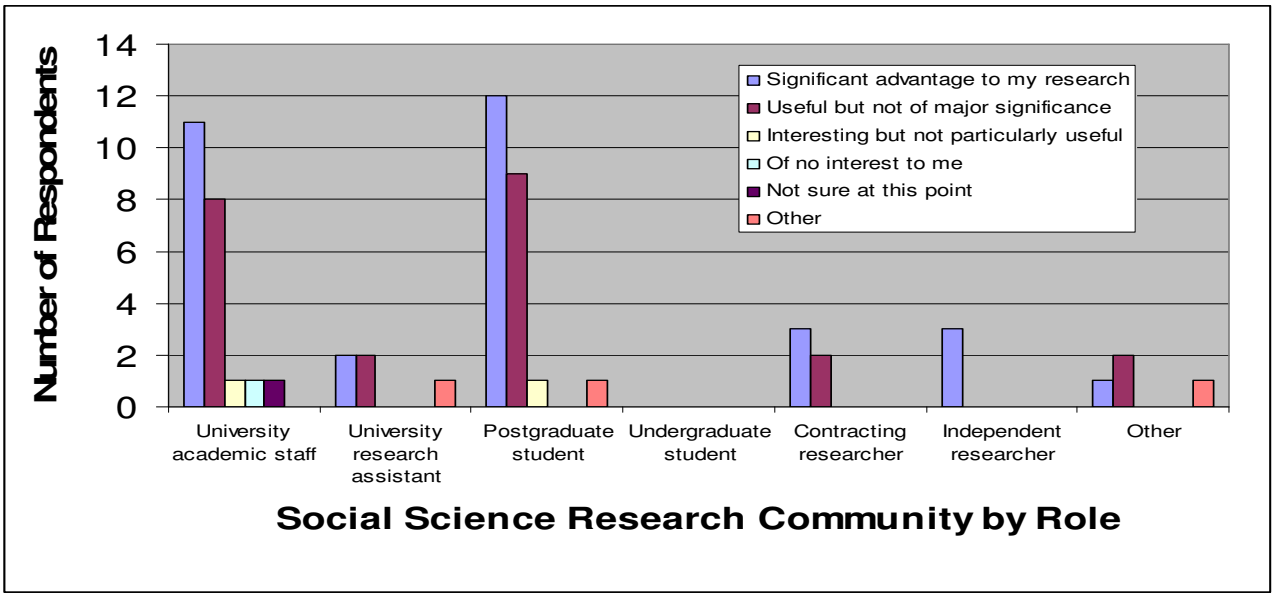
Figure 3: Link from Primary Data to Publication (Source-to-Output), by Role



NB More than 61 since more than one option could be checked in questionnaire

When the question was changed slightly (figure 4), to ask respondents about the use of linking from an online publication to the source data, a small change in perceived value occurred. Half of university academic staff (11) agreed with the notion that a direct link between an online publication and source data could provide a significant advantage while slightly less than half (9) thought it would be 'useful' or 'interesting'. Among postgraduate students, although slightly more than half (12) felt it would give a 'significant advantage', there was a higher proportion that were more lukewarm in considering the use of such links in their own research (10) compared to the idea in more general terms of being 'useful' or 'interesting' in Table 3 (7). Among other researcher roles, the numbers indicated a positive stance despite the smaller numbers.

Figure 4: Link from Publication to Primary Data, by Role



NB More than 61 since more than one option could be checked in questionnaire

When investigated in more detail through the in-depth interviews, the value of source-to-output repository links becomes more nuanced. Although there was a generally favourable opinion across most research roles represented in the survey, university academic staff tended to see it as more useful for others than themselves. More precisely they saw it as of particular benefit to postgraduate students for a range of reasons. B, an economic historian, and D, a social psychologist, saw such a facility as a means of methodology training for students, including in statistics and discourse analysis. E, a sociologist, concurred that such links would be more beneficial to her students than herself, while P, an economist, saw the value in providing students with data that they could use either to replicate previous work or establish what data already existed prior to undertaking their own research. Only one university academic staff member did not see such a facility as benefiting postgraduate students over her peers; she saw established researchers as those who would gain the most advantage, by providing the means to undertake secondary analysis and meta-analysis.

Among interviewed postgraduate students, the degree of support for source-to-output repositories was less clear than that in both the questionnaire or by university academic staff interviewees. For example, H, who studied public administration, was generally supportive of such a facility, but added a number of reservations. This included the prospect of group-think and slower research output as a result of data verification. O, a doctoral student in media and communications, also felt that such links could be helpful, but for quantitative forms of data rather than qualitative. By contrast C, an anthropology student, and G, a research assistant and sociology student, were more sceptical and distrustful of such links; C feared its use by non-academic interests, including commercial and government, while G claimed that such data would lack context and could result in scholarly attention being focused on his forms of data collection and methods rather than the findings he might make.

Of the other roles represented in the interviews, F, a contracting researcher, claimed that while such a facility would be of little use to her current research programme, she believed it could have been useful

to her doctoral studies and possibly for her future research. M, a development researcher at the ODI, thought it 'useful but not significant' in the questionnaire, claiming that he only read 2-3 articles a year where he might be interested in accessing the data concerned. Even then there would be constraints on pursuing such data: first, because of the number and size of variables that might exist in a database; second, the time it would take to go through the data.

3. Perceived missing functionality in source and output repositories

Social science respondents and interviewees were invited to comment on the limitations of source and output repositories as they saw them. In the questionnaire the majority did not venture an opinion and not all interviewees offered new lines of enquiry. However, a number of observations were made and have been split between questionnaire respondents and interviewees since questionnaire respondents were asked to offer individual comments on source and output repositories while interviewees were asked to make general statements and questions on both simultaneously.

3.1 Questionnaire respondents' attitudes towards functionality in source repositories

Four general themes were apparent: a lack of complete data, difficulty in navigating the data, problems in making sense of the data and accessing the data.

A. Incomplete data

- 'Different things are missing from different places'
- 'not always up to date'
- 'the lack of a meta system enable searching of multiple repositories simultaneously'

B. Difficulty in navigation

- 'lack of detailed cataloguing, keywords, full descriptions etc'
- 'Knowing what is there - often the data are hidden behind unhelpful titles. Often, it comes down to personal contact.'
- 'often very confusing in their output'
- 'severe shortcomings in metadata so no real understanding of how data were generated and how reliable it is'
- 'Also better searching: I find UKDA poor in this regard'

C. Incomprehensible data

- 'not known enough, not clear how one could submit data to them'
- 'Some that I have used are voluntarily maintained web pages (academic) and have no systematic content or metadata, so no instruction on content'

D. Accessibility of data

- 'Easy access to the data. There is usually too much red-tape and bureaucracy (e.g. with the Data Archive)'
- 'More online downloads - esp. for older datasets that one often needs to buy on CD'

3.2 Questionnaire respondents' attitudes towards functionality in output repositories

When social science respondents were invited to comment on the functionality of output repositories, three main themes emerged: the importance of linkages, a sense of incompleteness and lack of accessibility.

A. Links (source and output)

- 'Links to source data, which would be very useful.'
- 'Links to copies of the documents. Plus the fact you normally have to check several databases to make sure that all the references are covered.'
- 'Access from journal to raw data.'

B. Incompleteness

- 'not always up to date'

- ‘Timeliness’
- ‘networking with other repositories and also a more complete list of journals including less 'mainstream' sub-disciplinary ones rather than main subject journals which are not widely read’

C. Accessibility
- ‘lack of detailed cataloguing, keywords, full descriptions etc’
- ‘search with 'Author' should be easier. It is usually with 'title' that the search works best’
- ‘often very confusing in their output’
- ‘could be simpler in use’
- ‘access costs (financial)’

3.3 Interviewees’ attitudes towards functionality in source and output repositories

Interviewees were invited to make final comments or add their own contributions at the end of each interview. Of those that were willing to offer suggestions, the following themes were addressed:

A. Premature publication of results
Several interviewees remained concerned about the publication of data before sufficient use had been made of it by its producer (A, P). F, a contracting researcher, felt concerned about making it available in its ‘raw’ form.

B. Organization of data
E, a sociologist, asked how the source data would be organised. Would it be done in a format akin to a textbook? That is, ‘here’s the subject, here’s some related information related to it to look at’. Furthermore, how would the data be classified? For example, data on young women’s attitudes to higher education would be located under gender or education?

C. Accessibility
Interviewees reported a general acceptance of accreditation systems (B, F, G, H, I, J, M, N, O, P), even if N, a social policy lecturer, saw it was a ‘crashing bore’. B, an economic historian, argued that it should be possible for software to recognize a user so that typing in a password would become unnecessary. H, a postgraduate student in public administration, said it would be helpful if all licenses were on one system, perhaps through Athens. O, a postgraduate media student, suggested a ‘hierarchy of access’, which varying degrees of accreditation for staff, students and other researchers.

D. Use of data
I, a sports psychologist, spoke for several interviewees when she reiterated at the end of her interview her concern about how her data might be used. In particular she was concerned at non-academic usage and offered an example of a journalist taking her tentative conclusions to be fact.

E. Additional features and concerns
Both N and O pointed out that the development of a source-to-output repository might require additional demands to be met. O asked whether source-to-output repositories would be multilingual and multi-format (e.g. audio and video). N claimed that the qualitative source-to-output repositories would be complemented with computer-assisted data analysis given the new types of data formats that would arguably be stored in them. This would also give rise to new training needs for researchers in how to use them.

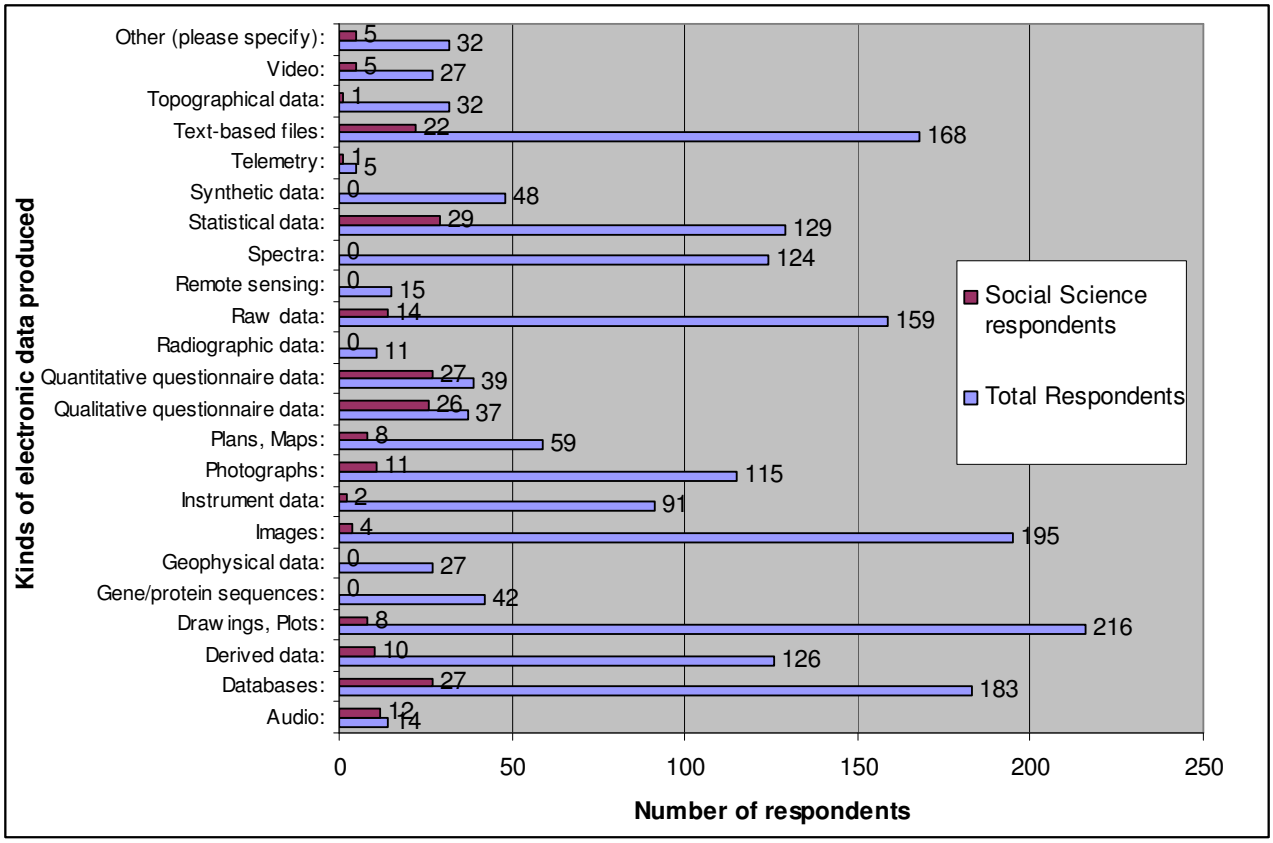
III. Source Data

The following section analyses the questionnaire and interview responses regarding the kinds of data produced by researchers, the formats they are kept in and the degree to which they access other researchers and make available their own. Social scientists were the research community most inclined to specify that their data was 'qualitative' or 'quantitative' compared to other researchers and made extensive use of formats such as Word, Excel and spreadsheet software. Like other researchers they would access others' research data to inform themselves of what is happening within their discipline, although it was notable that the extent to which they make use of others' data varies; a third claimed not to use it while nearly half said that they did so.

1. Producing data

Figure 5 shows that when asked what kinds of electronic data was produced by social science respondents, nearly half reported creating statistical data (29 instances), databases (27) or quantitative questionnaire data (27) and over a third claimed to produced qualitative questionnaire data (26). When compared against the responses made across the different disciplines as a whole, social science respondents predominated among those claiming to produce audio data (85.7% of all responses), qualitative questionnaire data (70.3%) and quantitative questionnaire data (69.2%).

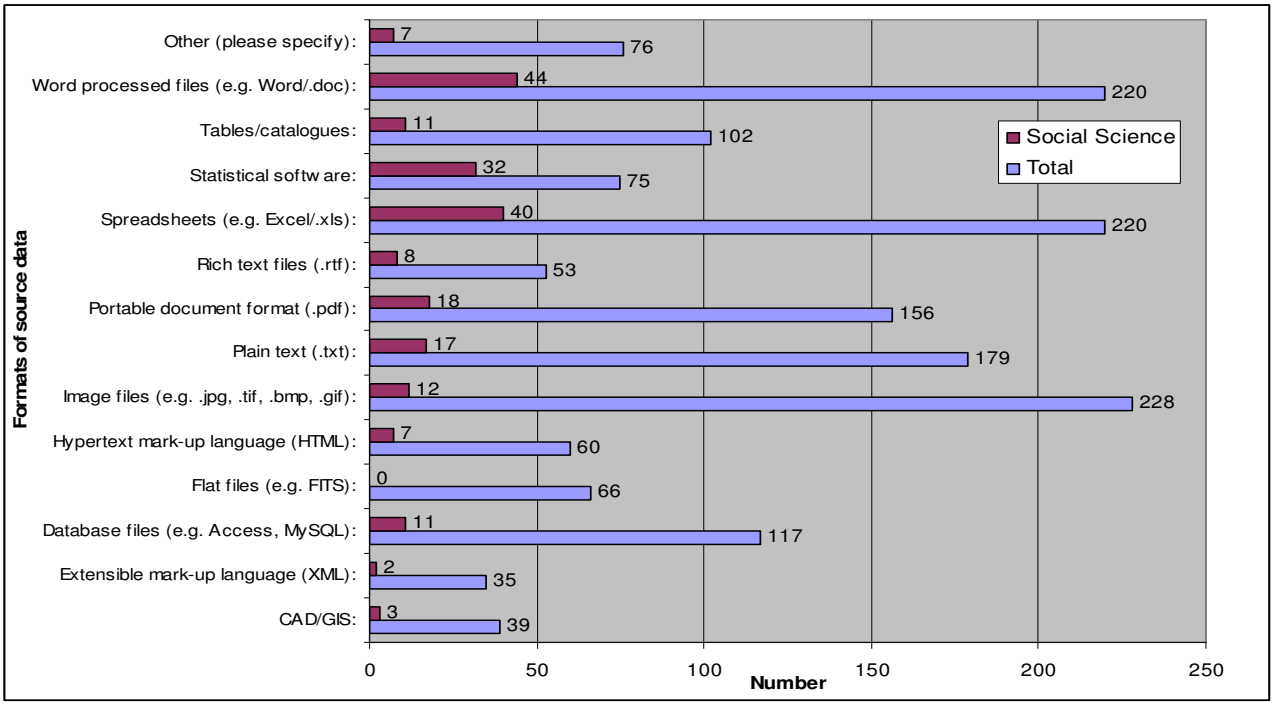
Figure 5: Kinds of electronic data produced by respondents



The most popularly cited format for source data used by social science respondents were Word processed files; two-thirds of them claimed to use the format (44 citations by the 61 respondents). A slightly fewer number claimed to use spreadsheets such as the Excel package (40) and half of social science respondents used statistical software (32) as shown in Figure 6. This was reflected in the more qualitative interviews as well, with Word documents being among the most commonly cited format used for text-based data, regardless of a researcher's role (i.e. university academic staff or postgraduate student) (A, D, F, G, H, L, N, P). Indeed, a researcher's role did not appear to be related to the types of format used, with not only Excel being cited (B, M, P), but SPSS as well (B, G, H, J, L, M). Other format types included mpg, digital video and audio, mp3 files (questionnaire) and saving emails and MSN conversations (C) and HTMP (O).

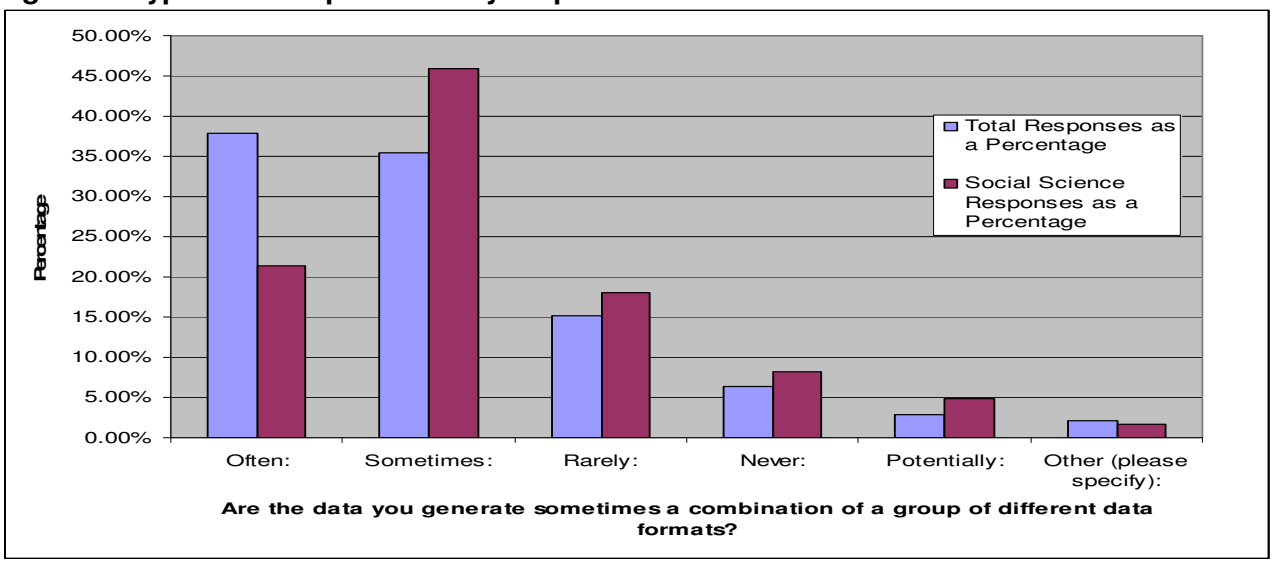
Interview work also highlighted the globalisation of converting data from one format to another. B, an economic historian, offered the example of a colleague who used Indian labour to transcribe data collected by digital camera at a cheaper cost than using a British-based research assistant.

Figure 6: Formats for source data used by respondents



Nearly half of social science respondents (45.9%) stated that 'sometimes' the data they generated was a combination of different data formats. This compares to just over a third of all respondents in the questionnaire (Figure 7). By contrast respondents of other disciplines appear more 'often' inclined to generate a combination of different data formats (37.9%) to social science respondents (21.3%). However, if the 'often' and 'sometimes' results are added together, there would appear to be little difference between social science and total respondents in terms of producing a variety of different data formats (67.2% and 73.4% respectively), indicating that a substantial majority of both types of respondents do use more than one type of data format.

Figure 7: Types of data produced by respondents



2. Accessing data

Respondents were asked why they might wish to access the research data from other research programmes. More than 80% of all respondents and nearly three-quarters of social science respondents felt that it was useful or necessary to access data for their own research (312 and 45 citations respectively). Slightly fewer respondents – just over half of all respondents and two-thirds of social science respondents – agreed with the statement that accessing others' data was needed to understand the broader context and orientation of one's own research (209 and 40 respectively) (Figure 8). Also notable was the finding that a quarter (24.8%) of all responses relating to the relationship between accessing data and making useful contacts were made by social science respondents. Finally, one free text response suggested that the reason for accessing others' data could be 'To see how the primary research data has been interpreted by the researcher who did the research or made the publication (trace the link between data and its interpretation)'.

Figure 8: Respondents' reasons for accessing research from other research programmes

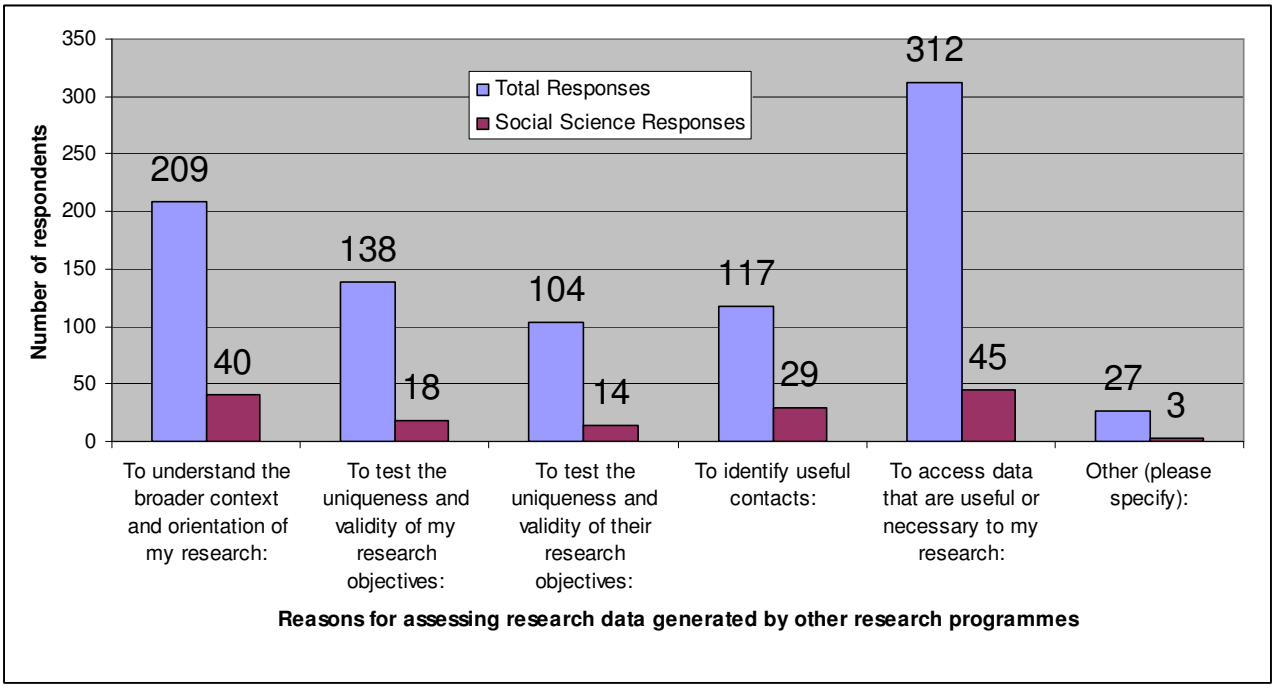
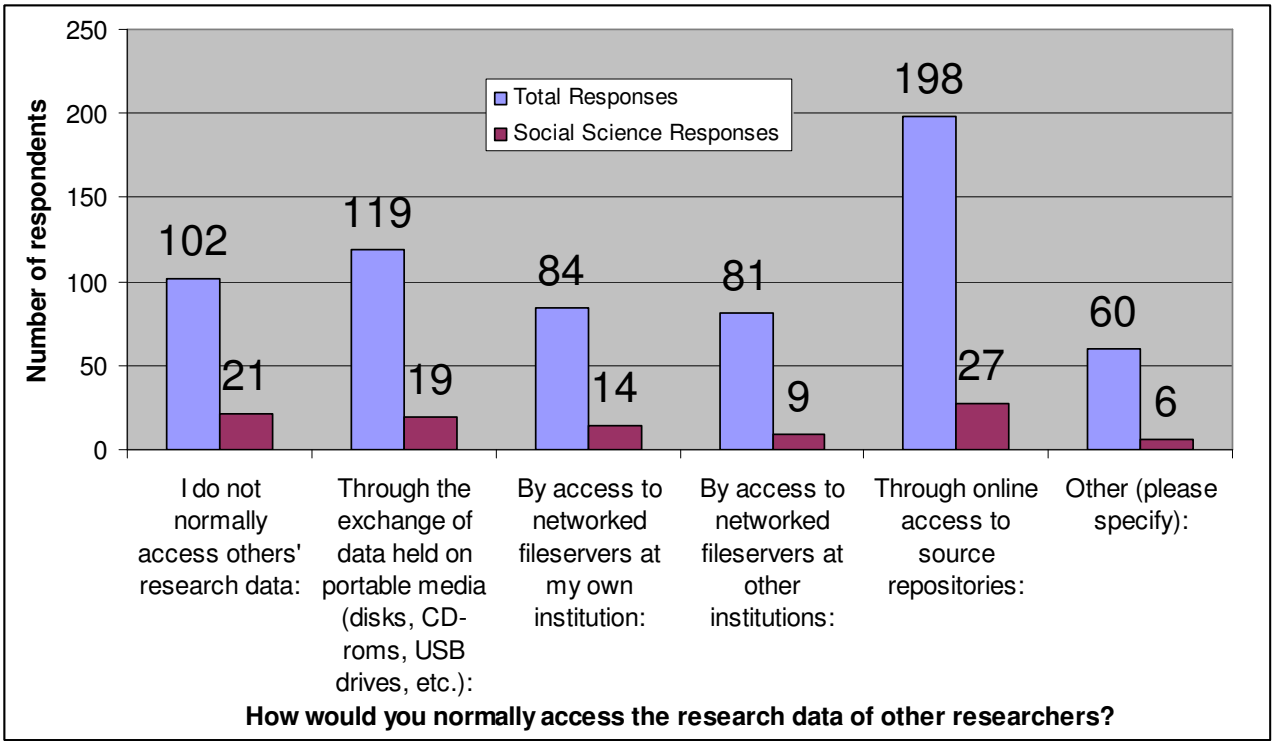


Figure 9 shows the ways in which respondents access other researchers' data. Half of all respondents and nearly half of social science respondents said that the most usual way in was through online access to source repositories (198 and 27 respectively). But proportionally more social science respondents compared to all respondents – one-third against a quarter – claimed that they both used portable media to exchange data (19 and 102 respectively) or that they did not use others' research (21 and 102). Indeed, put another way, social science respondents were responsible for a fifth of all references alleging non-use of others' data. Among the 'other' responses were references to 'word of mouth', 'by asking them for it', 'by contacting them directly, and using email to transfer data, or their websites' and 'by going back to the original sources'.

The relatively higher proportion of non-usage of others' data would appear to be contrary to the findings in figure 8 which indicate the reasons why respondents would wish to use such material. However, despite the various intentions given for social science respondents' reasons to access others' research, the interviews suggested that despite interviewees' general support with the principles of sharing data, there were various reasons for them either using it or not. Both B, an economic historian, and F, a contracting researcher, actively made use of others' datasets. Indeed much of B's work depended on the use of data collected by others. But F, who uses housing statistics available through the ODPM website, was especially concerned that the recent break-up of the department might have repercussions on the collection, storage and accessibility of such data in the future.

Other interviewees indicated that the use of others' research was not always evident. For example, an anthropologist, A, did not make much use of others' raw data, preferring to rely on her own findings. Much of E's work in sociology is based on interpreting others previous finished work, meaning that she neither produces or uses any source data. G, a research assistant and postgraduate sociology student, was wary about making 'strong statements' based on another researcher's data, given his lack of knowledge or awareness regarding the context in which it was generated. H, a postgraduate in public administration, said he had not pursued others' data since the researchers concerned had either moved away or had not responded to his requests. N, a social policy academic staff member, claimed that she had never approached anyone for access to their data since she felt they might be offended by this. A development researcher, M, articulated several reasons for why others might be unwilling to share their findings, including issues of ownership, its use and the potential complexity of ordinal, quantitative data through the many variables that may exist. Finally, L's work in business and management was geared towards generating both bespoke and new data, which meant doing something different from other researchers and teams.

Figure 9: Ways respondents access other researchers' research data



IV. Source Repositories

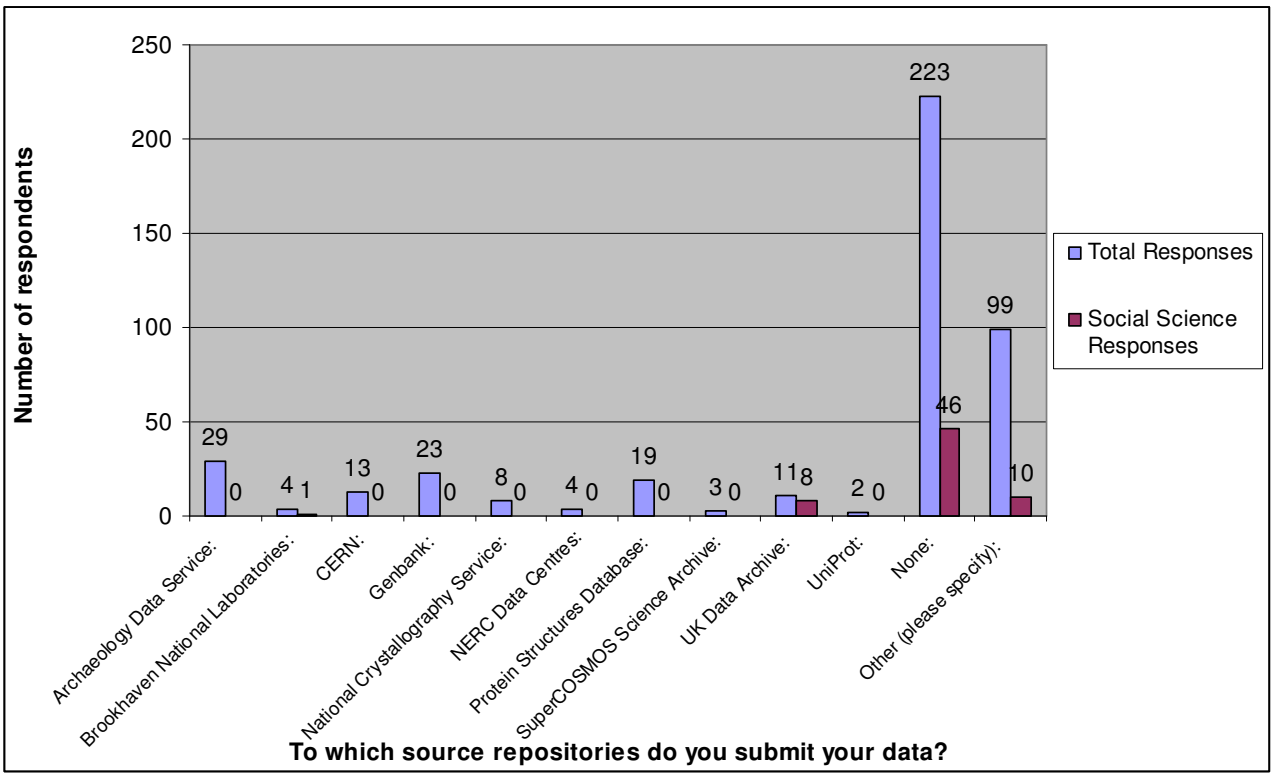
Both the questionnaire results and interviews suggested a relatively low level of data submission to source repositories by respondents. Analysis was run against these results, including the kind of data produced by the respondents and the format it was stored in. However, since the figures for those depositing data was so low the findings should arguably be treated with some caution.

1. Respondents' choice and frequency of depositing data in repositories

The vast majority of respondents to the questionnaire as a whole and within the social science sample indicated that they had never deposited data in a source repository (figure 10). However, among those social science respondents who had, the most cited repository was the UK Data Archives (8); among the 10 responses registered as 'other' included individual references to the EPSRC Environment Agency, ESDS (Economic and Social Data Service) International, the Global Entrepreneurship Monitor data at www.gemconsortium.org, Public Health Observatory and journal and project websites. In addition to the UK Data Archives, one postgraduate student concerned with education and trade claimed to have deposited data at Brookhaven National Laboratories.

The low level of data depositing by social science respondents was reflected in the qualitative interviews. Of the 16 interviewees, three had deposited data at the UK Data Archives (B, E and J), four had an awareness of the repository and had knowledge of it but found it did not meet their research agenda (F, K, N, P) while eight had either never used it or were unaware of its existence (A, C, D, G, H, I, L, M). This included a range of different kinds of researchers, including both university academic staff and postgraduate students.

Figure 10: Repositories to which respondents submit source data



Although figure 10 indicated that social science respondents had submitted data to the Brookhaven National Laboratories and the UK Data Archives, when asked to state how often they had deposited data, there was no response given for the level of frequency given by the postgraduate student for data submission to Brookhaven. Rather, the respondent claimed to 'frequently' submit data to the UK Data Archives (figure 11). However, despite the absence of detail regarding data submission to Brookhaven, the questionnaire findings revealed that a sociology postgraduate student claimed to

have submitted data to the CERN repository 'on several occasions' (Table 11) – the only social science respondent to do so.

A cross-tabulation between depositing data at the CERN repository and the kinds of source data produced indicated a wide range of formats used by the student, totalling 10 in number. This included audio material, databases, images, photographs, qualitative and quantitative questionnaire data, raw data, statistical data, text-based files and video material. A further comparison, based on the types of formats in which the source data are were held indicated a range of different types by the respondent, including image files, plain text, rich text files, spreadsheets, statistical software and word processed files.

Figure 11: Respondents submitting data to CERN repository

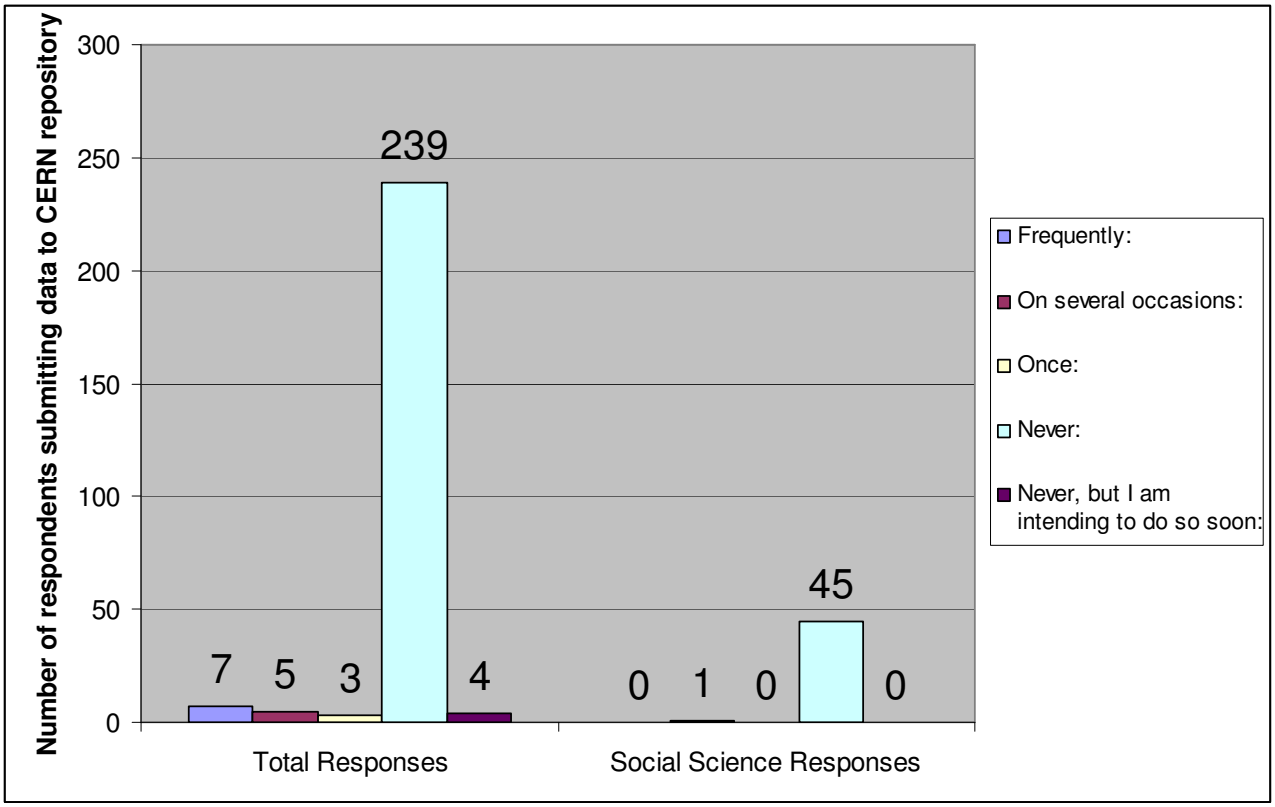


Figure 12 shows the frequency of data submission to the UK Data Archives. Of the eight responses claiming to have submitted data to the repository, one failed to provide an answer. Of the remaining seven, three university academic staff in economics and social policy claimed to have deposited data 'once'; a university academic staff member in politics and sociology and a postgraduate politics student said they had done so 'frequently'; and a sociology research fellow and the postgraduate student concerned with education and trade said they had done so 'on several occasions'. Of the five who said they had never deposited data but planned to do so, three were university academic staff and two postgraduate students; with regard to their fields of interest, two worked in sociology and one each in development, economics and public administration.

The greater preponderance of sociology and economics as the main fields of interests represented among the researchers who had submitted data to the UK Data Archives may arguably be explained by through the greater response rate of social science respondents in these two fields (see Identities section). However, the interviews also suggested a greater inclination to deposit among researchers in such fields: in addition to E, a sociologist, and J, a research assistant concerned with social surveys, an economic historian, B, who had not taken part in the questionnaire, also deposited data at the UK Data Archives. P, an economist, was a recipient of ESRC funding but was told that her results were too narrow and time-specific to be worthwhile depositing, while A had submitted samples in 2003 but had not heard back.

Figure 12: Frequency of respondents submitting data to UK Data Archive

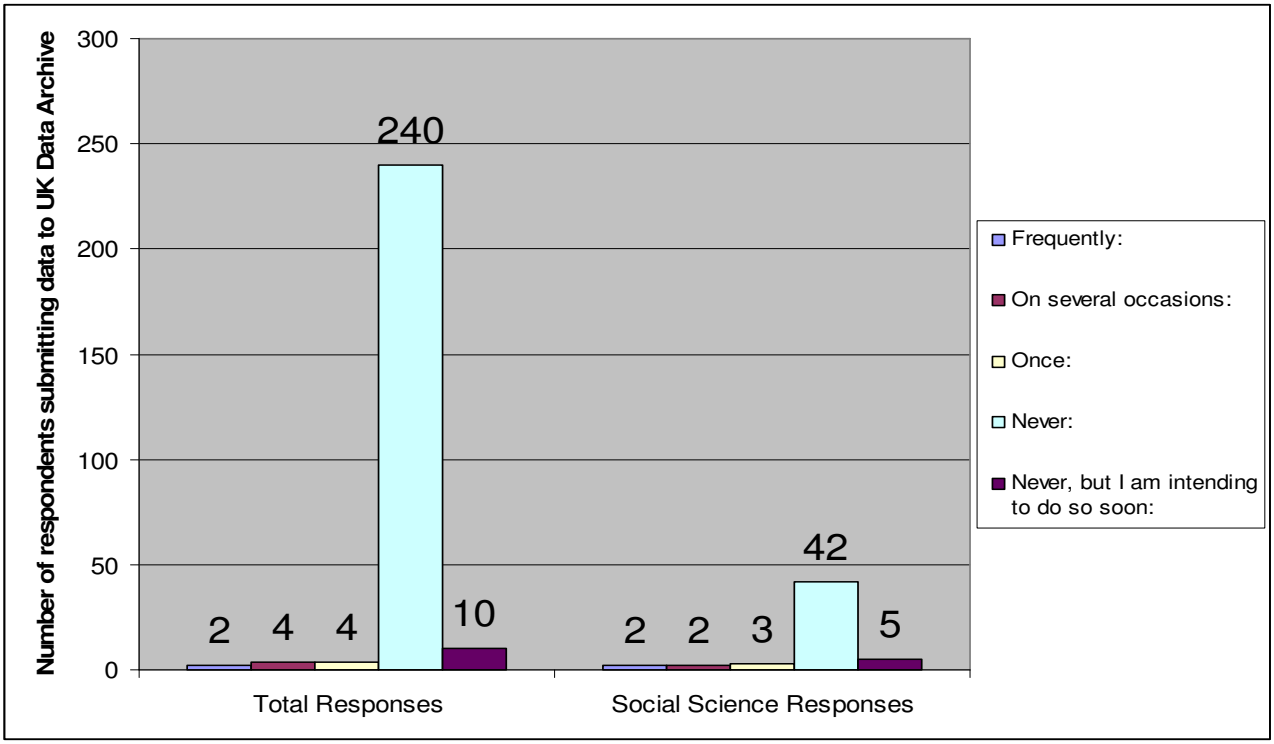
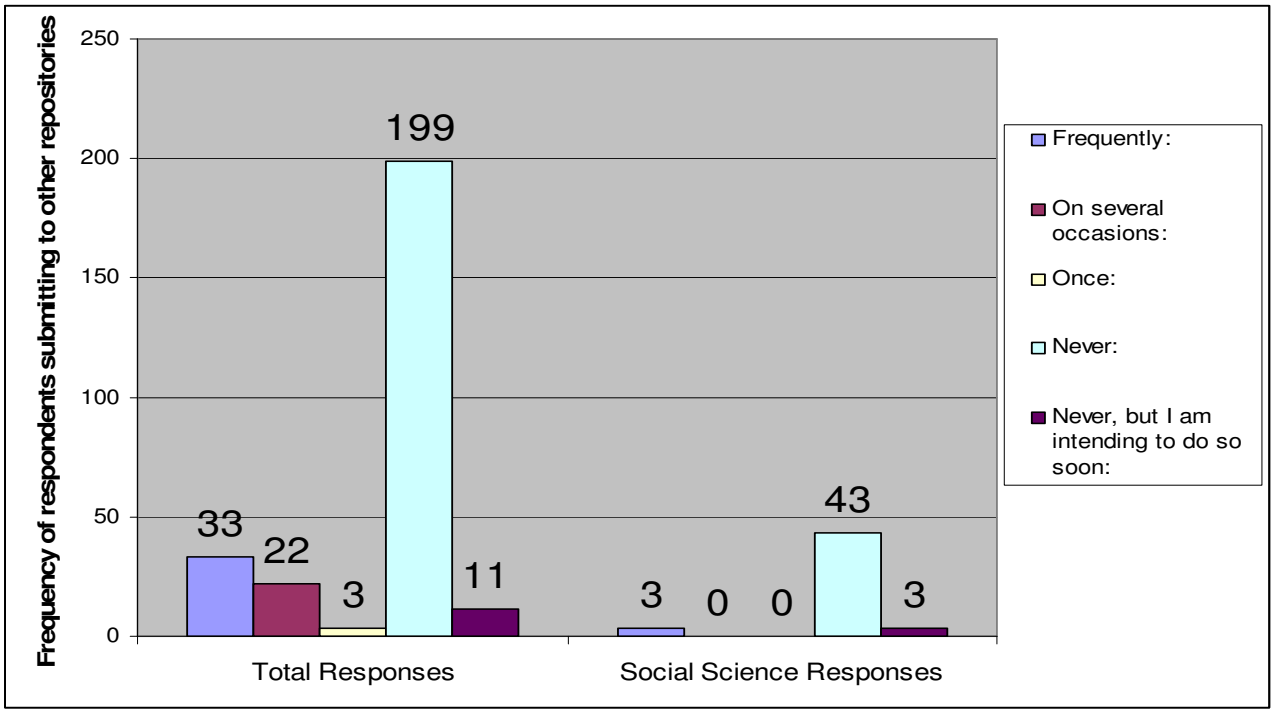


Figure 13 shows that of the three respondents who claimed to have deposited data in other repositories, two cited the UK Data Archives alongside 'our own project website' and 'open or unrestricted access web pages'. The third provided more specific detail, through the Global Entrepreneurship Monitor data at www.gemconsortium.org. No response regarding frequency was given from the respondent who claimed to have deposited data at the Economic and Social Data Service or the EPSRC Environment Agency while the Public Health Observatory respondent claimed that data had not yet been submitted, but was about to be done.

Figure 13: Frequency of respondents submitting data to other repositories



2. General perceptions regarding source-to-output links by repository usage

There does not appear to be a substantial difference in social science respondents' attitudes towards source to-output repositories, even when factoring in the extent to which they actually make use of repositories by depositing data in them. This is apparent in figures 14, 15 and 16 and 7 which provide a breakdown of the importance particular social science researchers attach to the possibility of linking from source data in a repository to publications derived from them. Only one response other than 'never' or 'no answer' (which may arguably be included among the 'never' responses) was made concerning the depositing of data in the CERN repository in figure 14.

Similarly, figure 15 shows that while there were more respondents that deposited data in the main social science repository listed in the questionnaire (UK Data Archives), as a proportion of the total these figures remained low. The few affirmative responses to depositing data at the UKDA (two 'frequently' doing so, two 'on several occasions' and three having done so 'once') concurred with the majority that had 'never' deposited data in the repository before, by generally having favourable opinions of source-to-output links. This was also reflected in figure 16, which examined the depositing of data in other repositories not listed in the questionnaire and the perceived value of source-to-output; of the three 'frequent' depositors, two saw a 'significant advantage' in the possibility of linking from source data to publications while one felt it was 'interesting but not particularly useful'.

Figure 14: Respondents attitudes to source-to-output repositories against frequency of depositing in the CERN repository

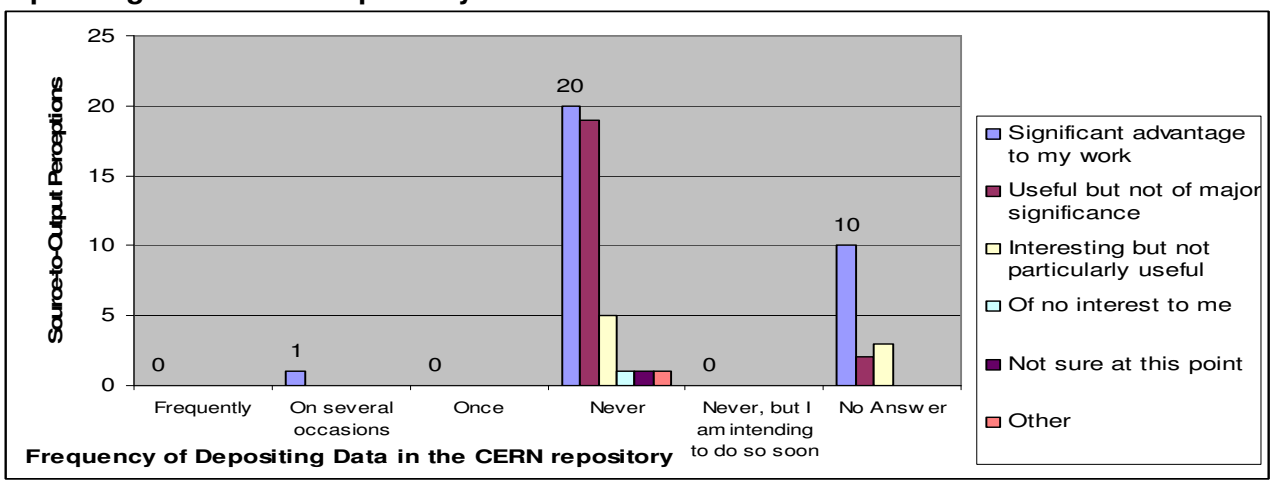


Figure 15: Respondents attitudes to source-to-output repository against frequency of depositing in the UK Data Archive

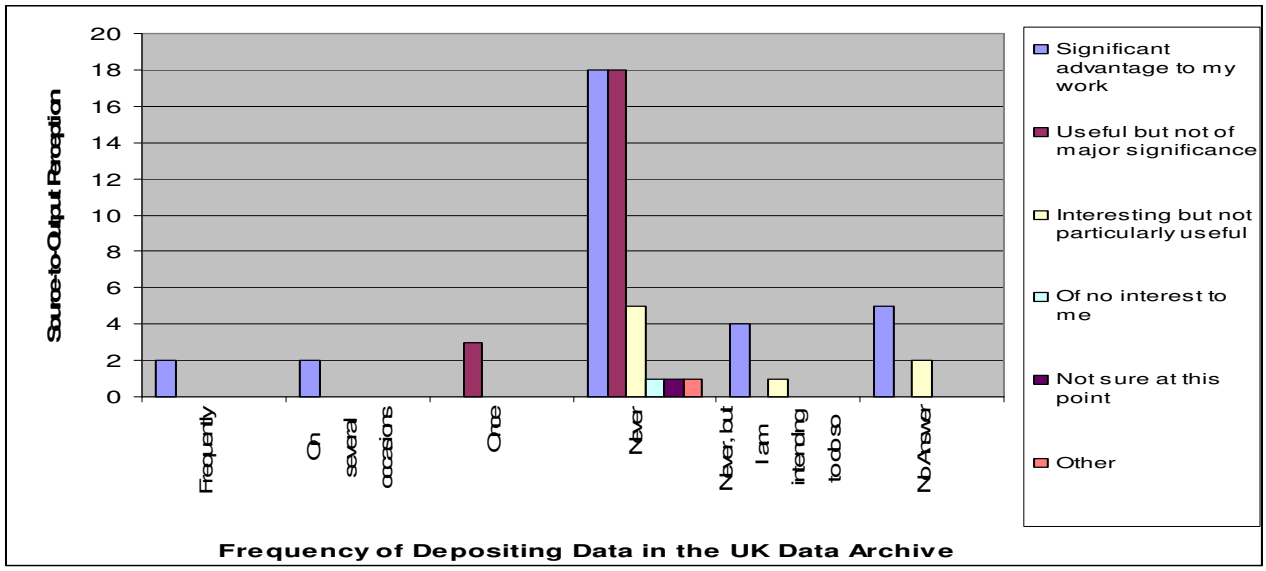
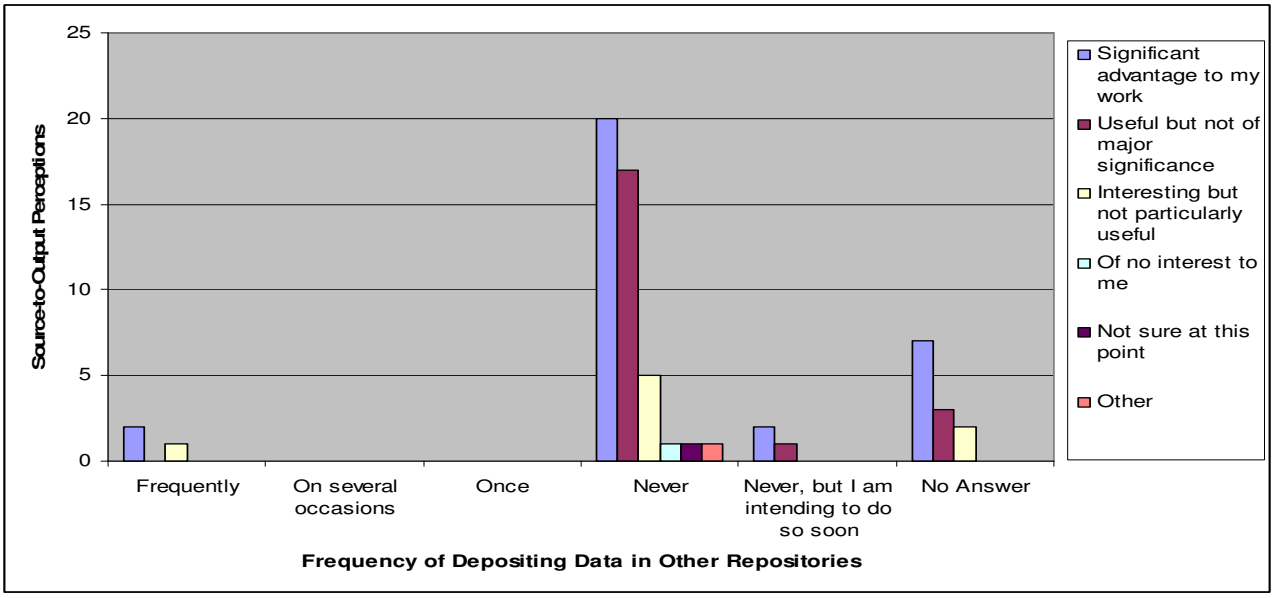


Figure 16: Respondents attitudes to source-to-output repository against frequency of depositing in other repositories



3. General perceptions regarding output-to-source links by repository usage

When the emphasis of the questions was changed, to consider the link between online publications and the source data they used, the lack of difference between depositors and non-depositors persisted. Figures 17, 18 and 19 show that the majority of responses indicated a virtually nil deposit rate by social science respondents despite the general approval that source-to-output links could be of 'significant advantage' or 'useful but not of major significance'. These figures were largely similar for the few respondents who had deposited data in the CERN, UKDA or other repositories as well; indeed, only one respondent who had 'frequently' deposited data in a repository not specified in the questionnaire thought that the facility would be 'interesting but not particularly useful'.

Figure 17: Respondents attitudes to output-to-source repository against frequency of depositing in CERN repository

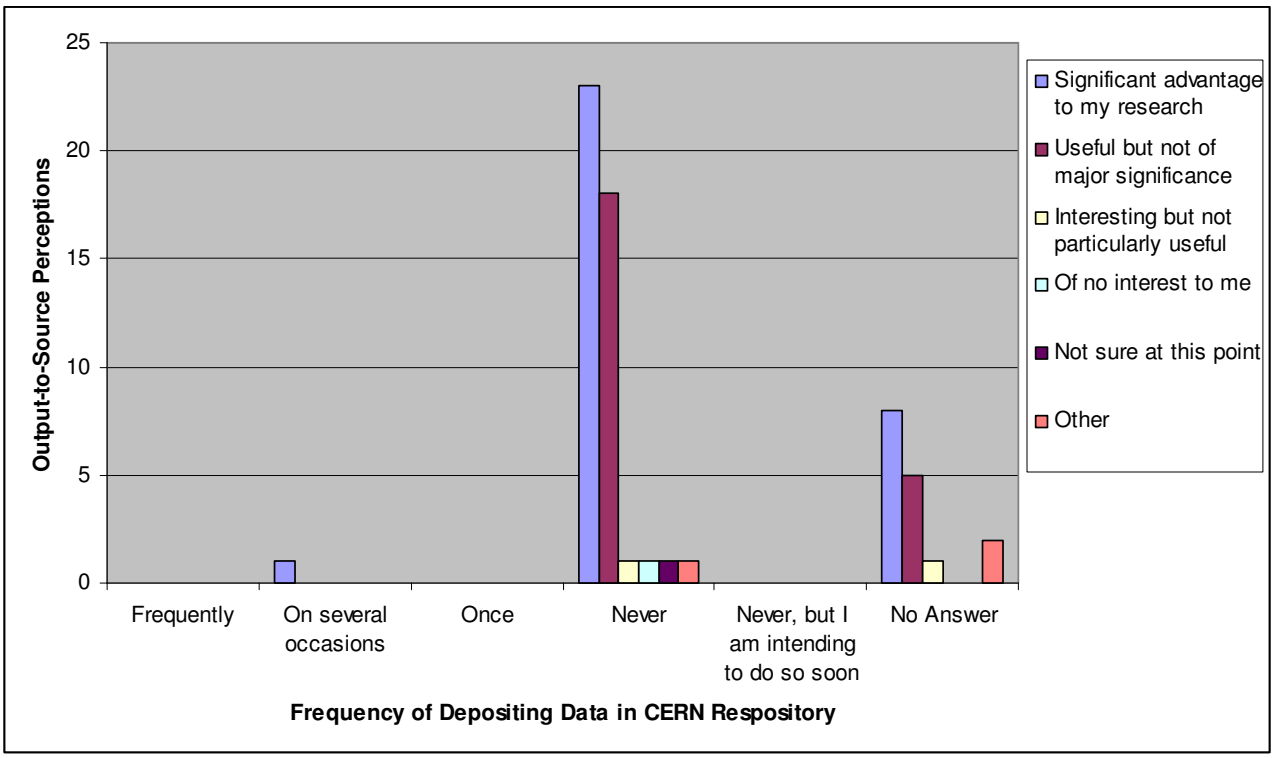


Figure 18: Respondents attitudes to output-to-source repository against frequency of depositing in UK Data Archive

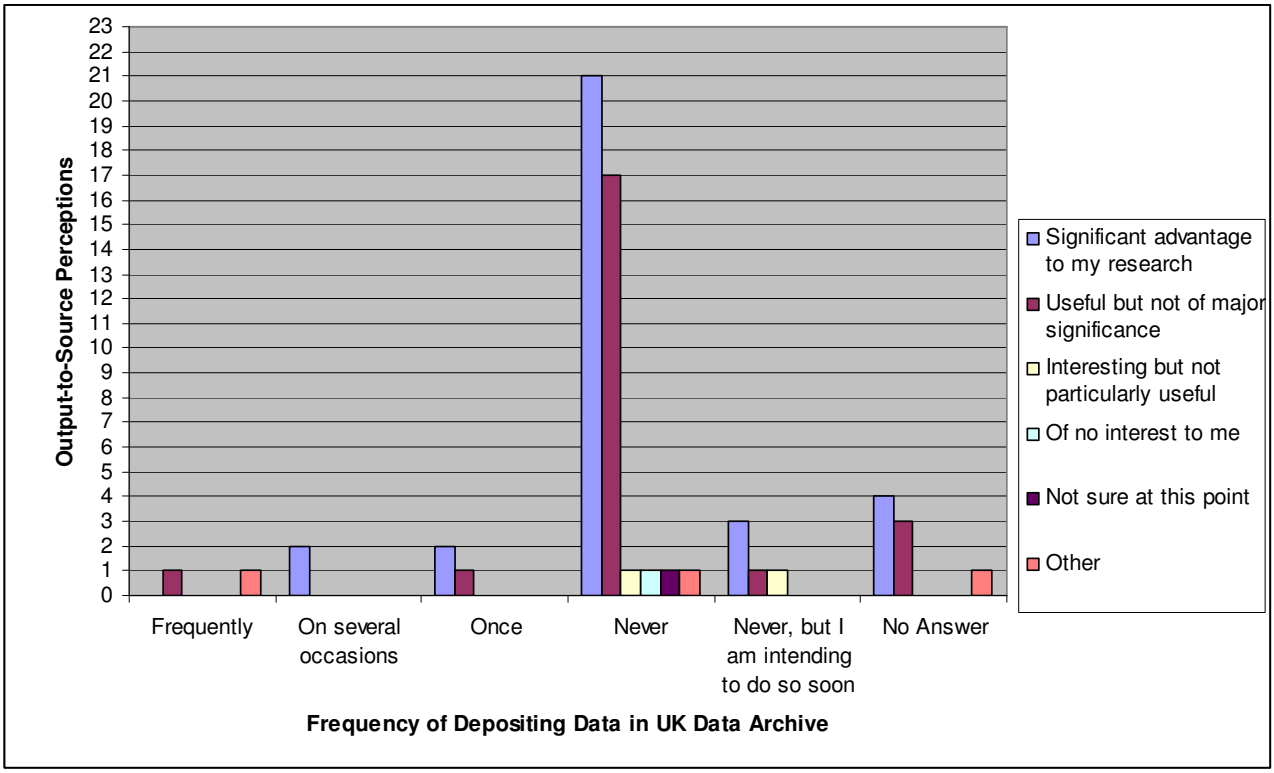
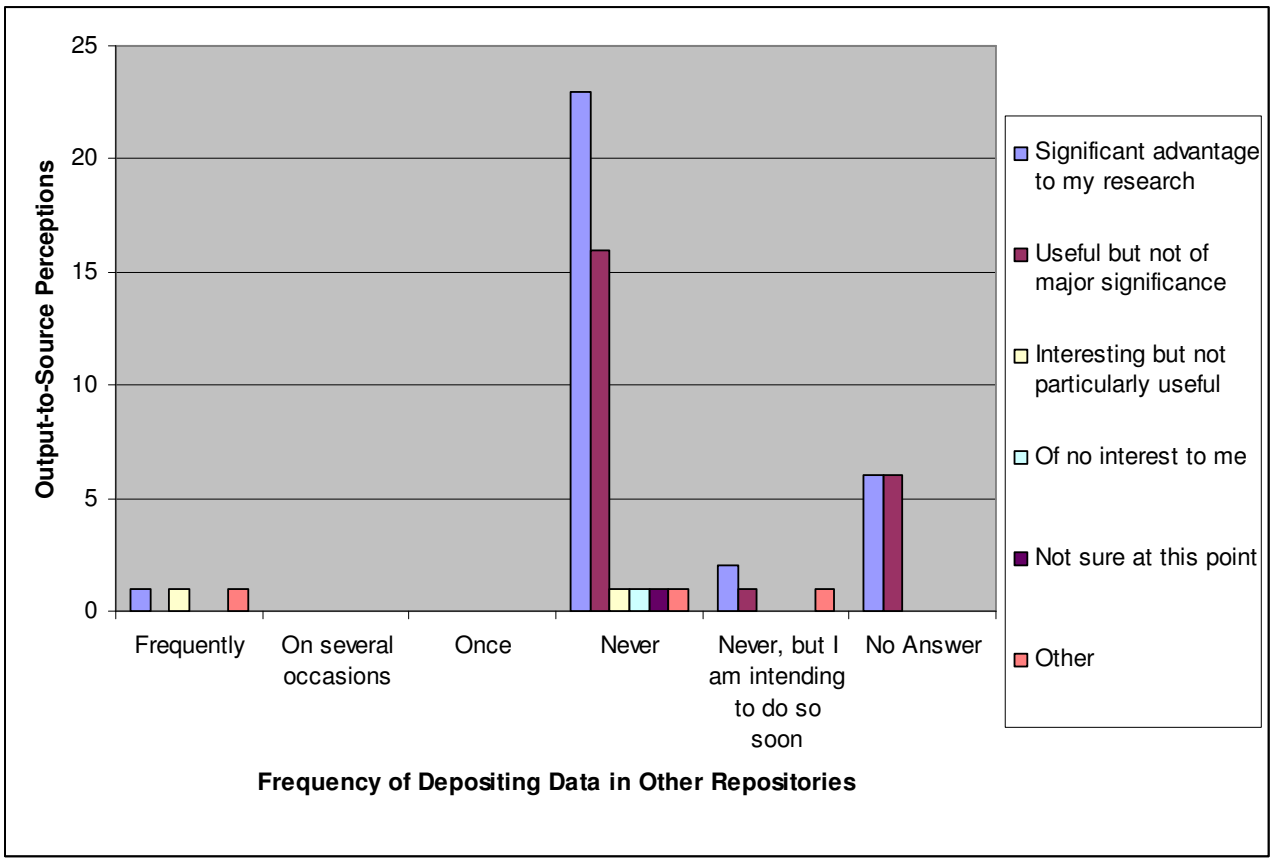


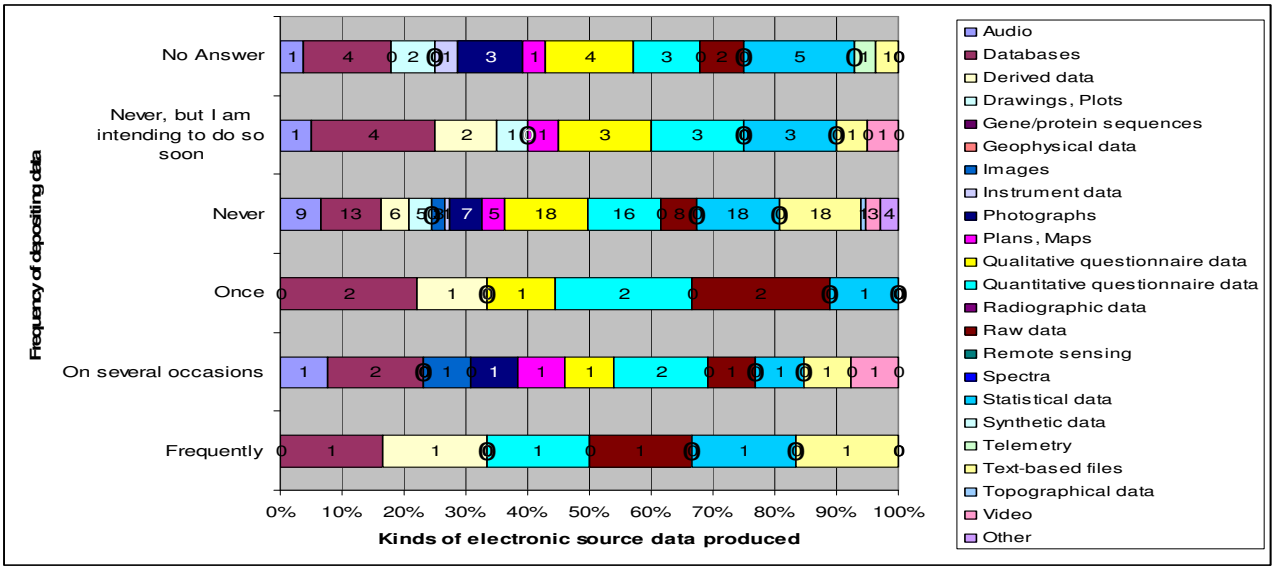
Figure 19: Respondents attitudes to output-to-source repository against frequency of depositing in other repositories



4. Relationships between kinds and formats of data produced by respondents and frequency of submitting data to repositories

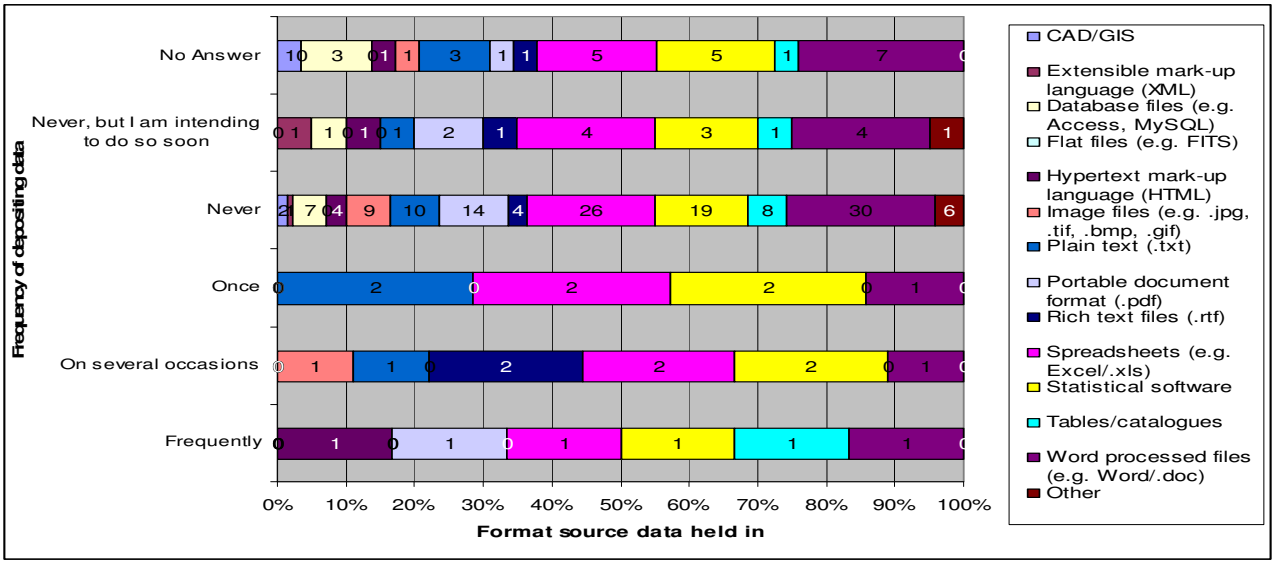
Much of the findings among the social science respondents indicates a greater use of ordinal-related data, including spreadsheets, statistics and other quantitative data. This is shown in the cross-tabulation between the frequency of respondents who had deposited data at the UK Data Archive and the kinds of data produced is presented in figure 20. It suggests that the most cited kind of data type generated by the seven social science respondents who had deposited data once or more were databases and quantitative questionnaire data (5 each respectively), raw data (4) and statistical data (3). Among those who had not yet deposited data but were going to do so soon, statistical data (5), databases and qualitative questionnaire data (4), quantitative questionnaire data and photographs (3 each) were cited.

Figure 20 – Frequency of data deposited in the UK Data Archive



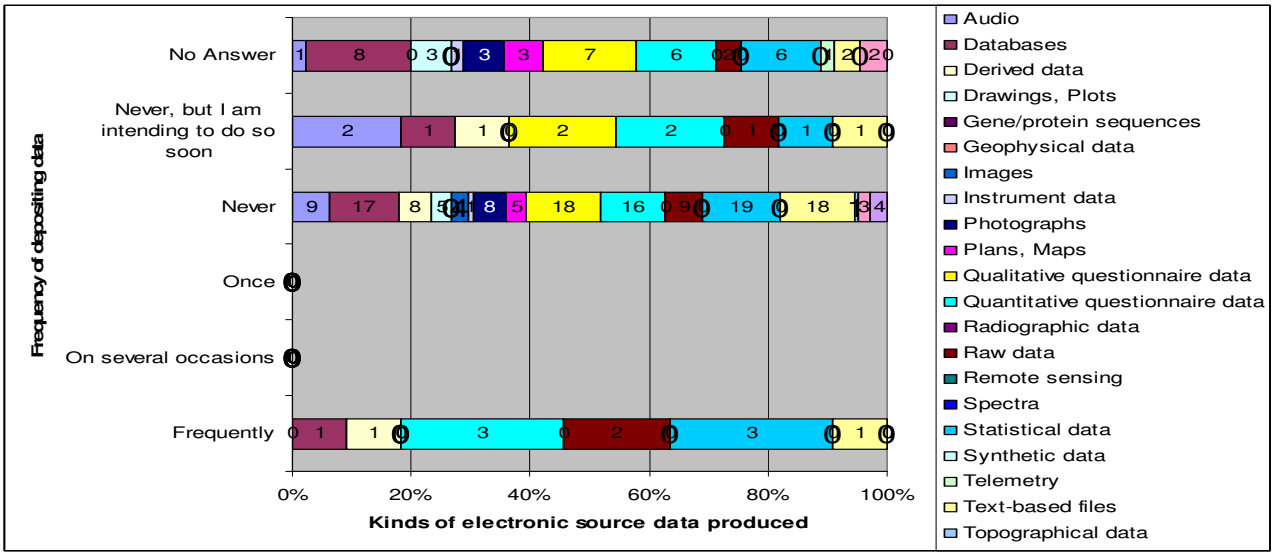
A further cross-tabulation between the frequency of data deposited at the UK Data Archive and the types of format in which the data is held and presented in figure 21 shows that among the respondents who deposited once or more the most common format used were spreadsheets and statistical software (5 each respectively). Of the seven who responded that they had not deposited data but intended to do so soon, word processed files and spreadsheets were cited four times respectively and statistical software three times.

Figure 21 – Frequency of data deposited at the UKDA



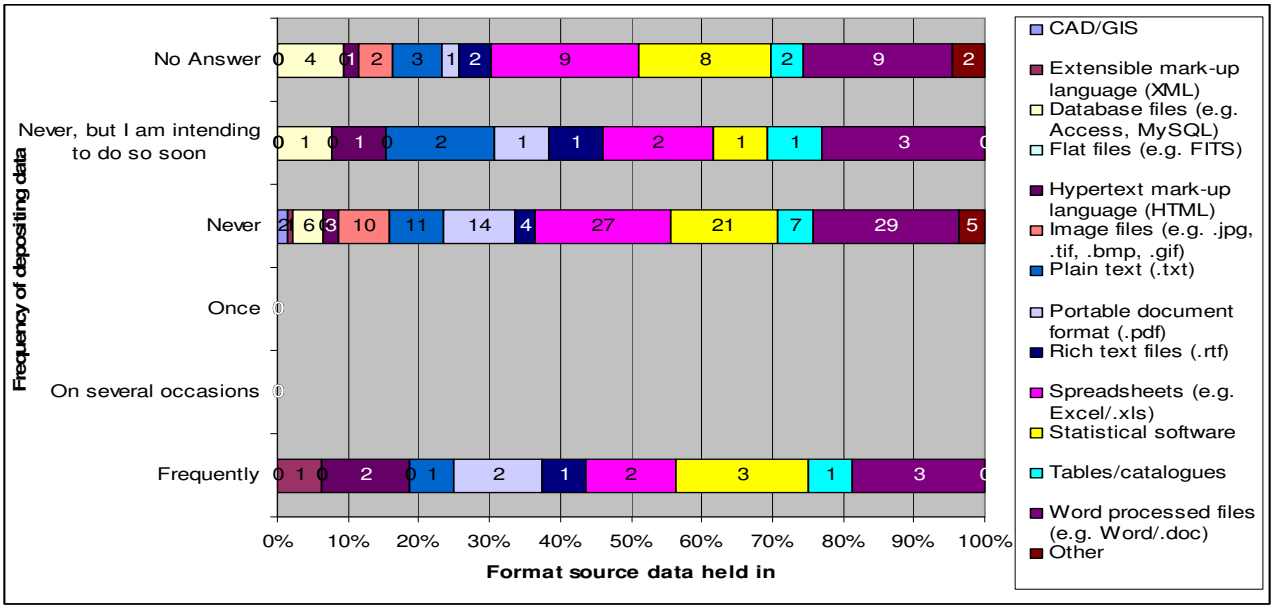
Of the social science respondents who claimed to deposit data 'frequently' in other repositories, all three claimed to produce quantitative questionnaire data (3), while two produced raw data (figure 22). among the three respondents who had not deposited data yet but were planning to do so, the most commonly cited kinds were audio, qualitative and quantitative questionnaire data (2 each respectively).

Figure 22 – Frequency of data deposited in other repositories by social science respondents



In terms of the format used by these respondents, figure 23 shows that all three had used statistical software and word processed files to store their data. Spreadsheets, PDF and HTML were cited twice each respectively. Of the three that had not yet deposited data but intended to do so soon, all three claimed to use word processed files while spreadsheets and plain text files were cited twice respectively.

Figure 23 – Frequency of data deposited in other repositories



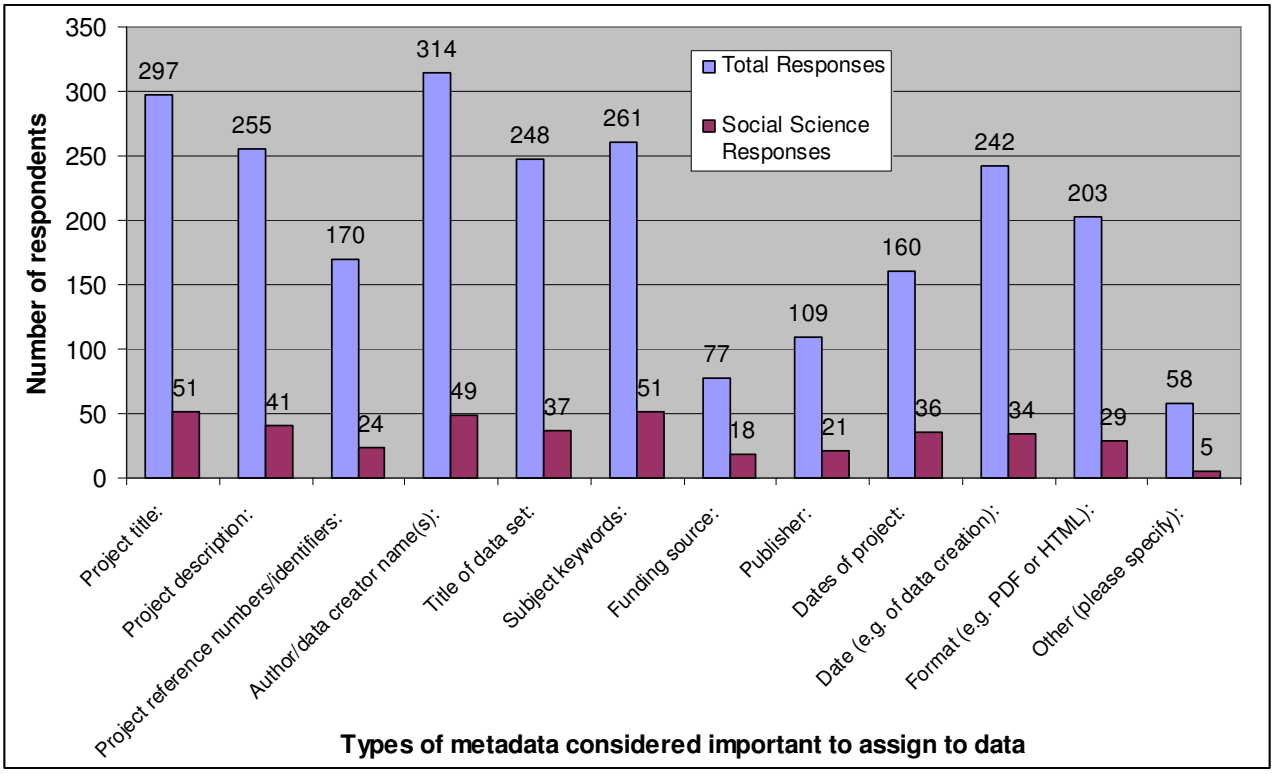
V. Metadata

Respondents were asked about their use of metadata with regard to their data in both the questionnaire and in interviews. Generally, social science respondents claimed that they assigned their own metadata with nearly a third claiming to do so during file-saving. The interviews enabled a more nuanced discussion to occur, in particular what social scientists understood metadata to mean. Through this process it was apparent that social science researchers' conceptions were substantially different to librarians; indeed, social scientists assumed that the file name given to a document on their computer was 'metadata'. Furthermore, social science interviewees indicated that metadata assignation was an ad hoc process, with the system they used devised for and prioritising their own concerns over that of other researchers.

1. Types, stages and responsibility of metadata assignment

The questionnaire listed a wide range of metadata options. When asked what metadata respondents assigned to their data, around two-thirds of all respondents claimed that the main ones they used were subject keywords (261), project description (255), title of data set (248) and dates (242). Just over three-quarters claimed to use the project title (297) as metadata (figure 24). For social science respondents, four-fifths used the project title and subject keywords (both 51) and author/data creator names (49) while two-thirds also reported using the project description (41). Among other responses given included 'copyright', 'country/time period etc', 'number of cases (quantitative data)' or none.

Figure 24: Types of metadata considered important to assign to data



Social scientists, like other respondents, most often claimed that they assigned metadata 'during file saving' (20 and 142 citations respectively), as shown in figure 25; this constituted slightly less than a third of social scientists and slightly more than a third for all respondents. However, beyond these two main references, whereas the responses in the questionnaire as a whole was spread across several different stages including 'as part of the indexing process' (98), 'when submitting data to the repository', and 'prior to data creation', the next most common responses for social scientists suggested that either metadata was assigned as part of the indexing process (14), not assigned or respondents were simply uncertain when this occurred (13 each respectively). Indeed, more than a third of questionnaire responses claiming that no metadata was assigned came from social science respondents (37.1%).

Figure 25: Stage at which metadata assigned to data

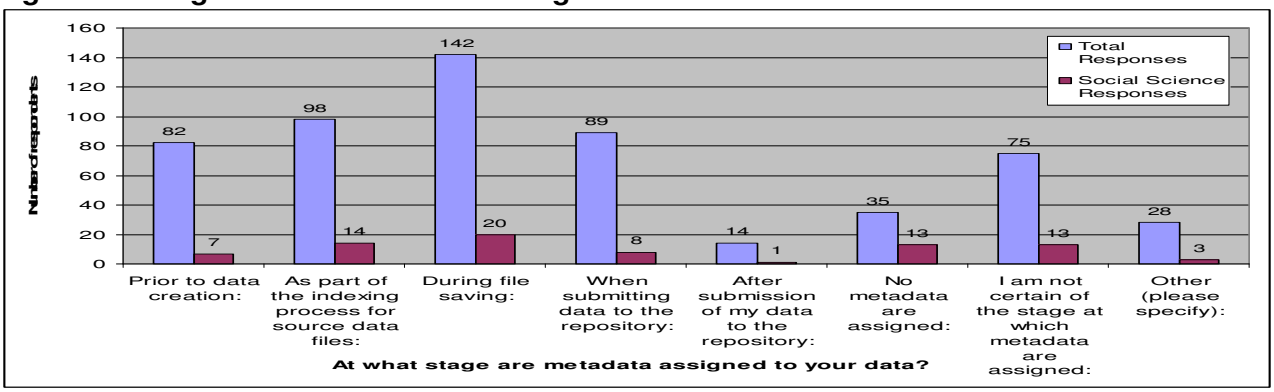
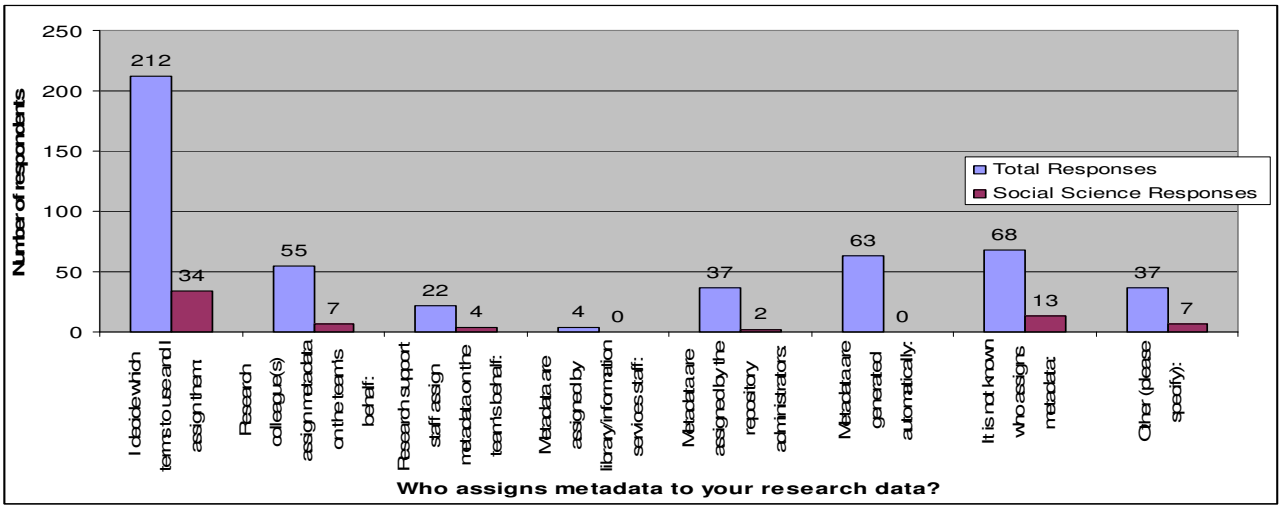


Figure 26 indicates that in the questionnaire as a whole and among social science respondents more than half see the allocation of metadata is a largely personal decision (212 and 34 citations respectively). But around a fifth of each sample did not know who assigned the metadata (68 and 13 respectively). But while a similar amount across all disciplines see metadata assigned automatically (63), no similar observation is made among social science respondents. Indeed, other metadata assignations are made either by research colleagues (7) or research support staff (4); the seven 'other' responses include four that metadata are not applicable or relevant.

Among the interviewees, the tendency towards self-assignation of metadata was similarly observed. B, an economic historian, admitted that he allocated metadata 'badly'. D, a social psychologist, and L, a university staff member in business and management, both relied on their research assistants to assign metadata. However, as the project leader both interviewees claimed the right to change it if necessary. For D though, this could prove problematic, especially when insufficient data has been collected by the assistants. E, a sociologist, assigns her own metadata to the raw material she unearths, even if it has already been given a label.

Both university academic staff and postgraduate students expressed concern with the prospect of obsolete technology and metadata. B gave the example of 8 inch floppy disks being currently unreadable by most of today's electronic hardware, while H, a postgraduate student in public administration, highlighted the potential change to government metadata that could have an impact on his work (e.g. Comprehensive Performance Assessment in the ODPM). More common though was an admittance by interviewees that little thought had been given to the allocation of metadata in the first instance or how subsequent users would be able to access it (G, J).

Figure 26: Responsibility for assigning metadata according to respondents



2. Relationships between metadata assignation and frequency of depositing data

The allocation of metadata when broken down by the frequency with which social science respondents deposit data in source repositories shows general usage of project and data set titles and dates, as well as the use of author/data creator names and formats as the preferred way of identifying data. The respondent who had deposited data in the CERN repository 'on several occasions' claimed to have used the following metadata: project title and description, title of dataset, subject keywords, publisher, dates of the project and date.

Of the seven social science respondents who had deposited data in the UK Data Archive at least once, the most commonly cited metadata were the project title and the date (six times respectively), as shown in figure 27. These were followed by the use of the author/data creator name and format (five times respectively). Among the five who were intending to deposit data but had not yet done so, all claimed to use the project title and author/data creator as metadata, followed by the project reference numbers/indicators, titles of the data set, the project date and format (four times respectively).

Figure 27: Types of metadata used in relation to respondents' depositing of data in the UK Data Archive

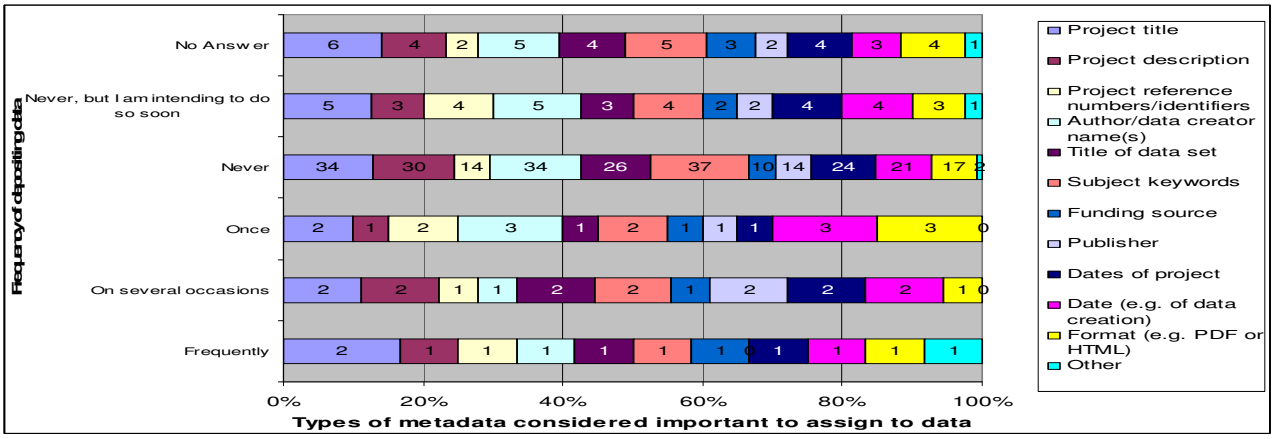
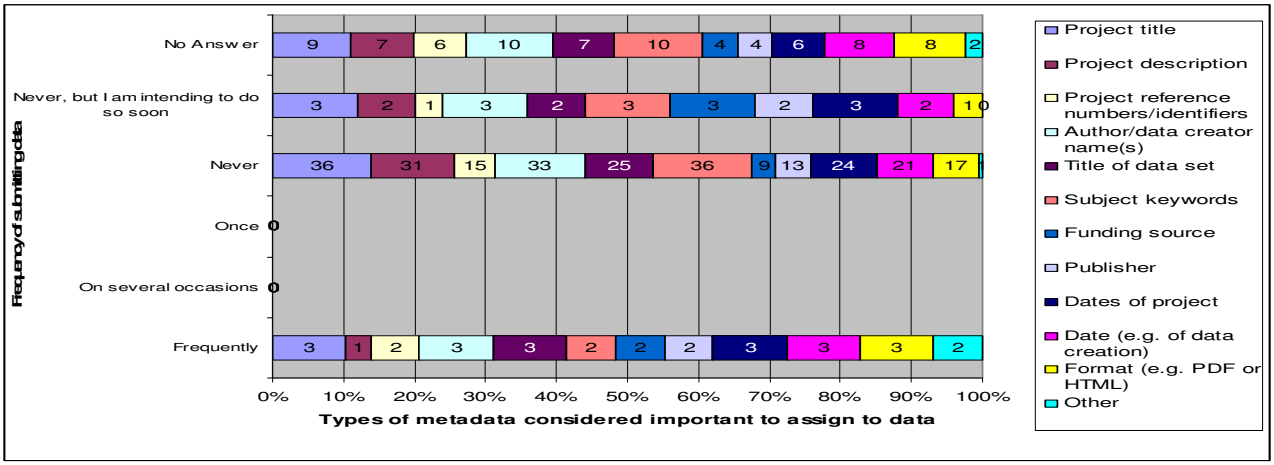


Figure 28 shows that of the three social science respondents who had deposited data in other repositories, all had made use of the project title, author/data creator names, title of the data set, project dates, dates and format. While not as widespread, the three who were also intending to deposit data in the future all claimed that they made use of the project title, author/data creator names and project dates.

Figure 28: Types of metadata used in relation to respondents' depositing of data to other repositories



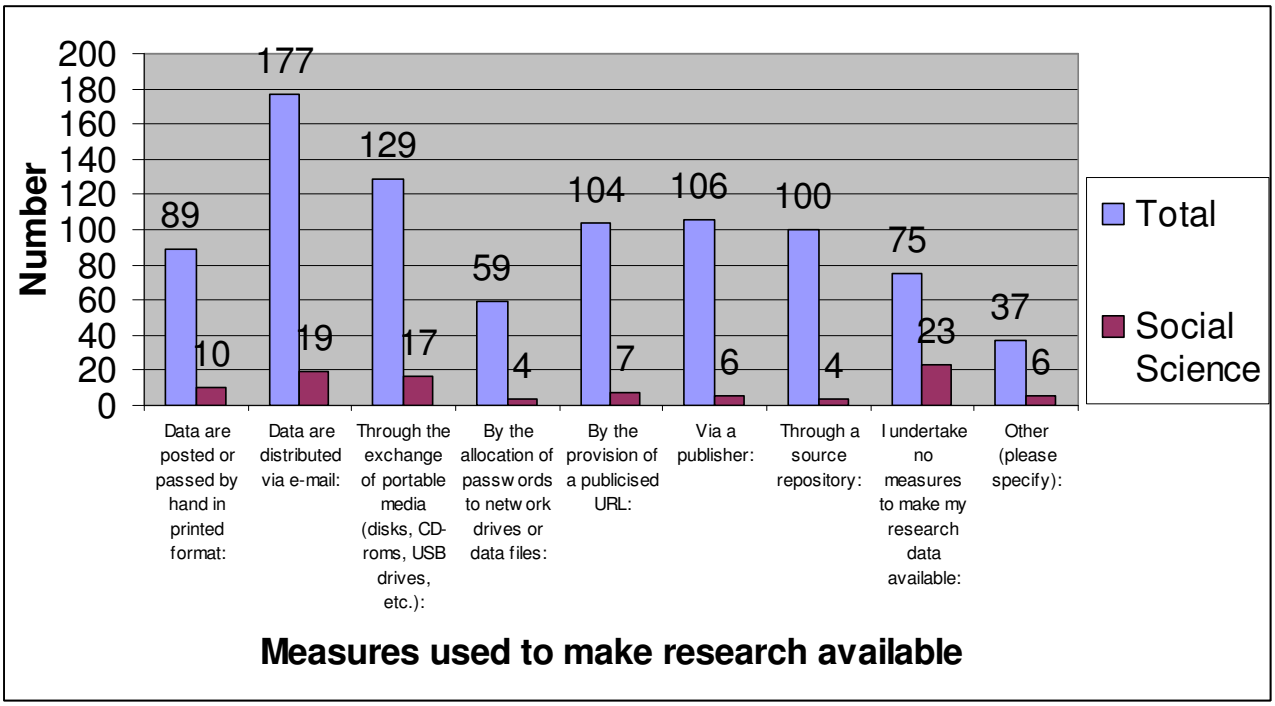
VI. Data Access and Sharing

The findings in the questionnaire and interviews suggested that social science researchers were supportive of sharing data in principle, although when applied to themselves this could be more awkward. In particular they were concerned that they gain the maximum use of their own research data prior to making it available to others. Furthermore, the way they presently share data remains largely informal; given the fact that most maintain their research data on closed networks, they tend to respond to personal requests for data and make a decision on that basis.

1. The accessibility and sharing of primary research data

Among respondents as a whole, nearly half claimed that they made research data available through email (177) while a third used portable media (129) (figure 29). By contrast more than a third of social science respondents said that they did not make their research data available (23). The lack of data sharing among social science respondents was found among those interviewed as well: both university academic staff and postgraduate students claimed to have made no effort to actively share their research data (H, I, K, L, M, O). Part of the reason may be attributed to a sense that the research was not yet complete enough to justify it being shared (e.g. O), which was also echoed in the comments made by questionnaire respondents. However, of the measures used to share data among social science respondents, the questionnaire findings suggested that a third reported using email (19) and portable media (17), ensuring that they were in line with the global findings as a whole.

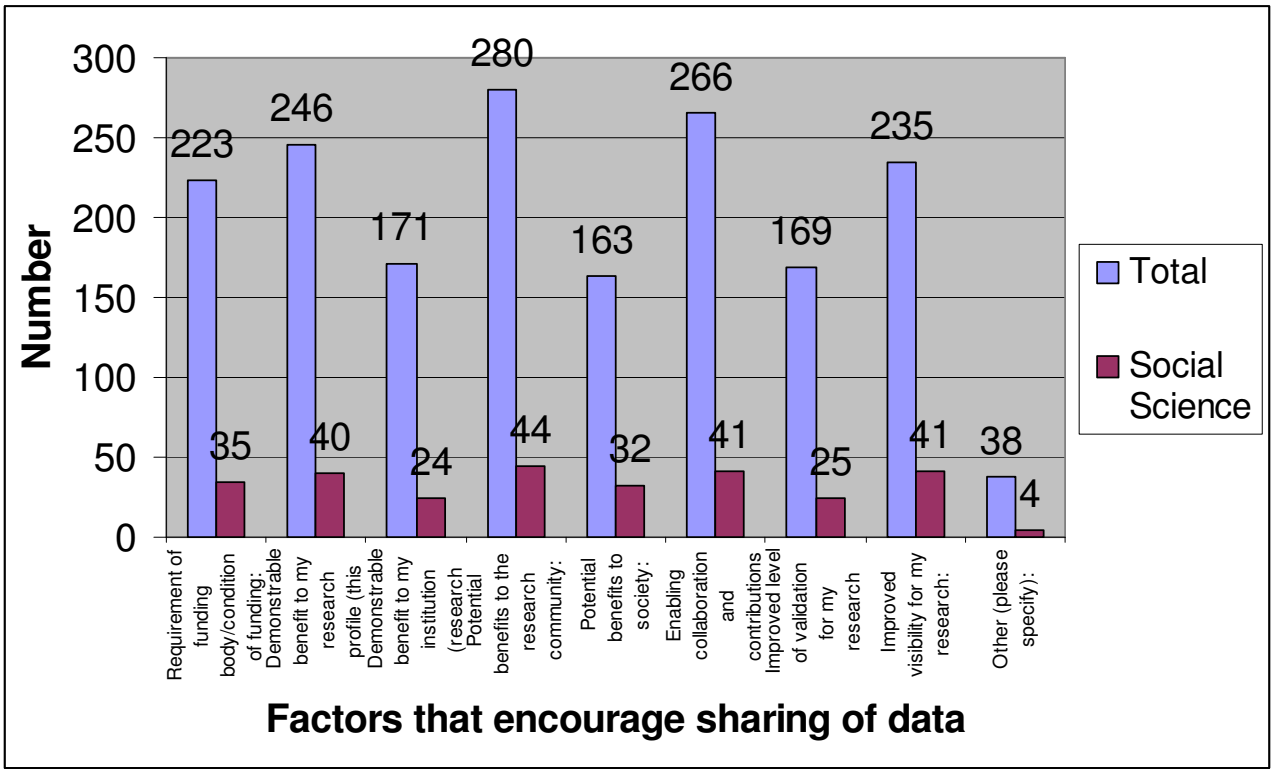
Figure 29: Ways of making research data available



Social science respondents were no different to respondents as whole when asked what factors might encourage them to share research data, as shown in figure 30. The reasons were varied, but the same themes occurred in both samples, including its potential benefits to the research community, enabling collaboration and contribution by others, demonstrable benefit to a researcher's research profile and improved visibility for a researcher's research.

Among social science interviewees, there was a general endorsement of the importance of sharing research; a postgraduate student in media and communications, O, likened data-sharing to 'file-sharing' in that contributing one's own material provided benefits to the wider community and offered the prospect of obtaining useful data for oneself; C, an anthropology postgraduate student, saw a tension at work between the benefits gained from data-sharing and the competition between researchers that might discourage it from happening.

Figure 30: Factors that encourage sharing of research data



Both social science respondents and respondents more generally shared common concerns regarding data-sharing. As figure 31 shows, both types of researchers fear its premature use and a perceived loss of ownership as among the most cited factors (with half of science respondents claiming so). For all respondents, the third most common factor was a question of time and effort required to share data (193), while for social science respondents it was a matter of data protection and confidentiality (29). Indeed, ethical considerations were notable as being proportionally of greater concern among social science respondents to all other respondents (39.7% of all responses).

The interviews with social science researchers reiterated these common themes and concerns, across both staff and student communities. C, a postgraduate anthropology student, saw a distinction between 'formal' interviews and 'informal' field notes. Various researchers, including D, F, G, H and N, all individually commented on the issue of confidentiality and the sensitiveness of the material they generated through their own interviews with informants. Indeed, D, a social psychologist, observed a difference between the British emphasis on 'efficiency' and sharing data against a presupposition that was not geared towards this in Southern Africa, through a greater concern with ethical practice and social justice in dealing with informants.

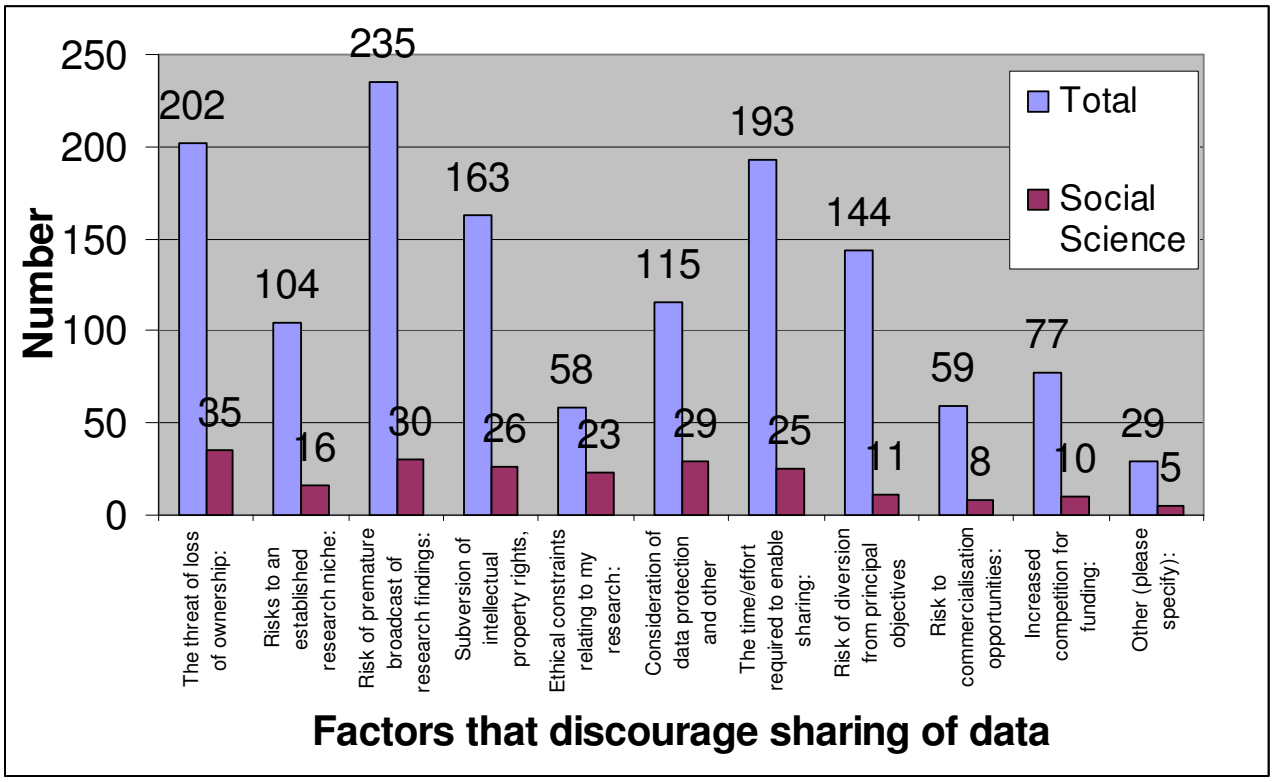
G, a research assistant and sociology student, said that he was concerned other researchers would focus criticism on his research data and methods rather than the publications and findings he

produced as a result. Others postgraduate students, K and O, expressed a concern at not being credited with the use of their data.

M, a development researcher, felt that not only were there issues of ownership to consider, but also the extent to which another researcher would be able to use that data without incurring its cost of production. N, a social policy staff member, suggested limitations with the Data Protection Act: namely that interview participants own the copyright of the material they produce while the researchers own the audio files in which they are kept.

These concerns highlight a tension among social science interviewees: whereas there was a general degree of support for sharing data across the community, there was a certain amount of personal self-interest: P, an economist, summarised this by distinguishing between being a 'producer' and 'consumer' of research data. Whereas sharing data was ideal for a consumer, for a producer there was an inclination to keep what was produced – at least until sufficient use of it has been achieved. B, an economic historian, offered a solution to this dilemma, by entitling the producer of the research data to a 'monopoly' or 'patent' on the data for a certain period to ensure the he or she derived maximum benefit before making it more widely available.

Figure 31: Reasons that discourage sharing of data

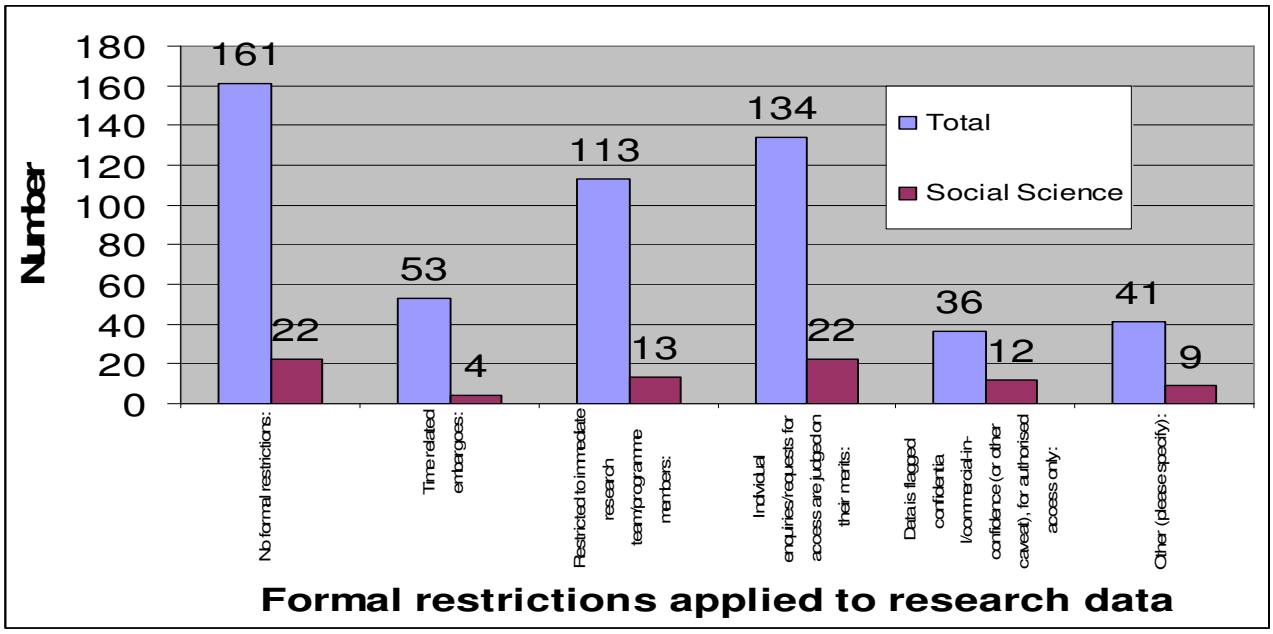


Social science respondents were not very different from respondents in general in the types of restrictions they claimed to want to see on their research data, as shown in figure 32. Both types of researchers advocated that there should either be no formal restrictions or that individual requests be judged on their merits. On these two points, just over a third agreed with these sentiments. Respondents also reported concern that 'any [data] sharing would de-anonymize the participants' and that 'when referring to individuals, some data is kept confidential'.

Among interviewees the response was largely similar. B, an economic historian, who was one of the most enthusiastic supporters for data-sharing (and who wished to see the ESRC enforce its data-sharing conditions as part of its funding agreements), said that despite not actively sharing his data, he did look at individual requests. He noted that particular requests took longer to be assessed by him, most notably journalists. Indeed, a sports psychologist, I, added to this reservation with the media by observing a journalist who had approached her to use her tentative findings in an article, only to argue for a more robust conclusion than was the case. L, a business and management lecturer, said that

she did not share data unless she trusted them and know how they were going to use it. However, H, a public administration postgraduate student, argued that sharing data was like lending money to a friend: once you had done so you could not complain how they used it.

Figure 32: Types of restrictions by respondents on access to research data

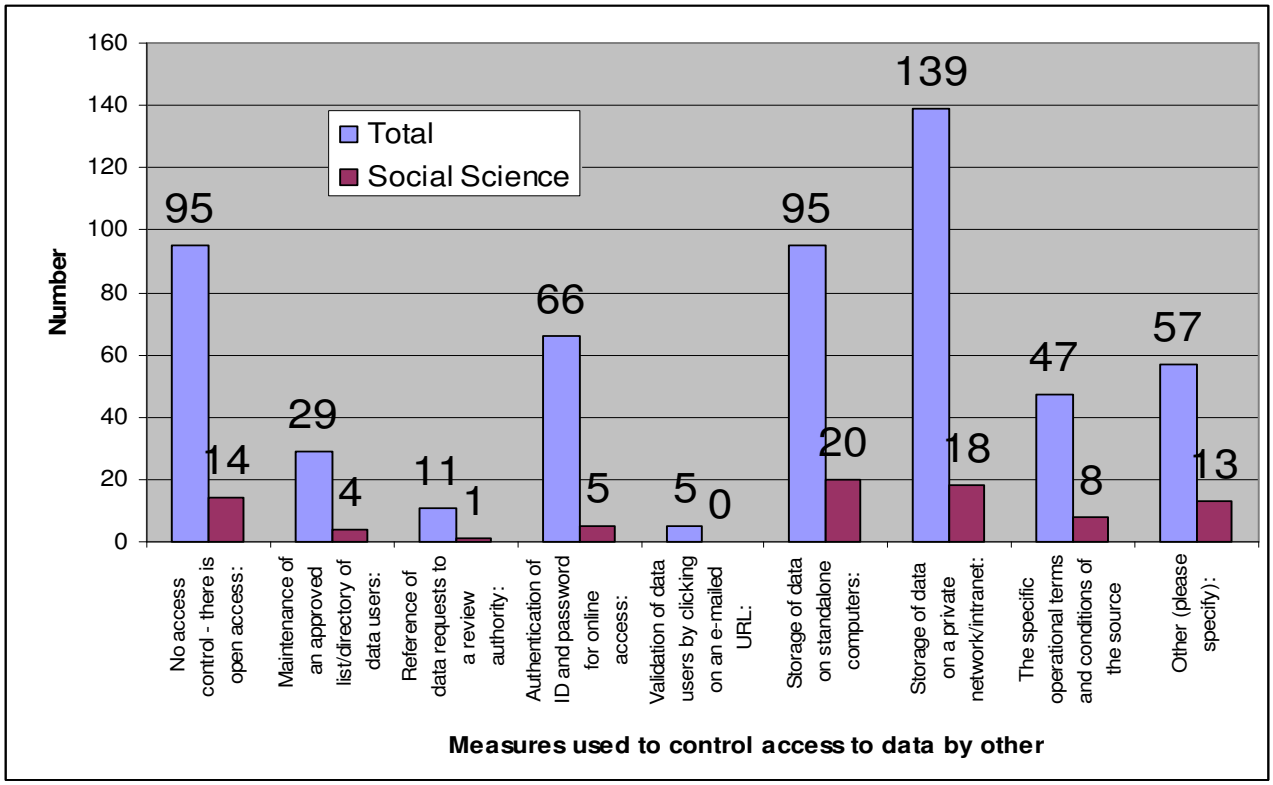


Around a third of all questionnaire respondents (139) claimed to store data on a private network/intranet to ensure the control of access to data by others (figure 33). This was followed by storage on standalone computers and no forms of access control (each 95 respectively). For social science respondents, similarly a third claimed to use standalone computers (20) or private networks/intranet (18), while just under a quarter (14) claimed open access to data. Among those that cited 'other', several common responses were observed relating to individual requests from one researcher to another, sometimes with time-delays.

Among the interviewees, the personal route was also in evidence. As noted above, several researchers observed their preference for an individual request for their data (B, I, L and H). Similarly, they tended to behave in the same manner when seeking data; if they wanted it they would approach a researcher for it if needed. Indeed, I said that she much preferred this route over all others, given the relative intimacy that direct contact afforded.

The interviews also revealed other ways of controlling access. C, a postgraduate anthropology student, said that she would not make available her interview transcripts in full, but rather parts of them. H, a public administration postgraduate student, said that depending on the request, he would only make available the material necessary for that individual or organisation, thereby ensuring the minimum was provided.

Figure 33: Methods of controlling access to research data by others

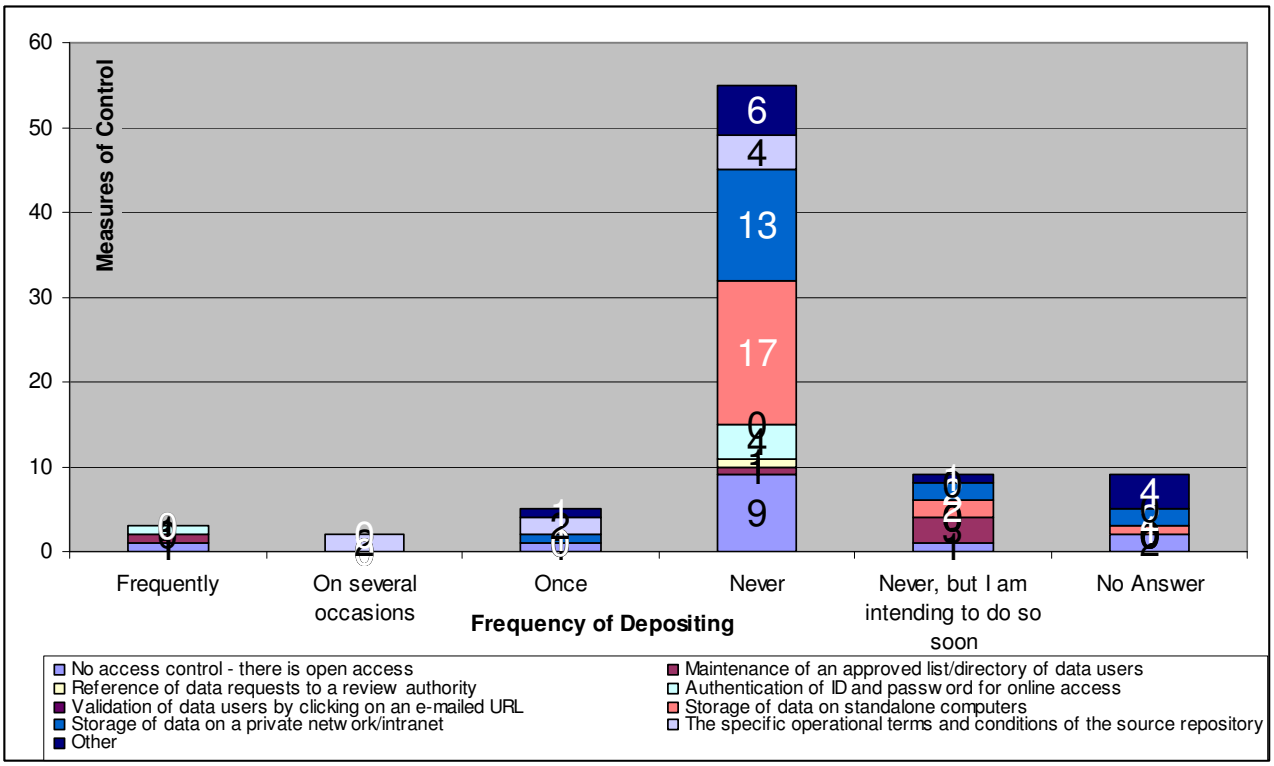


2. Measures to control access to data by frequency of depositing data in repositories

There was a wide range of opinions regarding control of data in repositories by social science researchers who had deposited data across different repositories. This ranged from deferring to the repository itself to determine those constraints through to not imposing any access restrictions at all.

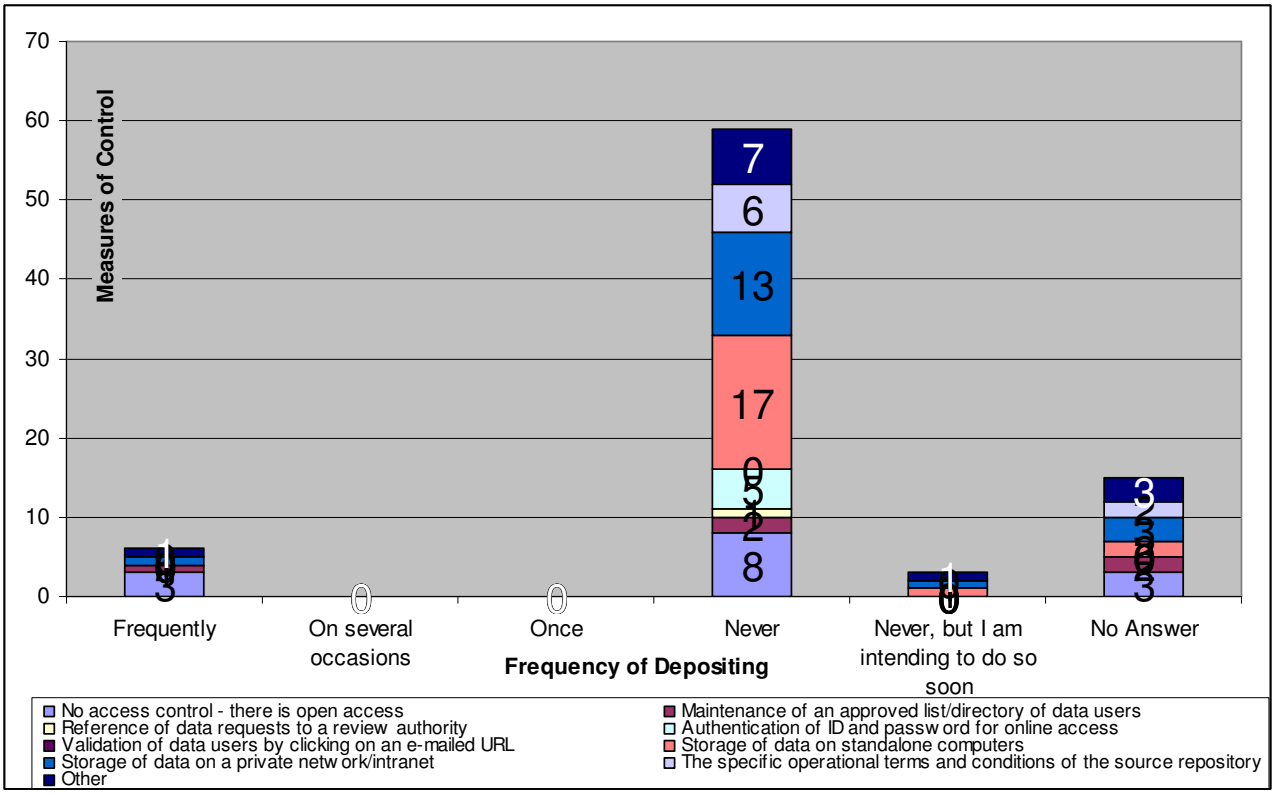
Only one social science respondent had claimed to do so in the CERN repository frequently. When comparing and contrasting the ways in which researchers control access to research data by the frequency with which they deposit data in repositories, this respondent cited the repository as determining the conditions of access. Among the seven who had deposited at the UK Data Archive at least once, the response was mixed, as shown in figure 34. The most cited response – four – was to say that the UK Data Archive decided upon the requirements, while there were two statements that there was no access control and one each for the maintenance of an approved list/directory of data users or authentication of ID and password. Among the five social science respondents who had not yet deposited at the UK Data Archive but intended to do so, there was a similarly broad mix, although three observed the maintenance of an approved list/directory of data users and two each for the use of standalone computers and storage of data on a private network/intranet.

Figure 34 : Respondents' methods of research data control by frequency of depositing at UKDA



When looking at the three social science respondents who had deposited data in other repositories other than the UK Data Archive and CERN, all three claimed that there was no form of access control, as shown in figure 35. One respondent maintained that data was controlled either through the maintenance of an approved list or through data storage on a standalone computer. Another noted that while there was open access, this occurred 'after time delay'. Of the three respondents who had not yet deposited data but would do so soon, one said that there was open access, one used standalone computers, another used a private network/intranet. One noted that 'this is difficult - we have a sharing scheme for post fieldwork anthropologists and we have to share with each other - but outside our circle it is not common - unless recommended by supervisor'.

Figure 35: Respondents' methods of research data control by frequency of depositing at other repositories



VII. Output repositories

The previous section considered social science researchers' attitudes towards sharing research data with particular reference to those who had deposited data in source repositories, or who were about to. In this section social science researchers' use of output repositories is analysed. For the purposes of the questionnaire, output repositories vary in type, from those based in a university (institutional), with a publisher or concerning a discipline. However, the interview phase that complemented the questionnaire suggested that social science researchers saw these distinctions as artificial. Rather they generally tended to associate them together in conversation, as a means to identifying and locating published material for their own work.

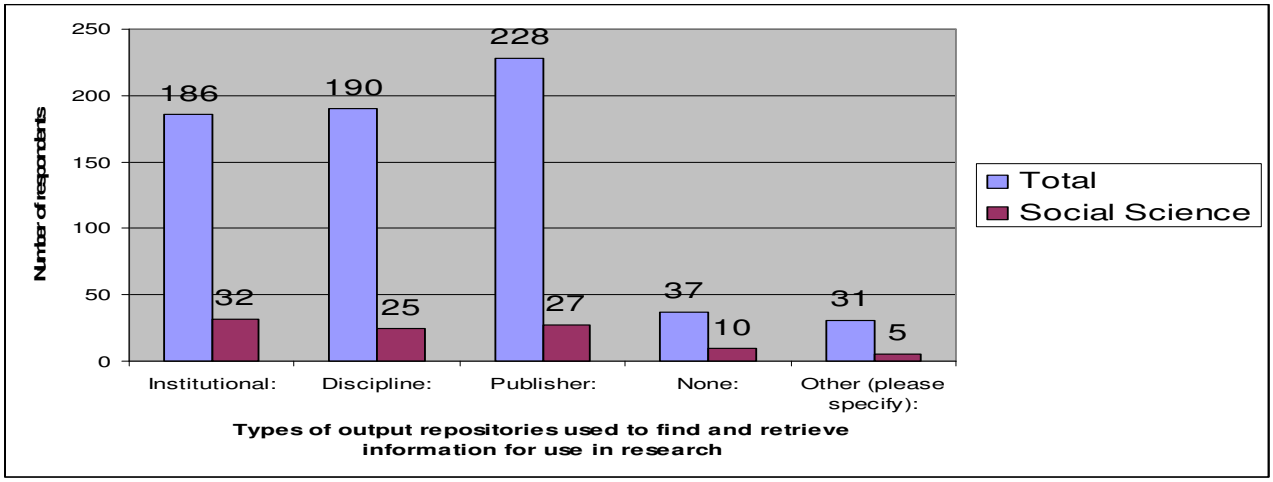
1. Respondents' use of output repositories

While two-thirds (228) of all questionnaire respondents claimed to use publisher repositories for research, less than half (27) of social science respondents reported doing so (figure 36). Slightly more than half of social science respondents claimed to use institutional repositories (32) while nearly a third of all responses claiming that no repository was used came from social science researchers. Indeed, there was little substantial difference amongst social science respondents regarding the use of different output repository types for research purposes. Among the few other repositories used, social science respondents cited the 'ESDS Penn World Tables at Toronto University', Google Scholar and 'mailing lists, personal contacts'.

For social science interviewees the most common type of output repository cited and used for research were publisher repositories. Particular ones cited included J-Store (C), BIDS (D, F, P), IBSS (M) and Web of Science (P). E, a sociologist, commented that her research involved analysis of previous work and publications; consequently, she spent much time browsing through the bookshops at the LSE (Economist Bookshop) and Goldsmiths College and the Amazon booksellers website. While the majority made use of the e-version of articles and publications founding these repositories, at least two (E, I), claimed that they used the online system to locate the relevant journal before turning to the paper-based format. Of the other two repositories, the discipline repositories interviewees highlighted REPEC (Research Papers in Economics) (P) and Psychlit, Sociofile and DAIS (N). Institutional repositories included the Office of National Statistics (F, H), the LSE (F, N, O), the British Library (O) and the work/employment relations repository at Sussex University (H).

Against these observations it was observable that in the interviews there appeared to be a low level of awareness and understanding of the differences between these various repository types; output repositories were generally assumed to be broadly the same. Indeed, one postgraduate student, O, confused source with output repositories with output, when asked where he found his data. He responded by saying that he made use of other researchers' articles.

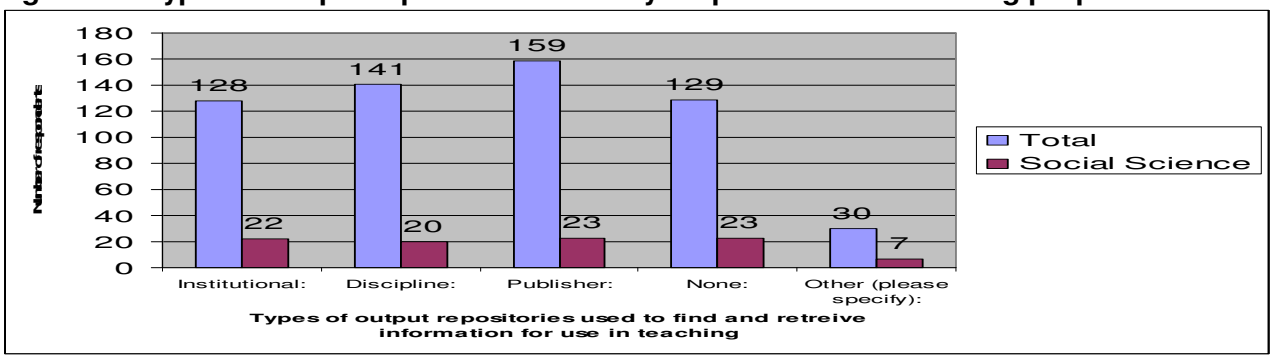
Figure 36: Types of output repositories used by respondents for research purposes



Among all respondents, publisher repositories was the most clearly cited source for teaching aides (159), followed by discipline (141) and institutional (128) ones (figure 37). Among social science respondents the findings were less conclusive; there was little substantial difference between all three types. In contrast to the repositories used for research purposes, it was also notable that a higher proportion of respondents claimed not to use any type of repository for teaching, making it broadly similar to the other responses provided (23).

There were few interviewees who commented on the use of output repositories as a means of locating material for teaching purposes (B, E, I). B, an economic historian, mentioned that he made his teaching material (both that produced by himself and others) available on his personal website, which was password-protected through his university. Neither E, a sociologist, nor I, a sports psychologist, made much use of output repositories for teaching. They were mainly concerned with teaching the core texts to students, so saw little need in accessing the most up-to-date material available. E added nuance to this approach, by distinguishing between undergraduates, who she felt did not need to go online, and postgraduates, who should be familiar with the core texts and therefore had a reason to access the latest work. B similarly saw a problem with students using e-resources. In particular he highlighted the fact that not all publications were presently available online, resulting in students assuming that certain articles did not exist rather than accessing the paper versions.

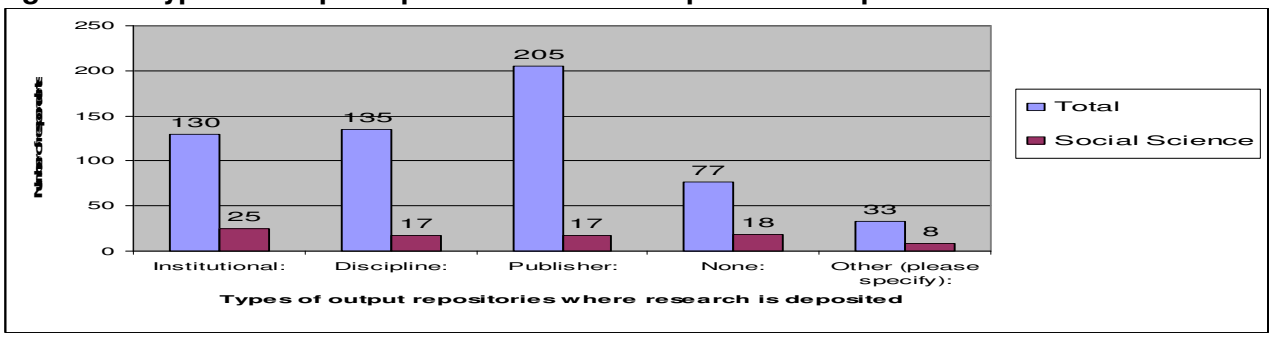
Figure 37: Types of output repositories used by respondents for teaching purposes



Among all respondents in the questionnaire, the most cited output repository where researchers deposited research are those hosted by publishers (205). Yet among social science respondents the position was less clear. Nearly half claimed to have deposited data in institutional repositories, while the same response rate was recorded for both discipline and publisher repositories (17). A similar number (18) reported never having deposited data in an output repository (figure 38).

Arguably the relative lack of awareness between different types of output repositories (see above) may be explained by the distance between them and researchers. This is evident in the 'other' responses that were given , including: 'None, but some things are picked up by IBSS and the like, as well as by services such as ID21 at IDS, Sussex' and 'I publish my research, and the publisher (including my institute, for informal publication) automatically acts as the 'depository' in your sense.' Another noted that 'when I publish it will be in anthropology journals'.

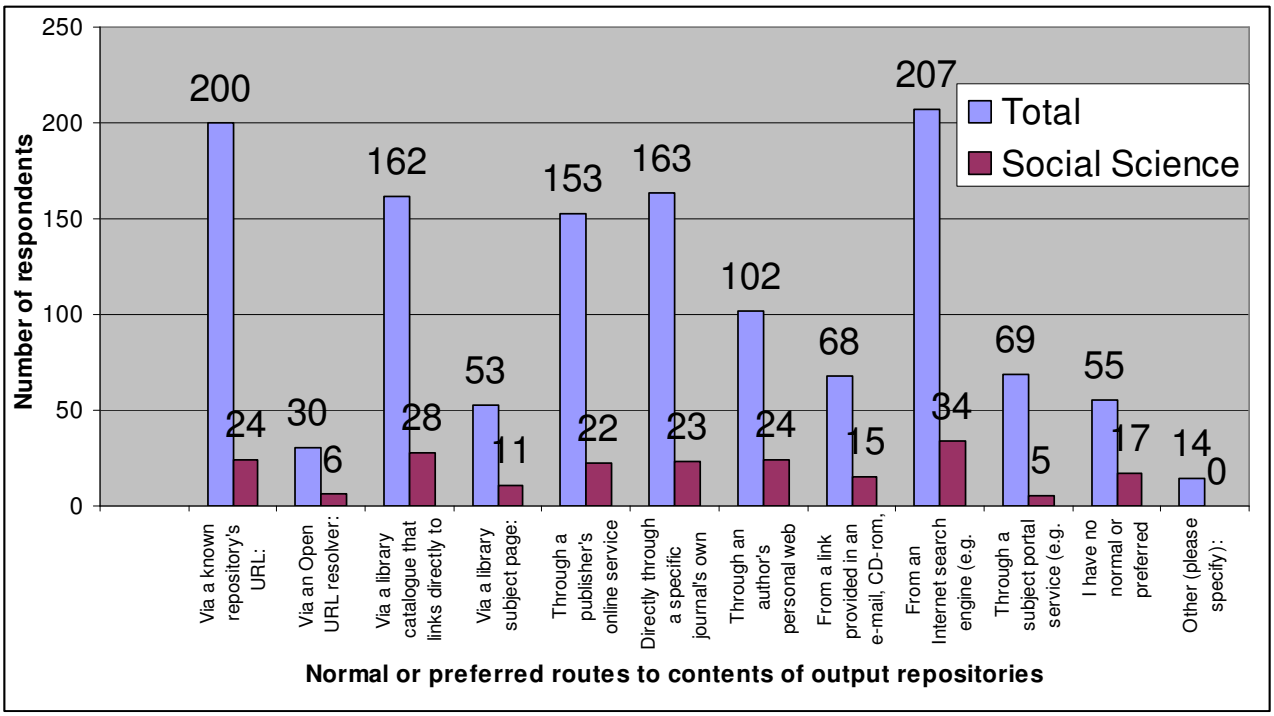
Figure 38: Types of output repositories where respondents deposit research



When asked how they sought out the contents of output repositories (figure 39), the most popular way for more than half of all respondents and in social science was through an internet search engine (207 and 34 respectively). Among the total 377 respondents the only other approach that garnered more than half was via a known repository's URL (200). Among social science respondents after internet searches there was little substantial difference with many of the other alternatives, with the exception of an open URL resolver (6) and a subject portal service (5).

The use of internet search and the other forms suggested in the questionnaire (e.g. a known repository's URL, library catalogue links, publisher's online services and an author's personal webpage) were cited by the interviewees. Although all social science respondents claimed to make use of search engines, the most cited one was Google (A, H, J, K, M, N, O). Despite the existence of an academic portal by the company, Google Scholar, interviewees generally commented use of 'Google'. One researcher, M, was not aware of Google Scholar while P, an economist, specifically stated that she used Google rather than Google Scholar. According to H, a postgraduate public administration student, Google is sophisticated and picks up non-cited material elsewhere. M, despite not using Google Scholar, claimed that Google was good at picking up what he missed from the predominantly Anglo-American centric IBSS. Similarly, O, a postgraduate, media student, highlighted the more international perspective of Google by commenting on his use of the Portuguese version for his research on Brazilian television. However, N, a social policy lecturer, expressed concern that she did not always know which version of a paper she found on Google was (which is incidentally being addressed through JISC's Versions project).

Figure 39: Respondents' normal or preferred routes to contents of output repositories



Nearly two-thirds of all respondents (223) and more than half of those in social science (35) claimed that they found simple search methods sufficient when browsing through output repositories (figure 40). Indeed, less than a third of respondents in either category said they made use of alternative search techniques.

This emphasis on simple search techniques was emphasised in the responses given by social science interviewees. Generally these individuals did not comment on difficulty accessing what they needed. Several reported using a range of different search engines – L, a business and management lecturer, used Ingenta, Emerald and Econlit to look for material while C, a postgraduate anthropology student used the American Anthropological Association, J-Store and other university repositories – to look for what they needed. However, this did contrast with the response from one social science respondent in the questionnaire, who claimed that 'It would be helpful if there was one database that would have

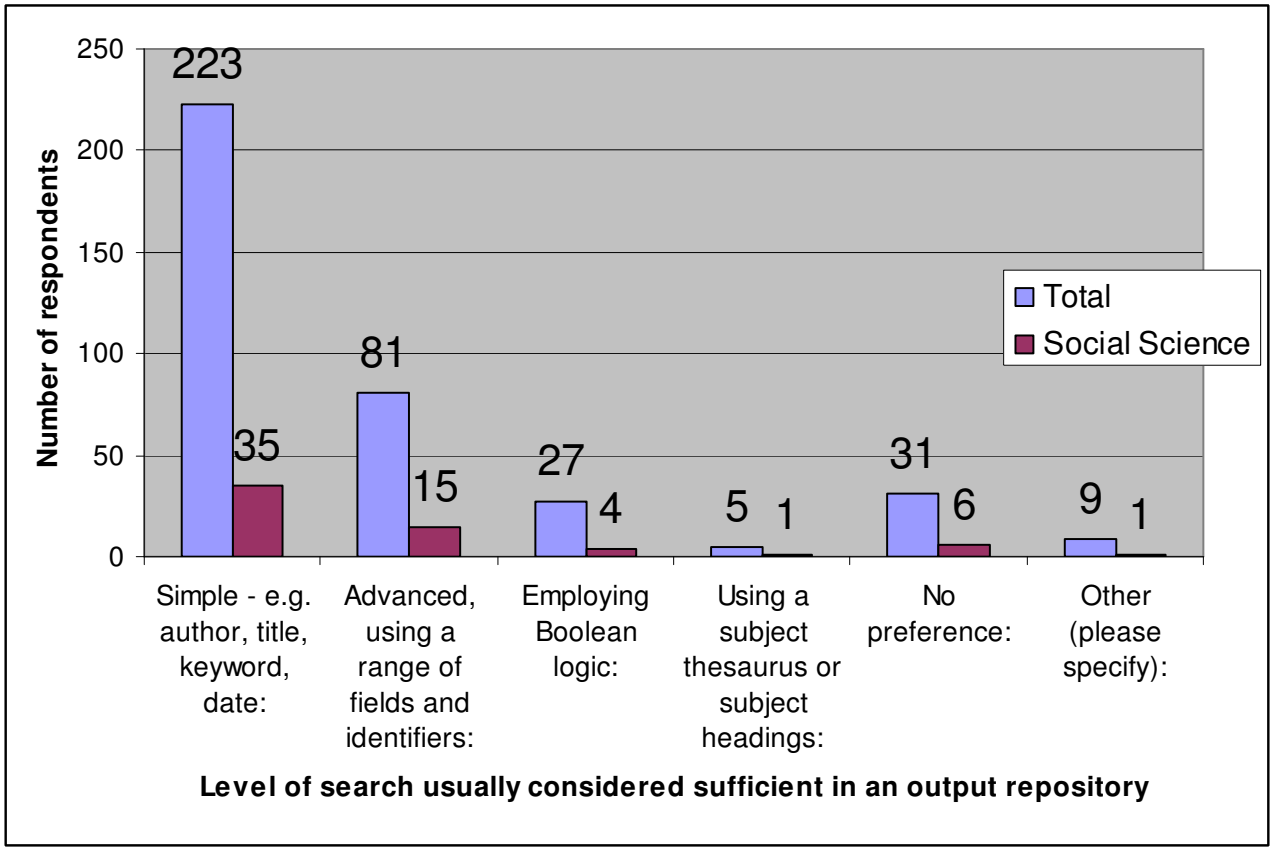
EVERYTHING on. As it is, I need to check at least 5-6 different databases as each comes up with different results. Also, links to the actual papers are v[ery]. v[ery]. useful. I normally have to go through Google Scholar for that type of service though.'

Besides using several different search engines, others noted that different combinations of words could be used within the search engines themselves to yield the necessary publications; G, a sociology postgraduate and research assistant, reported narrowing the field down from 'PhD' in his work to 'doctoral dissertations' while O, a media postgraduate, observed the limited material available on 'Brazil' as against 'Latin America'. H, a postgraduate student in public administration, said that he made use of keywords, authors, titles and dates to access the material he wanted.

Despite these measures taken by researchers, some commented on the limitations involved with search. N, a social policy lecturer, admitted that taking incomplete references from the found publications hampered her subsequent work by making them difficult to locate. B, an economic historian, felt that search was not as intuitive as it could be; for example, typing in 'craft' as an author's name might yield irrelevant material related to 'arts and craft' while typing in '1953' might exclude relevant material which was stored under '1950-55'. He suggested that greater consideration given to the metadata assigned to such material might be of use in this regard.

Indeed, given the suggestions on ways to improve searching, while some of the interviewees (B, H) and questionnaire respondents argued for 'enhanced subject search', 'more detailed cataloguing' and 'easier Boolean searching that doesn't unnecessarily screen', another claimed that 'given Google-type search engines, it is now usually a waste of time to design new indexing'.

Figure 40: Respondents' level of search in output repositories



2. Relationships between respondents' use of output repositories and depositing in source repositories

Among social science respondents who had deposited at least once in source repositories, the most cited use of output repositories were institutional ones, as shown in figures 41, 42 and 43. Of the 11 respondents who claimed to have deposited data in the CERN, UK Data Archive and other source repositories, six had used institutional repositories, three had made use of publisher and two of discipline-specific repositories. In addition two others had not made use of any output repository.

By contrast discipline repositories were the most used form of output repository by the eight social science respondents who had not yet deposited data but intended to do so soon. Seven of them claimed to have used discipline repositories with seven reporting use of institutional repositories and five of publisher repositories.

Taking these two findings together suggests that institutional repositories were the most popular type of repository used by social science respondents that had either submitted or were about to submit data to a source repository. This goes against the overall findings which indicate that amongst all social science respondents the most common form of output repository were those belonging to publishers.

The interviews did not yield any explanation as to why the difference between output repository popularity varied between depositors and non-depositors. Of the five interviewees who reported having submitted source data, four (A, B, J, P) reported using e-resources to find publications and articles; one (E) did not make use of such facilities and admitted being unaware of whether her university had an institutional repository or not. Only two reported using institutional repositories: B, an economic historian, placed teaching material in a password-protected section for his students while J, an economist, deposited her working papers and publications at both her university repository and on an external website.

Figure 41: Respondents' use of output repositories against depositing data in CERN repository

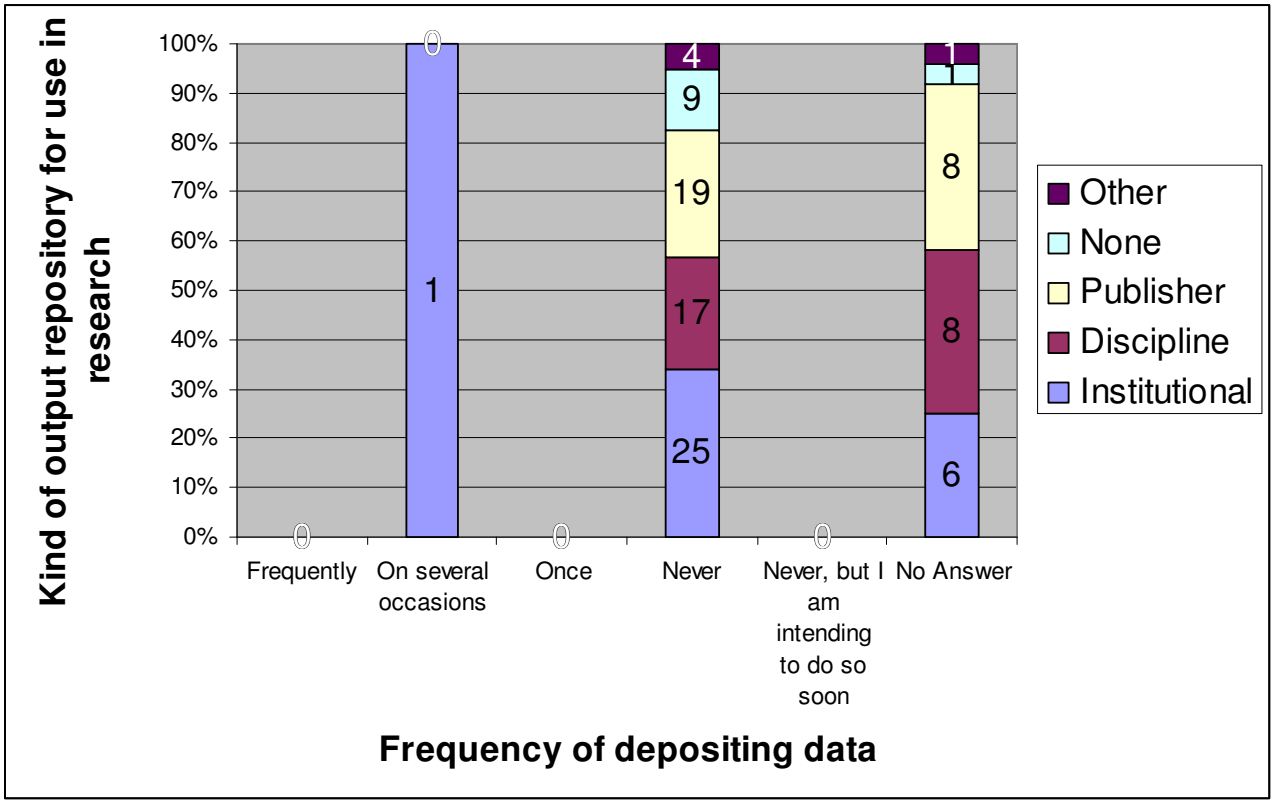


Figure 42: Respondents' use of output repositories against depositing data in UK Data Archive

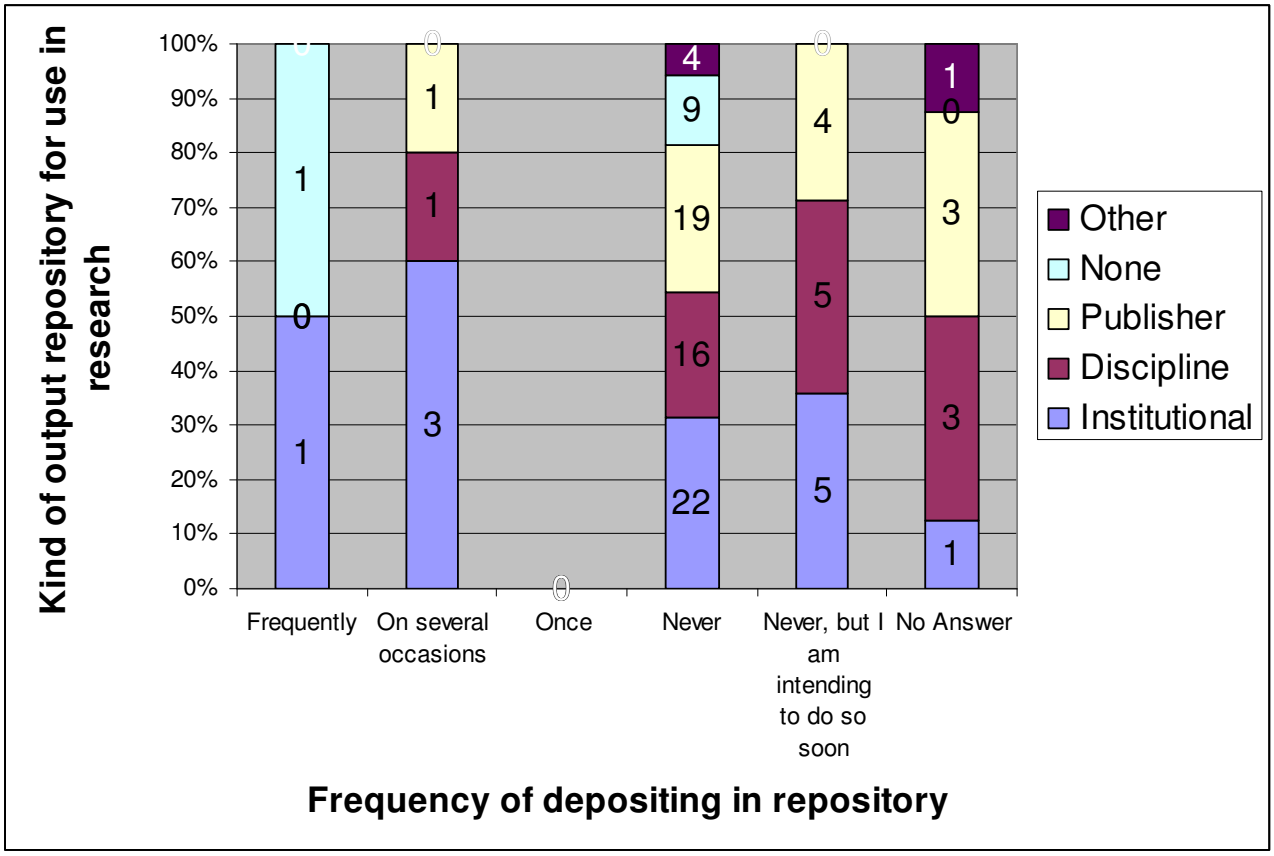
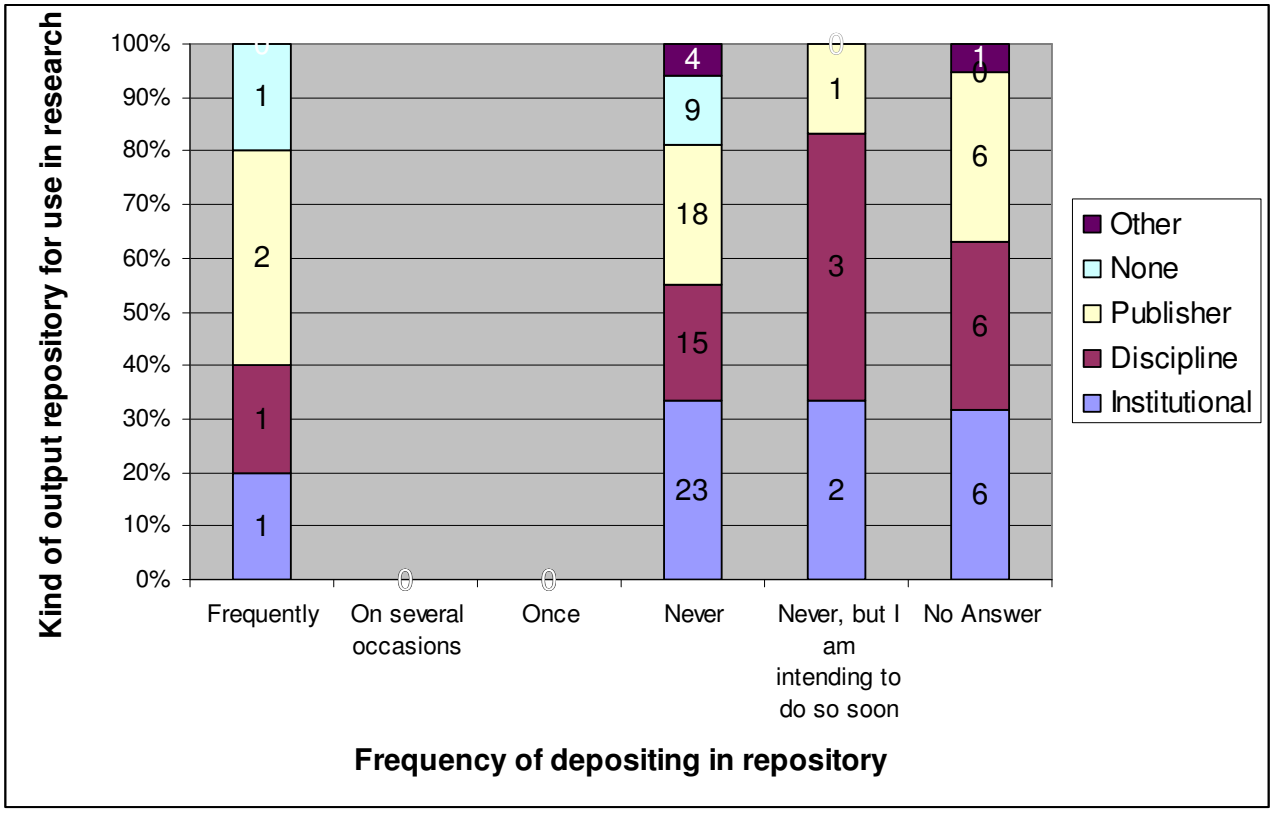


Figure 43: Respondents' use of output repositories against depositing data in other repositories



NB Despite three claiming to frequently submit data to other repositories, the questionnaire enabled more than one response to be ticked. Consequently, this resulted in a total greater than three.

Cross-tabulating the normal or preferred routes into the contents of output repositories against those social science respondents who have deposited data in a source repository at least once shows a wide range of approaches favoured (figures 44, 45 and 46). However, the most cited ways were through a known repository's URL and from an internet search engine (7 times each), which was also acknowledged across the three repositories which received data from social science respondents at least once (CERN, UK Data Archive and other repositories). These two main methods was followed by the use of an author's personal webpage and a library catalogue that links directly to the relevant article (5 instances each) and a publisher's online service and a specific journal's website (4 instances each).

Among the social science respondents who had not yet submitted data to a source repository, but intended to do so soon, there was a similarly broad spread of different techniques. However, the most cited was the use of a library catalogue that linked directly to the relevant article and an internet search engine (6 instances each), followed by a specific journal's website, an author's personal website and from a link in an email/CD-ROM/USB drive, etc (4 instances each).

Consequently, among the 19 depositors and soon-to-be depositors alike, more than two-thirds (13) claimed to use internet search engines; more than half (11) used library catalogues that linked to the article and just under half used an author's personal webpage (9) or a specific journal's webpage (8). This finding was not much different from the social science respondents as a whole, where 34 of the 61 in the sample claimed to make use of internet search engines and 24 and 23 made use of an authors' personal webpage and a specific journal's webpage respectively.

Figure 44: Respondents' preferred routes into output repositories against frequency of depositing data in CERN repository

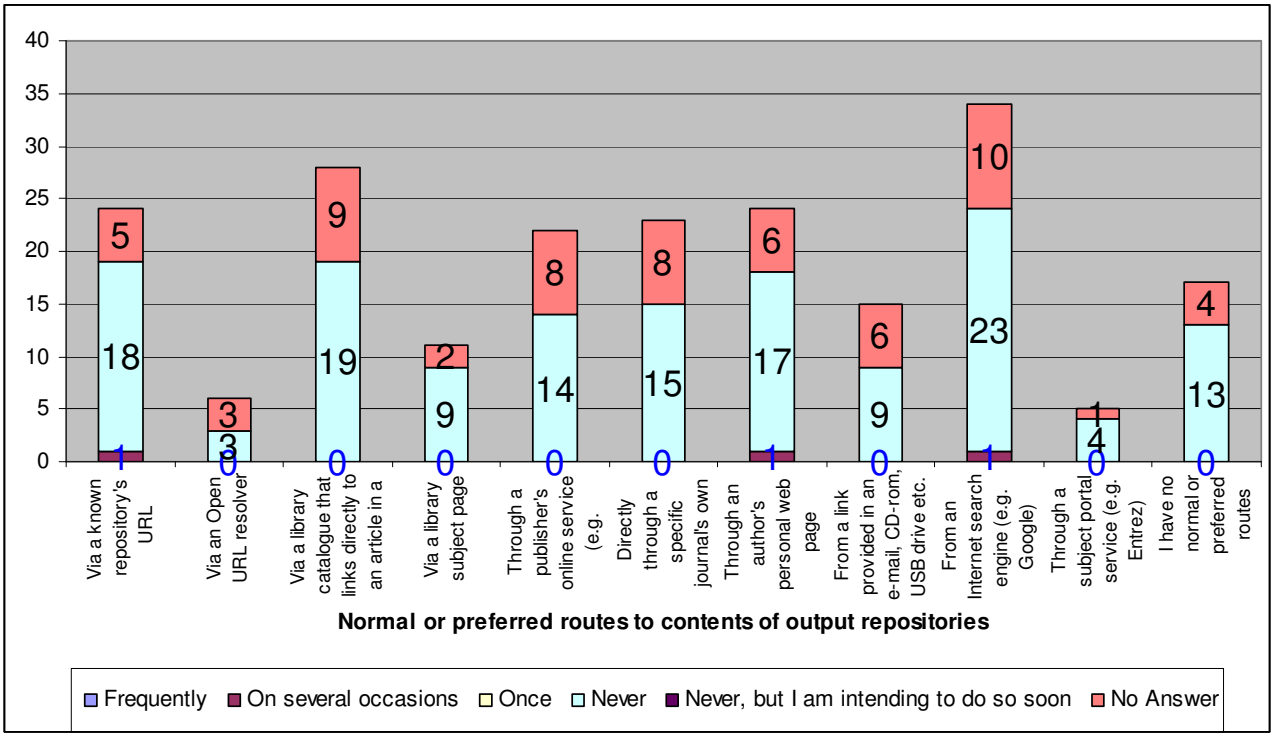


Figure 45: Respondents' preferred routes into output repositories against frequency of depositing data in UK Data Archive

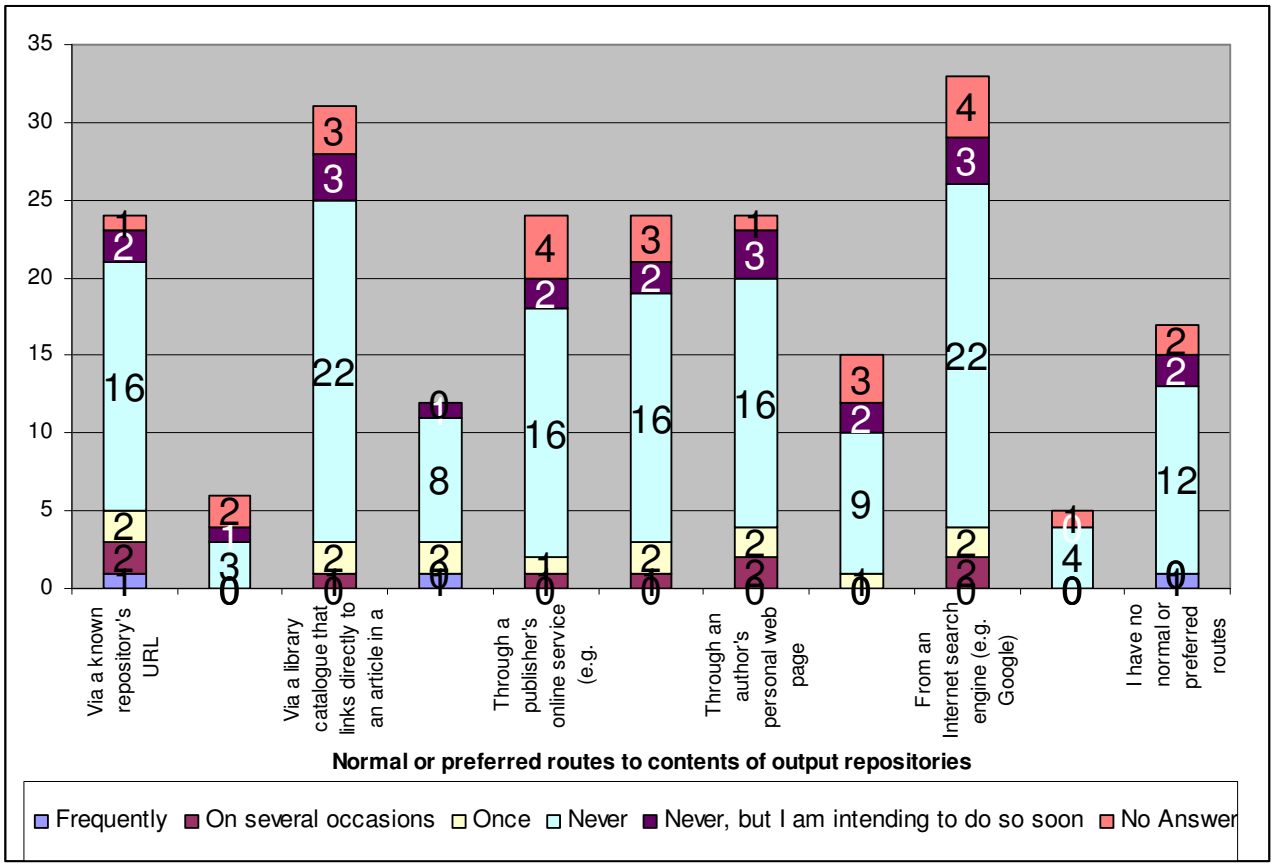
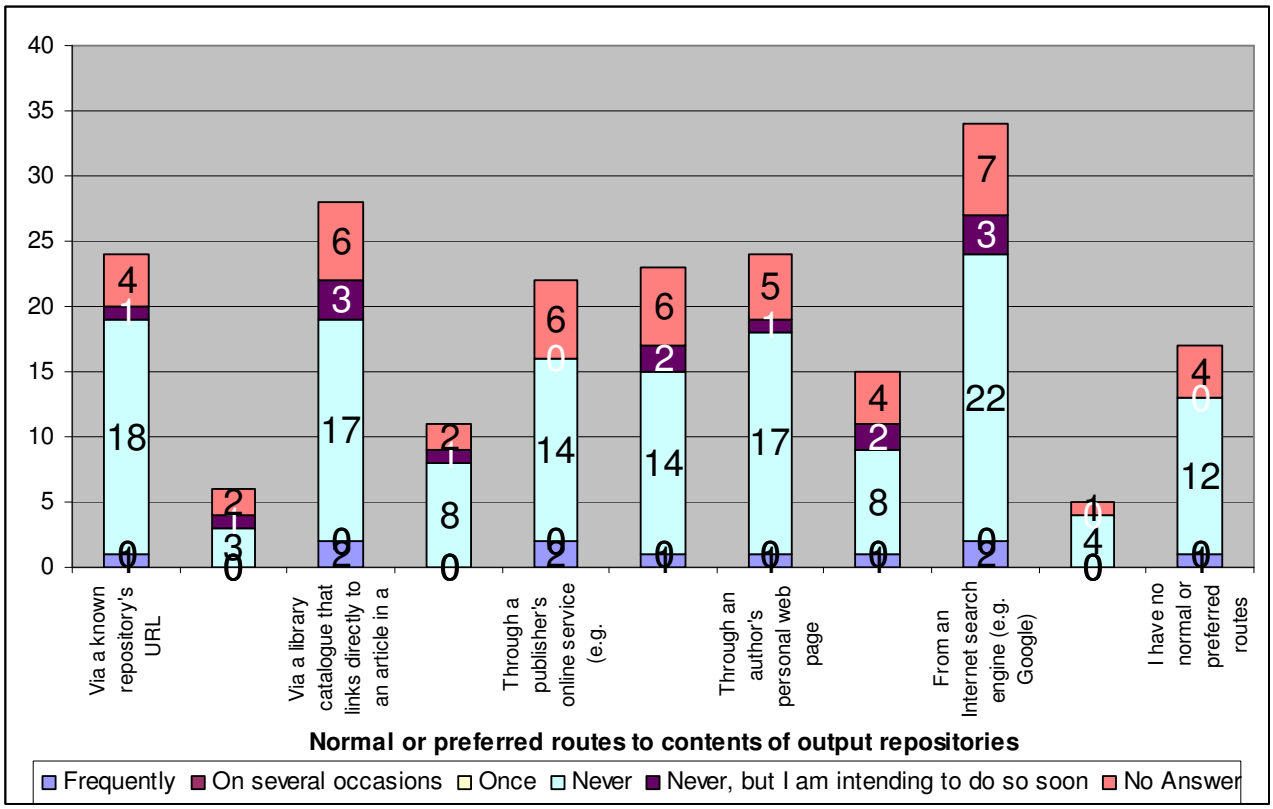


Figure 46: Respondents' preferred routes into output repositories against frequency of depositing data in other repositories

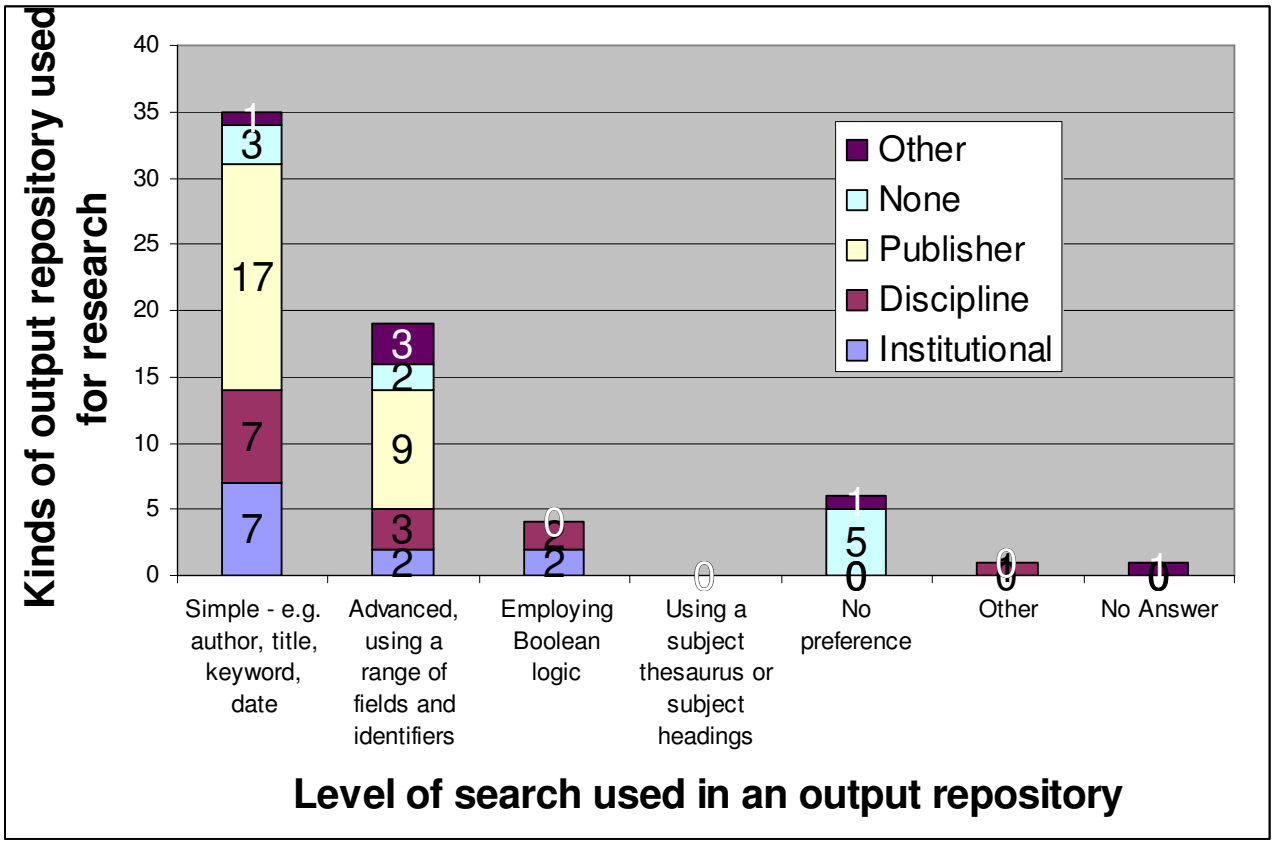


3. Relationships between respondents' use of output repositories and their search techniques

The most popular kind of search approach by social science respondents were simple ones, including authors, title, keywords and dates. This was cited in more than half of the 66 responses (35) and was regardless of the type of output repository used. Indeed, as figure 47 shows, of the 11 respondents who made use of institutional repositories, seven claimed to use a simple search technique, as did seven of the 13 who used discipline repositories and 17 of the 26 who used publisher repositories. Simple search was followed by advanced search across several fields and identifiers; again this was used by most of the respondents regardless of the type of output repository accessed.

As indicated by the analysis of the interviewees above, these approaches to search – simple and advanced – were the most mentioned among social science interviewees. In part this may arguably be due to a common theme across the interviewees and especially those working as academic staff: a lack of time. With such constraints researchers found themselves having to make use of what they could find; indeed, F, a contrasting researcher, spoke for many when she highlighted her method of search. In particular this involved using different search engines and typing in different combinations of key words and subjects to identify the most relevant literature for her chosen topic. This also had the added advantage of enabling her to take a 'sideways' look at the rest of the discipline, to gain a sense of what else had been written even if she did not ultimately use all the material.

Figure 47: Respondents' use of output repositories for research purposes against levels of search used



VIII. Support

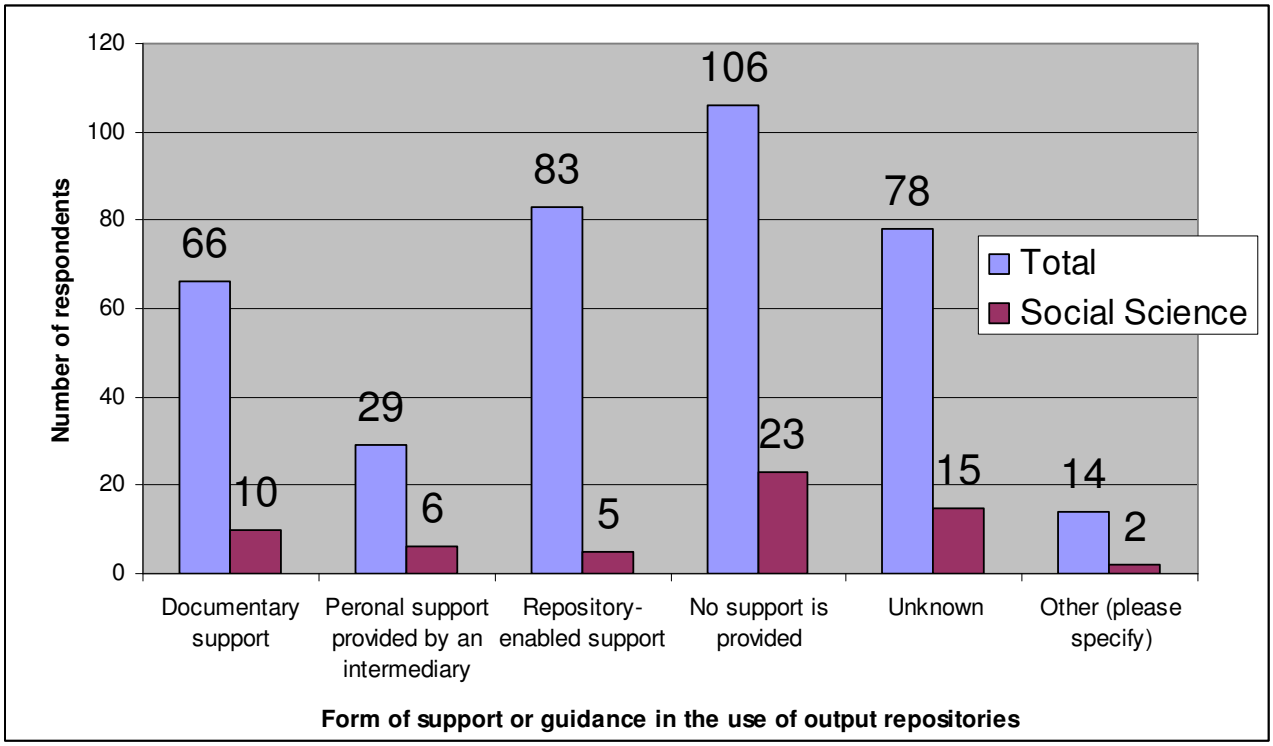
1. Respondents' perceptions of guidance and support

A third of all respondents (106) and those in social science (23) claim that they have not received any support or guidance with regard to using output repositories (figure 48). But while a quarter of all respondents (83) claim to use repository-enabled support, a fifth (78) said they did not know what support was available. These figures contrast to those for social science which show that almost a quarter of respondents did not know what support was available (15) followed by a fifth who used documentary support (10). The general attitude, therefore, was that support was there if needed, although researchers approached the question of research in a relatively independent and self-sufficient manner. Both the 'other' responses indicated this, with one respondent claiming that 'I have not looked for support but reckon it is available from my institution' and another saying that 'I expect that I would use repository enabled support, but I have not received any guidance so far because I do not use a repository'.

When compared to the interviewee responses, a general lack of knowledge or awareness of all forms of support or guidance was most apparent. Few had made much use of the services on offer; only B, an economic historian, and G, a sociology graduate student and research assistant, had done so, but these were geared towards the use of source data rather than output. B said the UK Data Archive provided good guidance on metadata and formats while G claimed that he had received help in using BOS software and tab datasets. Another postgraduate student, K, who studies development, claimed that she was aware of similar support and training for using data, but had not yet made use of it.

Among other interviewees, D, a social psychologist, said she responded to requests from the library but had very little contact with it otherwise. Similarly, L, a business and management lecturer, and M, a development researcher, reported no usage of support mechanisms that were available, while P, an economist claimed to deposit her own publications without the use of an intermediary or any training.

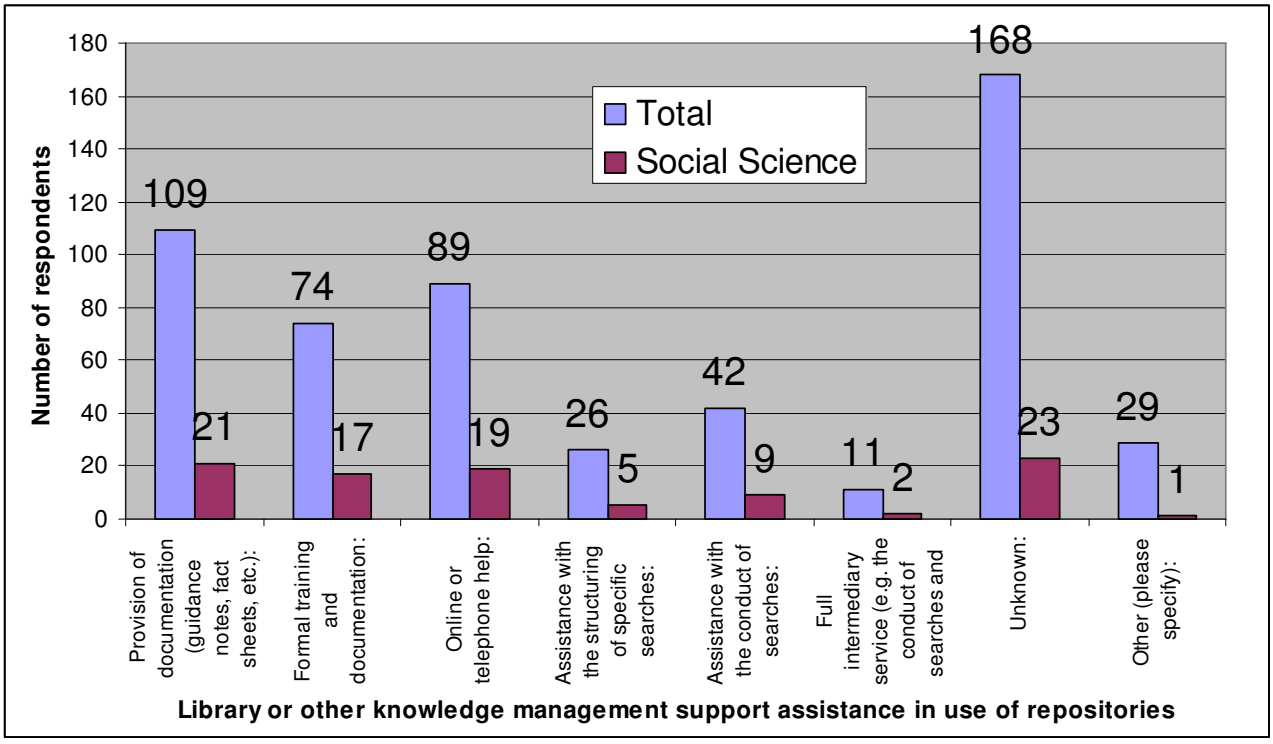
Figure 48: Respondents' perceptions of guidance and support for output repositories



The use of librarian or other knowledge management support for assistance in using repositories presented a mixed message among social science respondents (figure 49). This contrasted with the finding that nearly half of all respondents claimed not to have had any such assistance (168). In social

science the outlook was more mixed with no one response more evident than another. More than a third said they have received document-based support (21) or did not know what was available (23). Just under a third, claimed that they had had online or telephone help (19) or formal training (17).

Figure 49: Respondents' perceptions of librarian/knowledge management support for repositories



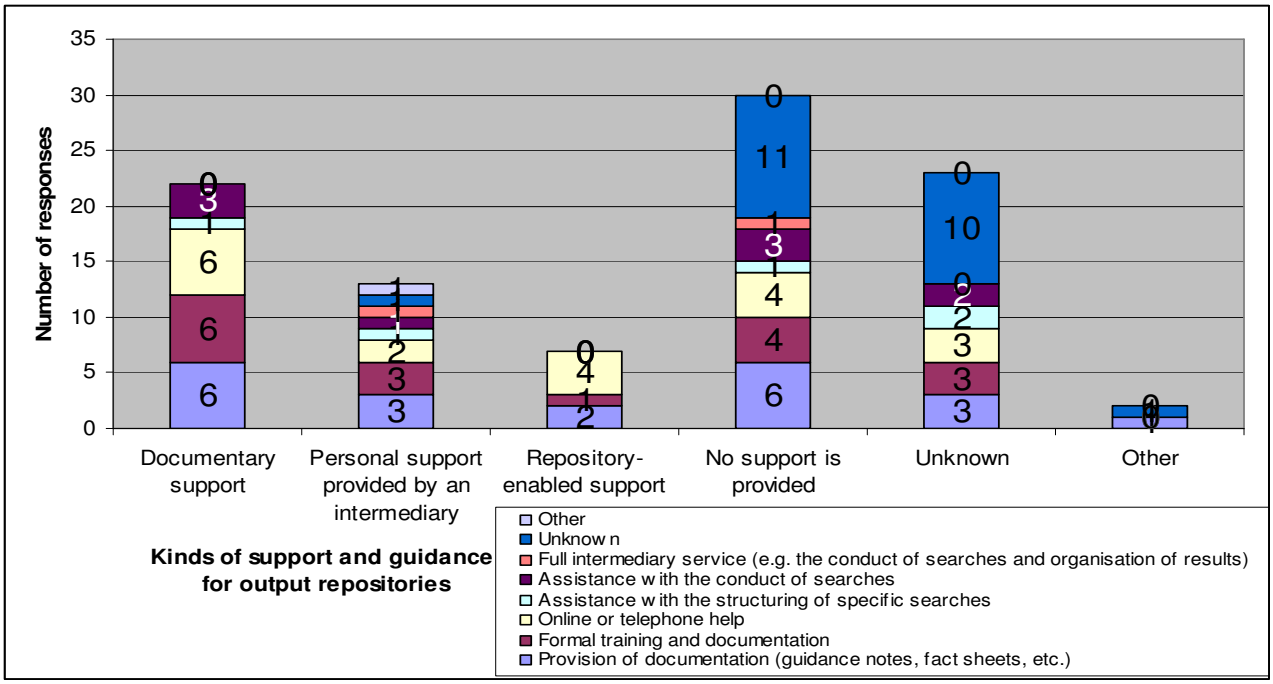
2. Respondents' perceptions of assistance for output repository use and assistance by librarian/knowledge management support

Social science respondents appeared to be relatively unaware of what assistance was available to them. A sizeable proportion of social science responses (53 out of 97) suggested that respondents claimed there was either no support available for output repository use or claimed not to know of it (86.9%) (figure 50). Within that figure, over half (30) were responses that there was no support or guidance in their use of output repositories. A third of this amount did not know what librarian or knowledge management support was available (11) while a fifth (6) claimed that documentary assistance was used. In addition 23 respondents claimed not to know what output repository assistance was available, with almost half of that number (10) similarly not knowing what librarian/knowledge management support was available for output repositories.

However, despite support being either not available or unknown for output repositories, respondents reported that what librarian or knowledge management support was available ranged across three main types: provision of documents (9), formal training and documentation, and online or telephone help (7 each).

These three main types of librarian/knowledge management support were similarly acknowledged my social science respondents who indicated awareness of institutional assistance for output repositories. The most cited type of output repository assistance was for documentary support (22). Within that figure respondents were split regarding the type of assistance that was provided by librarians or knowledge management support between the provision of documents, formal training and documentation and online or telephone help (6 each).

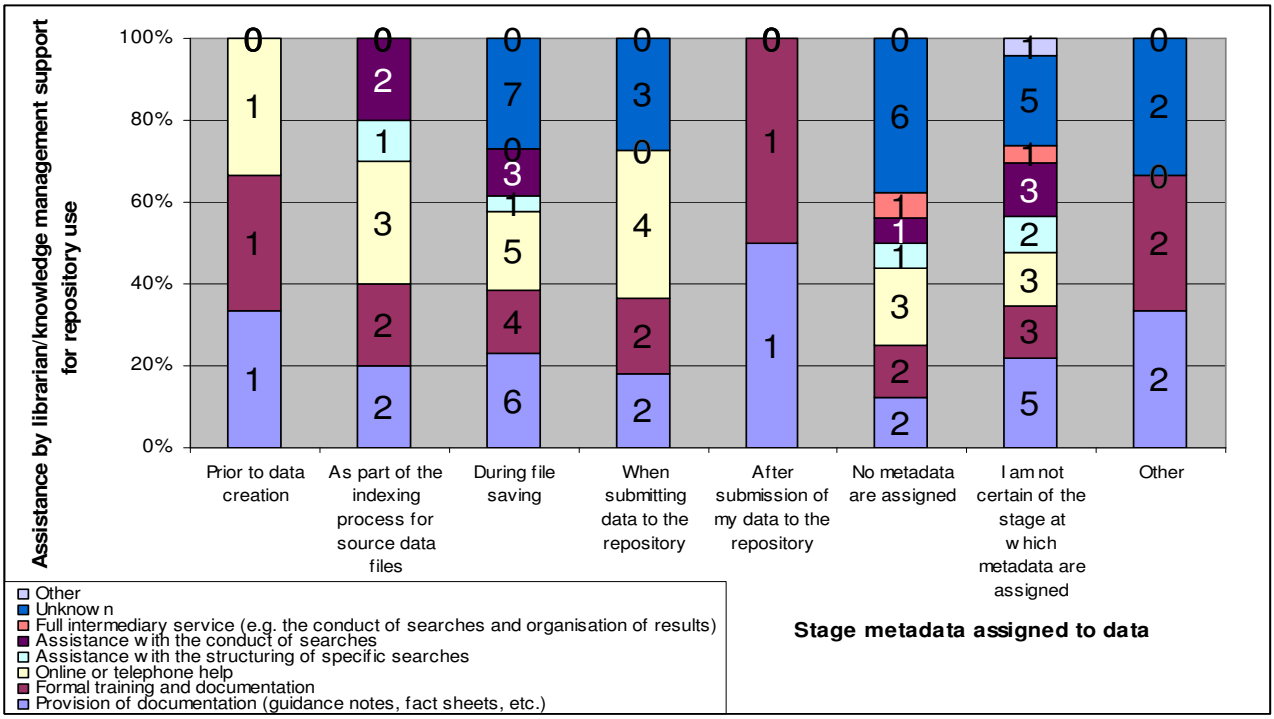
Figure 50: Respondents' perceptions of support and guidance for output repositories against librarian/knowledge management support



3. Relationships between guidance/support and metadata assignation

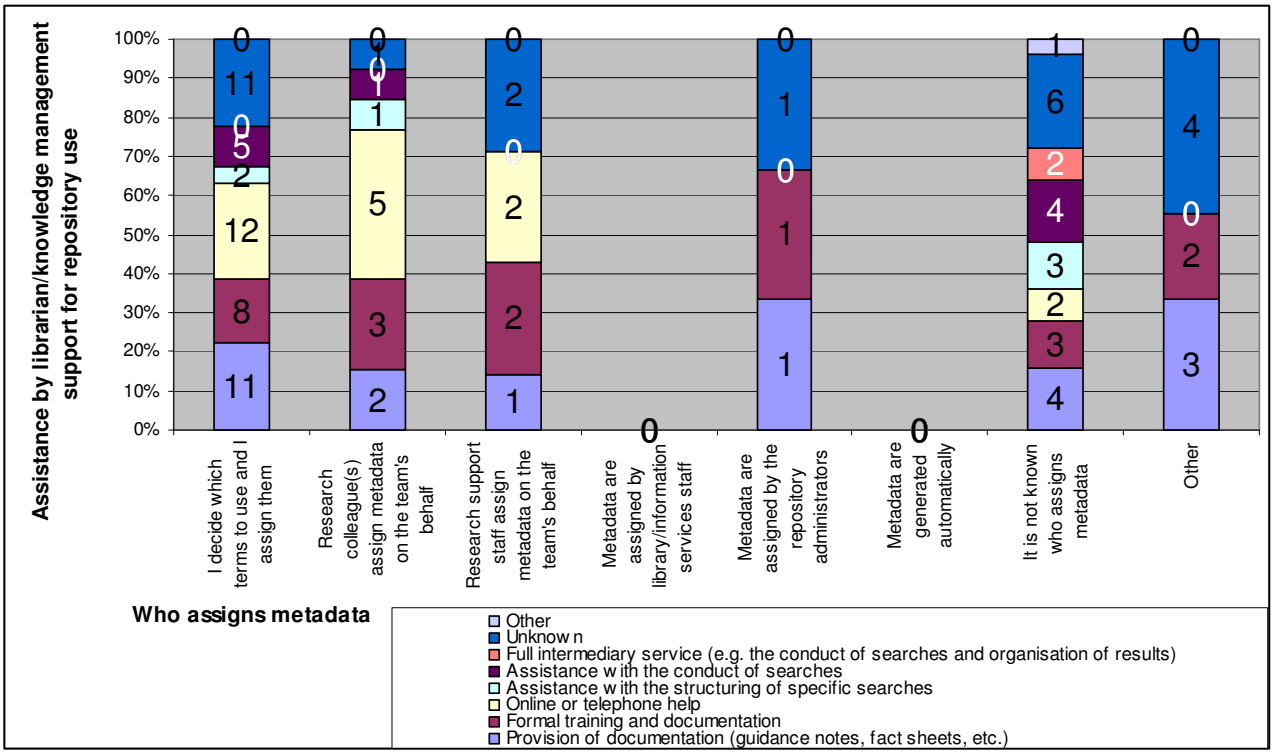
Social science researchers' usage of assistance and support can be analysed against their application of metadata to their own data (figures 51 and 52). When comparing the stage at which metadata is assigned and the kind of assistance used by respondents, it is not evident that there is much difference; rather, as figure 51 shows, respondents reported receiving the same proportionate kinds of support and guidance regardless of the stage at which metadata was assigned.

Figure 51: Metadata assignation against librarian/knowledge management support for repository use



Similarly, when the focus of analysis was shifted to consider the relationship between who assigned metadata and the kind of assistance utilised by social science respondents, it was also apparent that there was no clear relationship between one and the other. Respondents reported proportionally similar kinds of assistance regardless of who assigned data.

Figure 52: Responsibility for metadata assignation and librarian/knowledge management support for repository use



IX. Conclusions

The findings within this report indicate a generally high level of support and endorsement among the social science community for the aims of the STORE Project. This includes both the possibility of linking directly from a publication to the source data that it is based on and identifying what publications have been derived from a particular set of data.

However, these findings should be set against the way social science researchers appear to currently work. The impression given through the questionnaire and interviews is of a process that on the surface appears relatively organic, almost ad hoc. The way they use repositories is but one aspect of their work patterns, which sit alongside other aspects including their use of search engines to locate others' research and the production and storage of their own data. Indeed, social science researchers report certain functional limitations with both source and output repositories, including limited access, a sense of incompleteness, difficulty in navigation and a concern with the end-use of the material that is deposited in them. These themes suggest that repositories on their own may never replace the various ways in which they conduct their work, but they may be improved.

For the aims of the STORE Project to be achieved, it will be necessary for any future work to be successful to acknowledge the thoughts, comments and work practices of social science researchers. In particular these may be understood in three main areas: their current method of work; the distinction between ideal and actual practices; and differences within the social science research community.

1. Current ways of working

As noted above, social science researchers arguably work in an organic or ad hoc fashion. Social science researchers give the impression of being largely self-sufficient with regard to the production of data and its storage. Regardless of the different kinds of data that they produce, it was apparent that many take the decision of what metadata to assign to it as well as keeping it on a range of systems that are generally not open-access, from personal computers to networks based at their institutions. If they choose to make their data available it tends to be through personal channels, by deciding on the merits of a direct request to do so. This means they are able to use email to send their data or other portable media, such as CD-ROMs.

This independence from source repositories is reflected both in the few numbers of respondents who claimed to have submitted research data to them. It is also arguably indicated by social science researchers' relative lack of awareness of support and guidance that is available to them, particularly with regard to metadata assignation. By contrast, even if they did not use it, social science researchers did admit to being aware of librarian or other knowledge management support that could be accessed if needed.

Social science researchers' independence may also be seen in the way that they approach search tools during the course of their work. Without distinguishing very greatly between the different kinds of output repositories that they used, the interviews commonly indicated that social science researchers used different techniques and systems to find what they needed. This included using several different search engines and trying different combinations of terms, from title to author to date, for example. While the level of these searchers did not rise much above the simple, the interviewees reported generally being able to find what they needed. And if these processes did not elicit the desired results, they would opt for word of mouth, including contacting other researchers for their publications or research data.

2. Producing and consuming at the same time

The social science research community reported different approaches and attitudes to using others' research data and others using their own. In principle virtually all believed in the idea of sharing research, although in practice this could vary. In particular many were concerned that they would not reap the benefits of the data they had produced if it had to be made publicly available at the earliest stage possible. This attitude was evident even among the most enthusiastic researchers for data

sharing. Indeed, one argued that a 'patent' should exist whereby the producer of the data should have restricted rights to use his or her data before publishing it. Another commented on the tension between being a 'producer' and a 'consumer' of data at the same time; while making data more widely available was ideal for the consumer, this was less so for the data's producer.

However, not all social science researchers report using others' research – or at least not the source data that much of it based on. First, various pressures and constraints including a lack of time, teaching commitments and other responsibilities – means that they are unable to cross-reference colleagues' work. Second, using 'raw' data that has not been processed can make meaningful use of such material difficult. Third, they are too busy producing their own research data to be able to scrutinise or use another's.

How then to account for these contrasting views between researchers' belief in sharing research, their reservations to share data and their non-use of others' data? Arguably this may be explained by looking at the type of material that researchers actually use rather than claim to value. In particular they make use of search techniques not to find data in general, but publications related to their topic. By doing so they would both gain a greater appreciation of their discipline, but also ensure that they were on the right track in their own work. Such research tends to be more 'processed' than the 'raw' data that one would expect to find in a source repository; it is processed either as a finished article (or a working paper) or as data that has been contextualised.

3. Same community, different needs

The social science research community is not a homogenous group. This was reflected in attitudes towards the project's aims, which showed some variation between the two main constituencies, academic staff and postgraduate students. But it is not just in attitudes that these two groups differ; it is also highlighted through the interviews in the work patterns that each undertakes. In particular academic staff were less likely to make use of others' research in their own work than would postgraduates. This included not only publications but also the use of others' data.

The difference may be explained by the extent to which one section – the academic staff – are already 'established' in their particular field with tried and tested set patterns of work. By contrast the postgraduate community presents a more aspirant role; students not only seek out a method of research that works best for them, but also are engaged in a process of building knowledge in their chosen discipline. This latter aim requires a familiarity with previous researchers and their work, much of which may be located electronically and in different kinds of repositories.

Such differences within the social science research community suggests therefore that the development of source-to-output repositories would not necessarily benefit all researchers in equal measure. Indeed, as the interviews brought out, 'established' researchers saw source-to-output repositories as of less benefit to themselves than to aspiring researchers whose patterns of work had not yet been set. The interview phase presented a common academic staff view that such a medium would be as a training tool for postgraduate students. In particular it was suggested that it could enable postgraduate students to run particular analyses used by another researcher and see if they were able to replicate the findings. By contrast, among the social science postgraduate students themselves, there was less consensus regarding which group of researchers would most benefit from source-to-output repositories.

However, this does not mean that all future work in the STORE Project should be targeted solely at postgraduate students. Despite academic staff claiming not to have time to scrutinise others' data, several did concede that the project's aims could be used in peer-review processes, such as refereeing journal articles. In addition the development of a new functionality could give rise to new kinds of work within the research community, including analysis of research methods and publications based on the effects of this available research and data.

JISC Digital Repositories Programme

Scenario Collection Form 1

Title	**Source-to-Output and Output-to-Source**
Author	**Guy Burton**
Project	**StORe**
ID	
Date	**26 July 2006**
Narrative	Dave is an economic historian. To test a hypothesis which he has received ESRC funding he accesses economic history data from the UK Data Archives or the National Archives at Kew. The data is available in various formats, but Dave tailors it to fit into Excel. Dave labels the data by column titles. He conducts analysis of the data and drafts an article for an economic history journal. He submits the article to the economic history journal with the processed datasets supporting the findings (e.g. tables graphs). The article is published and following funding conditions, Dave submits the datasets to the repository he used to gather the data.
Other information	Dave finds the following problems with finding data from repositories: 1. *Oversight* – Dave and a colleague go through a list of the available datasets from a repository but fail to find the one they need given its sheer volume 2. *Unintuitive* – Dave finds the search facility in repositories unintuitive. He is unable to identify researchers by name (e.g. typing in 'craft' generates data associated with 'arts and craft' rather than 'Nicholas Craft') or date (e.g. a search for data relating to '1953' may be overlooked by a dataset titled '1950-55'). He is also unable to access previous searches done by himself. 3. *Obsolescence* – Dave finds that repositories exclude certain types of data, owing to the obsolete format that they are stored in (e.g. 8" floppy disks). This may mean greater costs by either the owner of the data or the repository to make it readable or a decision not to provide it. 4. *Associated Work* – Dave finds it difficult to see how data retrieved from repositories has been used. The list of datasets do not provide details of how they have been used. 5. *Accreditation* – Dave finds it frustrating to have to type his username and password every time he used a source repository. He would find it useful if the repository's software/middleware would recognise a researcher.

	6. *Subsequent Publication* – Dave places much of his teaching material on his password-protected website (university-based output repository), along with his working papers. However, he does not know how much of it he is allowed to make available, owing to funding conditions.

JISC Digital Repositories Programme

Scenario Collection Form 2

Title	**Source-to-Output and Output-to-Source**
Author	**Guy Burton**
Project	**StORe**
ID	
Date	**26 July 2006**
Narrative	Anita works as a contracting researcher for a university research unit concerned with social policy. She decides upon a topic of research, including a hypothesis to test or series of questions to answer. Anita begins by undertaking a literature review and tracking down 4-5 authors in a given topic (identified through catalogues she receives of new publications or details of conferences in her field). Anita checks out the references cited by these key authors and conducts an online search of publisher repositories (i.e. BIDS) for the abstracts listed in the references. She assesses the value of each abstract. For those considered relevant, Anita either accesses the article electronically, thereby giving her a sense of what is currently being published in the field. Anita conducts her own research in two ways: one, by fieldwork involving interviews with contacts and field notes of the community or area that she is based; two, through statistical analysis of data collected online. Anita collects her online data from official government sources, including the National Statistics Office, the Census website, Nomis and local authority websites. She stores the data produced in two main formats. First, the field notes and interviews in word documents on her computer, using the following metadata: policy area (e.g. health, crime, welfare), area (neighbourhood), time of interview and type of interviewee (e.g. resident, workers, community activist). Second, she bookmarks the accessed websites on her computer to facilitate return visits. Anita runs the analysis of these different data types and drafts a working paper based on the findings. The working paper is disseminated through conferences or an output repository and subsequently prepared for publication, via the research unit.
Other information	Anita observes the following issues with regard to her accessing and processing of data: 7. *Availability of Data* – Anita uses various source repositories which are subject to change by the organisations within which they operate. This may mean their dissolution or shift away from a particular URL, thereby invalidating her bookmarks (e.g. the breaking up of the ODPM into two different ministries with different websites and addresses may

	have implications on the location of social housing data). 8. *Awareness of Repositories* – Anita observed a lack of knowledge about particular source repositories, citing word of mouth as the main way of identifying relevant ones. Anita noted little knowledge of the UK Data Archive compared to Nomis. 9. *Data Sharing* – Anita favoured sharing of data generated from working papers and publications. However, she noted two main concerns with this: one, making it available prior to publication of her findings; two, lack of explanation in the deposited raw data for subsequent researchers (which might be resolved through the production and availability of a brief summary of core data/findings associated with the raw data to enable researchers to make sense of it).

JISC Digital Repositories Programme

Scenario Collection Form 3

Title	**Source and Output**
Author	**Guy Burton**
Project	**StORe**
ID	
Date	**26 July 2006**
Narrative	Linda works for a repository that is both source and output, dealing with social attitudes. The repository includes both primary data (through a contracted out quantitative survey) and publications relating to social attitudes. Linda receives primary data from depositors and 'cleans' it, making it accessible for researchers before placing it in the repository. The data is laid out in SPSS and table format (including cross-tabulations) and metadata is assigned to it in an informal fashion, with little consideration given to future requirements, needs or challenges. As well as data, Linda receives publications that are associated with the repository's field of interest from depositors. Like, the data, this is placed in the repository as open-access, with no requirement for licensing use or password protection prompts. Linda and her colleagues are able to see which datasets are most popular by counting the number of 'hits' they receive. Linda and her colleagues ensure that funders receive details of data three months in advance of its availability in the repository.
Other information	Linda noted the following observations regarding source and output repositories in relation to her work: 1. *Knowledge of Future Use* – Linda observed that the open-access format of the repository (i.e. no user registration or password requirement) made it difficult to identify the use of primary data by researchers accessing the repository. Despite discussions with the UK Data Archive, no agreed protocol to resolve this issue had been achieved. 2. *Open-access* – Depositors and funders were made aware of the repository's open-access status prior to depositing of data and publications. 3. *Delayed Depositing* – Linda noted that there could be substantial periods of time between the production of an output (i.e. publication) and making the dataset it was based on available in the repository.

JISC Digital Repositories Programme

Scenario Collection Form 4

Title	**Non-Repository User**
Author	**Guy Burton**
Project	**StORe**
ID	
Date	**26 July 2006**
Narrative	Elaine is a lecturer in business and management. She and her colleagues in the research team identify an area of research. She identifies what has been previously published by looking through the search results for publications within the same field. This she does by using the Ingenta, Emerald and Econlit search engines, typing in various keywords. Although she finds differences between these search engines she does not have any problems gaining a good overview of a topic since the number of journals relating to her field of interest is relatively small. If she is on campus she uses the links in the search engine to go directly to the articles she wishes to look at. Elaine and her research team carry out original research based on quantitative and qualitative data that she and her research team carry out through questionnaires. The quantitative data produced by the research team is stored in SPSS format while the qualitative findings are stored in Word documents. Each member of the research team takes individual responsibility for assigning a name to the dataset, which is stored on the university server (and therefore accessible by all in the team). Elaine and her research team use the data they produce to draft working papers and articles. Following the publication of her articles, she submits all her publications (including working and conference papers) to the administrative staff in her school/faculty for use in output repositories. Elaine is not sure how this material is subsequently used.
Other information	Elaine noted the following points about her use of data: 1. *Sharing data* – Elaineand her research team do not make their data available to others since it is bespoke and new. She would respond to individual requests from researchers interested in her data. However, in principle she is not averse to sharing her data and would do so if it was made obligatory. She favours keeping hold of data until she has finished analysing it before depositing it in a repository. 2. *Metadata Guidance* – Elaine claims not to have received any guidance with regard to assigning metadata. 3. *Controlling Access to Data* – Elaine believes that using the university server helps maintain control over data and its circulation. She says

	that sending data by email to those who request it is another way of avoiding indiscriminate dissemination.

JISC Digital Repositories Programme

Scenario Collection Form 5

Title	**Source and Output Repositories**
Author	**Guy Burton**
Project	**StORe**
ID	
Date	**26 July 2006**
Narrative	Philip is a development researcher at a think tank. He poses questions that he then tests with research. Much of these questions are derived from commissioned work, which is not long-term. Philip's research primarily include both the use of primary data and meta-analysis of past publications (occasionally he and his institution produce quantitative datasets in Excel or SPSS format for their own use). For primary data he access datasets at the FAO and World Bank, either on-line or through the associated CD-ROM that his institution subscribes to. For other publications he uses IBSS, Google Scholar and the online journal subscriptions that his institution has, along with five personal journal subscriptions. In addition Philip receives email updates from particular bodies and institutions he is associated with, including both datasets and publications that are newly available. Philip analyses the data and publications he accesses to produce working papers and publications for his institution. These are made available online without restriction.
Other information	**Issues** Philip observes the following issues with regard to data: 1. *Variables* – Philip finds the number of variables and cases in some datasets too large, which can create problems in deciding which ones to use and the constraints posed by time in analysing them. 2. *Ownership* – Philip observed that dataset producers (including himself) could be possessive about sharing, especially if the cost of its production is not carried by an external user. 3. *Accessing Data* – Philip noted that if he wanted to access another researcher's data that was not readily available but he knew it existed, he would approach him or her directly. He does not do so generally for a lack of time to effectively analyse such material. 4. *Publications* – Philip observed that IBSS did not cover all journals and was mainly concerned with Anglo-American publications;

	consequently, it missed a lot of non-US/UK-related material. He also noted that Google Scholar picked up publications beyond journals, including conference and working papers.