UNIVERSITY OF BIRMINGHAM

Project StORe (Source to Output Repositories)

Work Package 2: Survey of Researcher Use of Repositories

Survey Report Part 2: Physics Report

Stephen Bull

Project StORe Researcher

Information Services

University of Birmingham

August 2006

Abstract

Results are presented on the Physics Survey of Researcher Use of Repositories which constitutes the culmination of Work Package 2 (in Physics) of Project StORe (Source to Output Repositories). The data were obtained by Project StORe from an online questionnaire and individual interviews during the period March 2006 to June 2006. A comprehensive study of the project's principal aim of linking source to output repositories (and vice versa) is given along with a detailed study of the associated topics of source data, source repositories, metadata, data access and sharing, output repositories and support.

A brief overview of Project StORe, the Physics user survey and a summary of significant observations from this survey are given. The in-depth results and commentaries from both the Physics questionnaire and interviews are detailed. A brief summary identifying consistent messages and potential follow-up actions is given.

Acknowledgements

Thank you to everyone who has helped, in whatever way, with this project and in particular all those who responded to the questionnaire and who participated in an interview.

List of Contents

List of Figures

List of Tables

1 Introduction

1.1 Project StORe

Project StORe (Source to Output Repositories) [1][1] is part of the JISC Digital Repositories Programme [2]. The principal aim of Project StORe is to add significant value to the output repositories of research publications by enabling them to interact with source repositories of primary research data (and vice versa). In support of this, Project StORe aims to address and learn more about researcher behaviour and attitudes on the themes of source data and source repositories, metadata, data access and sharing, output repositories and support. The project is multi-disciplinary in scope and covers seven scientific disciplines: archaeology, astronomy, biochemistry, biosciences, chemistry, physics and social sciences.

User surveys, in the form of a questionnaire and researcher interviews, have been conducted in all of the seven scientific disciplines to determine required functionality in both source and output repositories and to make them more useful to researchers when using primary data in source repositories and at the point of submitting to or downloading papers from output repositories. A business analysis of the survey results will establish general principles for middleware to link source and output repositories with a pilot demonstrator developed and tested in one of the subject domains. Finally, a full and extensive evaluation of the project will be carried out in order to inform the JISC [3] of the best options for future development in this area.

In this report, the details and results of the Physics user survey are presented and discussed.

1.2 Physics User Survey

As with the other six scientific disciplines covered by this project, the Physics user survey consisted of a questionnaire followed up by researcher interviews. The online questionnaire was developed for use with all seven disciplines and covered all of the major themes of the project. The questionnaire aimed to introduce the project to researchers, gain an overview of how researchers manage and use repositories as well as to identify researchers willing to take part in an interview. The questionnaire was open for participation between Monday 13th March 2006 and Friday 21st April 2006. Invitations to the Physics communities at 34 UK universities were made via e-mail, using addresses obtained from university websites, through Heads of Research Groups (25/34 universities), Heads of Departments (3/34 universities), Research Group Secretaries (2/34 universities) and through individual researchers (4/34 universities). Recipients in the first three groups were encouraged to circulate the invitation to all members of their group or department. It is therefore hard to assess exactly how many people received the request, but it is known that 568 people were sent an e-mail invitation.

The interview phase of the survey immediately followed the questionnaire, with the interviews aiming to develop the themes explored in the questionnaire as well as to address issues that had arisen from the questionnaire and required further exploration. In particular, the interviews were used to gain a much more detailed view of how researchers use and manage source and output repositories, as well as to identify functional enhancements to both types of repository that researchers would like to see implemented. An interview script of 56 questions was developed (given in full in appendix A). A structured script was necessary to allow comparison of results from the interviews, although it enabled flexibility (including the omission of questions) depending on how individual interviews developed and the time constraints of the

i.

[1] This and future numbers given in [ ] refer to items in the bibliography at the end of this report.

interviewee. Individual e-mail invitations inviting participation in an interview were sent out to the 20 Physics researchers who had indicated in the questionnaire that they would be happy to participate in an interview. Interviews, both face-to-face and via telephone, were arranged at times that were mutually convenient to both interviewee and interviewer. The interviews took place between Monday 22nd May 2006 and Monday 12th June 2006. They lasted between ½ and 1 hour, with most interviews lasting 40-45 minutes. Where possible, the interviews were recorded and the main points transcribed at a later date.

The results from the questionnaire and interview phases of this project are detailed in the remainder of this report. In the following chapter, a summary of the significant observations from both the questionnaire and interviews is given. Chapters 3 and 4 present in detail the significant observations from the questionnaire (chapter 3) and interviews (chapter 4). Chapter 5 contains additional observations from both questionnaire and interviews. In chapter 6 a brief summary along with some recommendations for future work is given. Three appendices are given. The first (appendix A), as already mentioned, provides a full copy of the interview script, the second (appendix B) details a historical description of the three nominated repositories for Physics adopted by this project [1] whilst the third (appendix C) presents a number of scenarios and a use case that attempt to depict some parts of the Physics research process.

2 Summary of Significant Observations

In this chapter, a summary of the significant observations from both the Physics questionnaire and Physics interviews are presented. For a more complete discussion of all the points raised in this chapter, readers are directed to chapters 3, 4 and 5.

2.1 Identities

A total of 63 Physics researchers responded to the questionnaire. The researchers were from 19 Universities across the UK and consisted of academic staff, postgraduate students and research assistants / fellows. A wide-range of Physics fields of interests was represented.

Of the 63 questionnaire respondents, 13 agreed to participate in an interview. This smaller group of researchers were from 5 Universities across the UK and again consisted of academic staff, postgraduates and research assistants / fellows. 6 different Physics fields of interest were represented.

2.2 Project Aims

The principal aim of Project StORe (to enable direct links from source to output repositories and vice versa) was well received: 60% of questionnaire respondents thought that the source to output linkage would be either of 'significant advantage to my work' or 'useful but not of major significance' whilst 67% of respondents selected these two options for output to source linkage.

The support for the linkage that emerged from the questionnaire was largely supported in the interviews, where again researchers were slightly more positive towards the output to source (over that of the source to output) linkage. The specific example of this linkage that was particularly well received was that of being able to obtain the source data that makes up figures presented in publications. Many felt the linkage could also be an advantage with the processes of data comparison and understanding the work of others' presented in publications. The main objections to the linkage were often practical in nature, with perhaps the biggest concern being the level (or stage) of source data to which it would be useful to link to. Many felt that the source data that may be the most useful to link to is the final Physics results produced towards the end of a particular analysis and that in most cases linking to the 'raw' or 'unprocessed' data would be of little use to others.

2.3 Source Data

Physics researchers produce a wide variety of electronic source data which they hold in a variety of formats. Whilst in many cases researchers will store their data in well known formats and analyse their data using 'standard' software, it is not uncommon for Physicists to store their data in less well known formats and write their own analysis software: this is particularly true in the case of High Energy Physics[2]. It follows, that Physics researchers will often generate data in a combination of different data formats. There is a huge range in the size of Physics source data, with final Physics results being stored in files as small as kilobytes (10^3 bytes), whereas raw data can be as large as petabytes (10^{15} bytes).

Many Physics researchers do not access others' research data. Those that do are most likely to do so to improve the quality of their own research, for example for cross-checking and comparing results.

i. _____

[2] Also known as 'Particle Physics' and 'High Energy Particle Physics'.

2.4 Source Repositories

Many physics researchers do not use source repositories: the notable exception being High Energy Physics, where their use is the norm – although it should be noted that these are mostly private source repositories where access is restricted by collaboration and laboratory membership. Of the High Energy Physics source repositories that were mentioned, CERN [4] was the most popular. The questionnaire restricted the discussion to the source repositories to which researchers submit data; however from the interviews it became clear that their usage is slightly larger when also considering the researchers who extract data from this type of repository. There is a large variation in the frequency of use of source repositories. Users of source repositories seem fairly happy, although better documentation is an improvement a number would like to see.

2.5 Metadata

Metadata is most popularly assigned to Physics source data 'during file saving' and 'as part of the indexing process of source files'. It is most commonly decided and assigned by the researchers themselves or it is done automatically. The types of metadata which researchers feel it is important to assign to their data consist of a number of generic terms (9 out of the 11 that were suggested in the questionnaire were considered important) and a number of terms specific to the Physics field of interest. It was clear from the interviews that the types of metadata that should be assigned to Physics data may vary depending on the level or stage of the analysis.

2.6 Data Access and Sharing

Physics researchers employ a variety of methods to make their research data available, although nearly a third of questionnaire respondents said 'I undertake no measures to make my research data available'. Questionnaire respondents also cited a range of factors that would encourage or discourage them from sharing data. Many researchers would be encouraged to share their data to enable collaboration, benefit the research community and raise their own research profile whilst premature broadcast of results and threat of loss of ownership are key reasons that would discourage them. The time taken to enable data sharing is also a concern to many. On the issue of the formal restrictions that researchers apply to their data, many fell into one of two groups: those who apply maximal ('restricted to immediate research team / programmer members') and those who apply minimal ('no formal restrictions') access restrictions. A number of methods to control access to data are used, although 'storage of data on a private network / intranet' is easily the most popular.

In the interviews it became apparent that, although in principle many researchers are not against the idea of sharing data, there are many practical issues (particularly with regard to raw data) which deter them.

2.7 Output Repositories

The vast majority of Physicists make use of output repositories for their research, although the numbers using them for teaching is much less. All three types of repository: publisher, discipline and institutional were cited by questionnaire respondents as being well used, although it became clear from the interviews that there is some ambiguity in the understanding of the term 'institutional repository' and that the use of this type of output repository is probably less than the results from the questionnaire suggest. Publisher repositories are the most popular type of output repository in which Physics researchers deposit their research publications. The choice of repository is generally dictated by the relevance of the article to the output repository's coverage as well as the impact factor of the output repository. The issue of open access was discussed during the course of the interviews, where it was found that researchers were, in principle, supportive although before submitting their own work to such a repository (for example an institutional

repository) they would require a number of preconditions to be met. The biggest concern about this type of repository was that there should be appropriate peer-reviewing which to most is regarded as essential.

Physics researchers use a number of routes to access output repositories, although 'via a known repository's URL' and 'from an internet search engine' are the two methods most used. When searching for material in output repositories, researchers show a preference for using a simple level of searching. Many expressed the importance of an excellent search facility.

2.8 Support

Many Physics researchers appear self-sufficient when assistance is required with their use of repositories, with nearly one third of questionnaire respondents having used 'no support'. Of those who do receive support, repository-enabled support is the most popular. Where assistance is provided by librarians or other knowledge management support, the provision of documentation along with online or telephone help are the most popular services. There exists a clear lack of awareness of what assistance is available from librarians and other knowledge management support by a significant proportion of the Physics community.

3 Significant Observations from the Questionnaire

Detailed in this chapter are the results from the Physics questionnaire and accompanied by a commentary. Answers to the five 'free text response' questions that the questionnaire asked are given in chapter 5. A full version of the questionnaire along with the multi-discipline results of the whole survey can be found in Appendix A of Part 1 - Cross-discipline report. A number of the questionnaire questions allowed researchers to select more than one answer: where this is the case, care has been taken to specify the relation of any quoted percentages.

3.1 Identities

Of the 377 responses received, 63 were from Physicists, representing 16.7% of the questionnaire population. Using the 568 Physics researchers who were sent an e-mail, the response rate to the questionnaire in Physics was 11.1%. Whilst this may be seen as a somewhat disappointing result from the high numbers of e-mails that were sent out in Physics, the response is comparable with the other disciplines covered by this project as shown in Table 3.1.

Table 3.1: Questionnaire responses per discipline

Discipline	Number of Responses	Percentage of Questionnaire Population (%)
Archaeology	64	17.0
Astronomy	64	17.0
Biochemistry	46	12.2
Biosciences	40	10.6
Chemistry	38	10.0
Physics	63	16.7
Social Sciences	61	16.2
Non-attributable	1	0.3
Totals	**377**	**100.0**

Respondents were from 19 universities distributed across the UK. Table 3.2 shows the role of the researchers who answered the questionnaire: no undergraduates, contract researchers or independent researchers were contacted. The majority of the respondents were academic staff (51%) and postgraduates (33%), followed by research assistants / fellows (13%). The remaining 'other' researchers were a 'computer officer' and a 'research scientist in a laboratory'.

Table 3.2: Role of respondents to the questionnaire

Role	Number of Respondents	Percentage (%)
Academic Staff	32	50.8
Research Assistant / Fellow	8	12.7
Postgraduate	21	33.3
Undergraduate	0	0.0
Contract Researcher	0	0.0
Independent Researcher	0	0.0
Other	2	3.2
Total	**63**	**100.0**

The physics fields of interest of the respondents were from right across the physics spectrum including: Atomic Physics, Computational Fluid Dynamics, Condensed Matter, Grid Computing, High Energy Physics,

Lasers, Nanoscale Physics, Nuclear Physics, Optics, Organic Electronics, Plasma Physics, Semiconductor Physics, Solid State Physics, Surface Physics, Theory and Transmission Electron Microscopy.

3.2 Project Aims

The questionnaire asked two questions which addressed the projects principal aim of linking source to output repositories and vice versa. The first of these questions asked: 'Source repositories contain primary research data. If a standard feature of such repositories was the ability to identify and link to the publications that had been developed from these data, how advantageous would you find it?' The second question asked 'How advantageous to you would it be if it were possible to go directly from within an online publication (electronic journal article or other text) to the primary source data from which that publication was developed?' The results to these questions are shown in Figure 3.1 where 'source to output' represents the first question, 'output to source' the second.

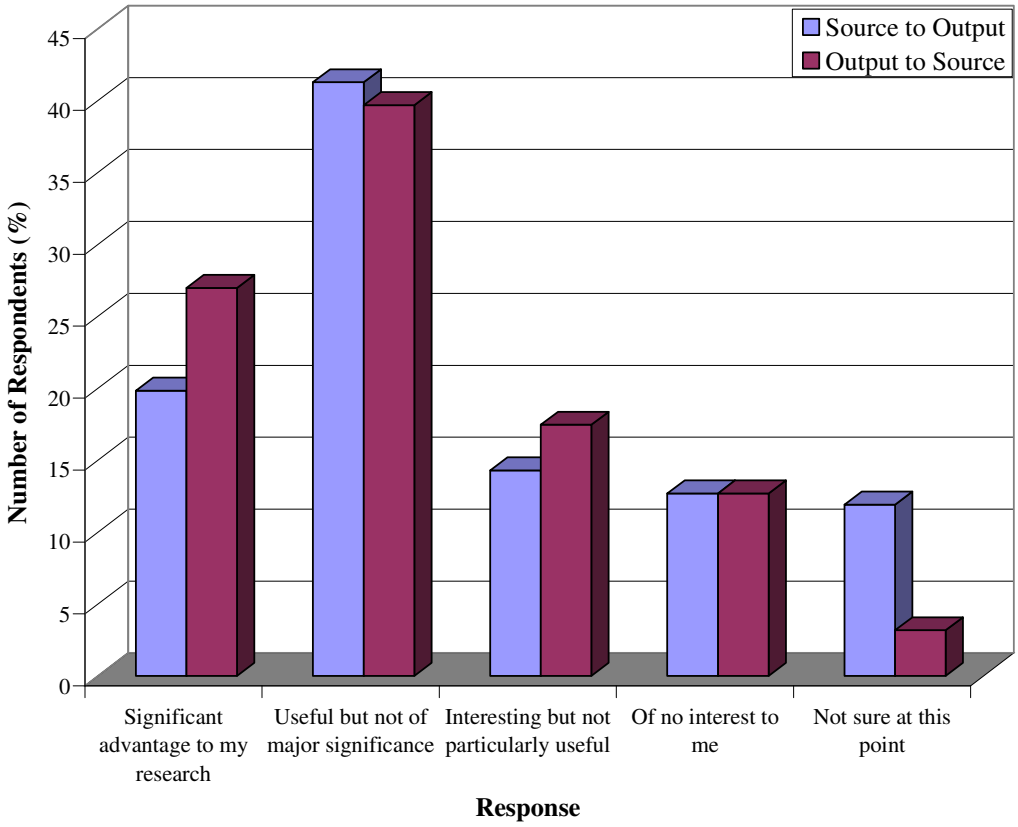

Figure 3.1: Perceived researcher benefit of being able to (a) identify and link from a source repository containing primary research data to the publications that had been developed from these data (source to output) (b) identify and link from within an online publication to the primary source data from which that journal was developed (output to source)

From these results it is clear that there is interest in the proposed linkage with the most popular response being 'useful but not of major significance' for both questions, followed by 'significant advantage to my work'. Only 25% of respondents said the source to output operation was 'of no interest to me' or 'not sure at this point', with even less (16%) selecting these two options for the output to source operation. Overall respondents showed slightly more interest for the output to source than for the source to output operation, with similar numbers selecting the 'useful but not of major significance' option (40% and 41% of respondents respectively), but greater numbers selecting the 'significant advantage to my work' (27% compared to 20%) and 'interesting but not particularly useful' (17% compared to 14%) options.

In addition to considering these distributions of perceived values of source to output repository links (and vice versa) it is also interesting to consider the perceived value of source to output repository links (and vice versa) relative to different research roles. This is done in Table 3.3 for the source to output linkage and Table 3.4 for the output to source linkage. It should be noted that the half-integer values that appear in Table 3.3 are due to two respondents who selected two options (which was not intended) for this question. Accordingly, where this is the case, each response has been equally weighted so as the total for any one respondent equals '1'.

Table 3.3: Perceived value of source to output linkage relative to different research roles

	University academic staff	University research assistant / fellow	Postgraduate student	Other	**Totals**
Significant advantage to my work	5	4	3	0.5	12.5
Useful but not of major significance	9	3	13.5	0.5	26.0
Interesting but not particularly useful	7	1	1	0	9.0
Of no interest to me	7	0	0	1	8.0
Not sure at this point	4	0	3.5	0	7.5
Totals	32	8	21	2	**63.0**

Considering the results of Table 3.3 first: research assistants / fellows seem the most enthusiastic group of researchers in favour of this operation, with 50% of them saying 'significant advantage to my work', 38% saying 'useful but not of major significance' and the remainder opting for 'interesting but not particularly useful. Most postgraduates have opted for 'useful but not of major significance' (64%), followed by 'not sure at this point' and 'significance advantage to my work' (17% and 14% respectively). Academic staff are fairly evenly distributed between the four options, with the remaining 13% 'not sure at this point'.

These results may reflect the stage that the researchers are in their career. Postgraduates who have less experience of the world of research are, understandably, less sure how such operations could be of benefit. Research assistants / fellows are more experienced and often in an early part of their careers: they may be able to see a need and show more enthusiasm to new operations that could potentially benefit them. The academic staff category is likely to cover the largest spectrum in career progression and thus could explain the largest range in responses.

Table 3.4: Perceived value of output to source linkage relative to different research roles

	University academic staff	University research assistant / fellow	Postgraduate student	Other	**Totals**
Significant advantage to my work	5	5	7	0	17.0
Useful but not of major significance	12	2	10	1	25.0
Interesting but not particularly useful	7	1	3	0	11.0
Of no interest to me	7	0	0	1	8.0
Not sure at this point	1	0	1	0	2.0
Other	0	0	0	0	0.0
Totals	32	8	21	2	**63.0**

Considering now the results of Table 3.4, where it can be seen that the results for the research assistants / fellows follow a similar trend as in Table 3.3. There is more enthusiasm amongst the postgraduates for this operation: although 'useful but not of major significance' is still the most selected answer for this group of researchers (48%), 'significant advantage to my work' follows behind much more closely (33%). Academic staff are also more positive about this operation, with fewer unsure respondents and a greater number

selecting 'useful, but not of major significance' (38% - compared with 28% in Table 3.3). The remaining three options for academic staff remained as in Table 3.3.

Further, it is interesting to consider the perceived value of source to output repository links (and vice versa) relative to different source repository communities. Source repositories will be discussed in detail in section 3.4, but for now it is sufficient to say that the named source repositories for Physics were pre-determined as being Brookhaven National Laboratory (BNL) [5] and the CERN laboratory [4]. Table 3.5 shows the perceived value of the source to output linkage relative to different repository communities, whereas Table 3.6 shows the perceived value of the output to source linkage relative to different repository communities. Again, as for Table 3.3, the half-integer values account for the two respondents who selected two answers for their perceived value of source to output linkage. The totals in this table (and indeed many of the future tables presented in this report) total a number greater than the 63 questionnaire respondents due to the fact that in a lot of the questionnaire questions it was appropriate for respondents to select more than one answer.

Table 3.5: Perceived value of source to output repository links relative to different repository communities

	Significant advantage to my work	Useful but not of major significance	Interesting but not particularly useful	Of no interest to me	Not sure at this point	**Totals**
BNL	0	0	0	0	0	0
CERN	5	2	2	1	2	12
None	8.5	19.5	5	5	4	42
Other	3	5.5	3	4	1.5	17
Totals	16.5	27	10	10	7.5	**71**

From Table 3.5 it can be seen that the biggest response by those who use the CERN source repository to the idea of linking source to output repositories is 'significant advantage to my work' (42% of CERN users). The remainder of CERN users are distributed evenly across the other 4 options. Those who use 'other' source repositories are not quite so enthusiastic, with the most popular answer here being 'useful but not of major significance' (32%). The remainder of the 'other' source repository users are fairly well distributed across the other 4 options. The majority of those who don't use source repositories again go for 'useful but not of major significance' (46%), followed by 'significant advantage to my work' (20%).

The enthusiasm amongst the non-source repository users is perhaps surprisingly high. However, the perceptions of this group of researchers may be very valid as, having zero or very limited experience of source repositories, they may be able to offer a more unbiased view and be able to see the potential benefits (without considering associated problems or issues) of such an operation that may tarnish the opinions of current users. In addition, this group of people are unlikely to contain many of the High Energy Physics (HEP – also known as Particle Physics) community (who produce masses of data) and therefore can perhaps see the proposed operation working in their field where their source data is not so large (and perhaps easier for others to interpret).

Table 3.6: Perceived value of output to source repository links relative to different repository communities

	Significant advantage to my work	Useful but not of major significance	Interesting but not particularly useful	Of no interest to me	Not sure at this point	**Totals**
BNL	0	0	0	0	0	0
CERN	3	3	4	2	0	12
None	12	19	7	3	1	42
Other	5	5	0	5	2	17
Totals	20	27	11	10	3	**71**

Table 3.6 shows that the CERN users are perhaps not as positive as for the source to output operation of Table 3.5, with less respondents opting for the 'significant advantage to my work' category. The numbers in the 'interesting but not particularly useful category have doubled. This slight shift in opinions may be due to the impracticalities that members of the HEP community can see, particularly in linking to the raw or experimental data (a topic which will be expanded upon further in section 4.6). 59% of 'other' users either opted for 'significant advantage to my work' or 'useful but not of major significance', with 24% stating 'it is of no interest to me'. The 'none' source repository users were again very positive, with 74% selecting 'significant advantage to my work' or 'useful but not of major significance'. Again, their positive responses may be due to their ability to consider the operation without being biased by practicalities or issues they may have experienced which could mean such an operation would be difficult.

3.3 Source Data

Figure 3.2 shows the types of source data that is generated by Physics researchers. As well as showing the variety that Physicists produce, this figure also shows the popularity of each media, as the size of each sector is proportional to the number of respondents that produce that particular type of source data. The most popular type of source data produced is raw data (with 73% of respondents producing this type of data), followed by drawings and plots (63%). Other popular source data produced include: text-based files (44%), derived data (43%), instrument data (43%), spectra (43%), databases (38%), images (38%) and statistical data (32%). The 'other' responses include: flash files (vector animation software), multidimensional images and simulation data. Two respondents were theorists who "use but do not generate any [source data]".

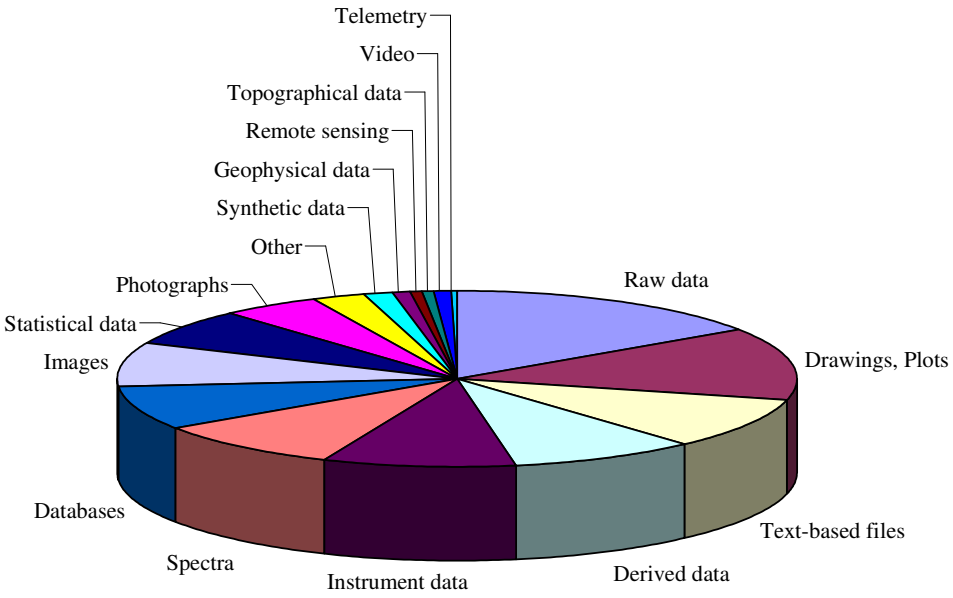

Figure 3.2: Type of source data produced by Physics researchers

Figure 3.3 shows the formats in which the source data generated by Physicists are held, revealing that Physics researchers produce source data in a wide variety of formats. The three most popular formats are: plain text (with 56% of respondents holding data in this format), image files (52%) and spreadsheets (48%).

As can be seen from Figure 3.3 the 'other' format is the option that was selected fourth most often (by 46% of respondents). The most popular 'other' format is 'special database files' such as 'Root' [6] (which is an object-orientated data analysis framework) and 'PAW' [7] ntuples. The former was selected by eleven respondents, the latter by two. One of these respondents stated:

"It is stored in a database, but nothing as simple as an Access file! It's one of the largest databases in the world! The format is Kanga / Root … I think it is of the order of petabytes in size."

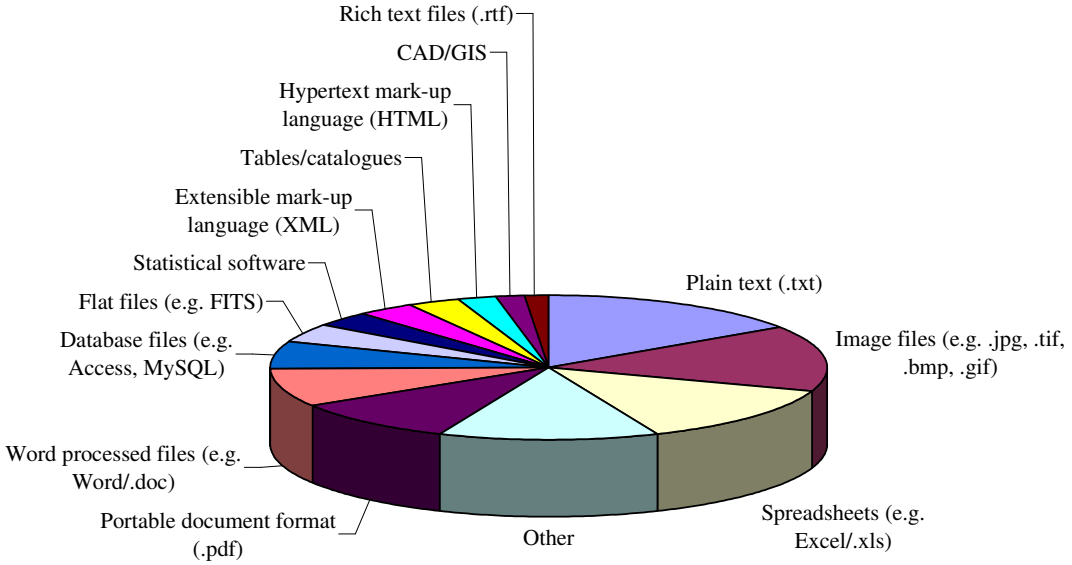

Figure 3.3: Format in which the source data produced by Physicists are held

Other formats that were selected in this category include open source software (such as gnuplot and xmgrace format for graphics and png format for images); own formats (such as individual or collaboration format), ascii format, C++, LaTex, binary files, postscript, kaleidagraph, Mathmatica notebooks and Origin (standard data presentation and processing program) workbooks.

As in section 3.2 it is interesting here to consider both the different types of source data generated and the formats in which they are held according to different repository communities. These results for the source data generated are shown in Table 3.7 whereas the formats in which they are held are shown in Table 3.8.

Across the CERN, 'other' and 'none' source repository users, the percentage usage of some of the media, such as drawings, plots, instrument and raw data are similar. Other data media differ across the repository users: for example databases are used by 67% and 59% of CERN and 'other' users respectively compared to 29% of non-repository users. Conversely non-repository users produce more images (43%) compared to 25% and 24% of CERN and 'other' users respectively. This trend is also true for photographs: 29% of non-repository users produce photographs, compared to 8% and 6% for CERN and 'other' users respectively. There is a big range in the production of statistical data across the three source repository users: 26%, 42% and 59% of 'other', CERN and non-repository users respectively produce this type of data. The biggest range exists for spectra with 0%, 18% and 60% of CERN, 'none' and 'other' repository users respectively producing this type of data.

From Table 3.8 it can be seen that the file formats of generated source data vary a lot across the three repository types, with statistical software, portable document files, flat files and hypertext mark-up language showing the smallest ranges. Some of the biggest differences exist for database files which are produced by 67% of CERN users, 47% of 'other' users and only 7% of non-users. The trend is the other way around for spreadsheets formats which are used by 17% of CERN users, 41% of 'other' users and 57% of non-repository users. Plain text format is used by approximately twice as many non-repository users (67%) compared to 33% and 29% for CERN and 'other' repository users respectively.

Table 3.7: The different types of source data generated according to different repository communities

	BNL	CERN	None	Other	**Totals**
Audio	0	0	0	0	0
Databases	0	8	12	10	30
Derived data	0	7	16	10	33
Drawings, Plots	0	7	30	10	47
Gene/protein sequences	0	0	0	0	0
Geophysical data	0	0	1	1	2
Images	0	3	18	4	25
Instrument data	0	4	18	7	29
Photographs	0	1	12	1	14
Plans, Maps	0	0	0	0	0
Qualitative questionnaire data	0	0	0	0	0
Quantitative questionnaire data	0	0	0	0	0
Radiographic data	0	0	0	0	0
Raw data	0	10	32	13	55
Remote sensing	0	0	1	1	2
Spectra	0	0	25	3	28
Statistical data	0	5	11	10	26
Synthetic data	0	1	2	1	4
Telemetry	0	0	1	0	1
Text-based files	0	5	20	5	30
Topographical data	0	0	2	0	2
Video	0	0	2	0	2
Other	0	0	5	3	8
Totals	0	51	208	79	**338**

Table 3.8: The different types of source data formats according to different repository communities

	BNL	CERN	None	Other	**Totals**
CAD/GIS	0	0	3	1	4
Extensible mark-up language (XML)	0	3	2	3	8
Database files (e.g. Access, MySQL)	0	8	3	8	19
Flat files (e.g. FITS)	0	3	4	3	10
Hypertext mark-up language (HTML)	0	2	2	2	6
Image files (e.g. .jpg, .tif, .bmp, .gif)	0	5	25	5	35
Plain text (.txt)	0	4	28	5	37
Portable document format (.pdf)	0	3	17	5	25
Rich text files (.rtf)	0	0	2	1	3
Spreadsheets (e.g. Excel/.xls)	0	2	24	7	33
Statistical software	0	1	6	2	9
Tables/catalogues	0	0	3	5	8
Word processed files (e.g. Word/.doc)	0	1	16	2	19
Other	0	10	12	15	37
Totals	0	42	147	64	**253**

Perhaps the most interesting difference across different source repository users is the ‘other format’ option, where 29% of non-repository users produce other data formats, compared to 83% and 88% of CERN and ‘other’ repository users. For CERN users this is probably a reflection of the specific software tools and formats that High Energy Physicists have to produce in their collaboration in order to understand the specific and complex nature of the data obtained from their detectors. The ‘other’ source repository category may also have a high ‘other format’ percentage due to the High Energy Physicists that are not included in the CERN repository category, for example, those who submit to DESY [8], SLAC [9] and Fermilab [10].

As well as generating data in a variety of formats, it is not uncommon for Physics researchers to generate data which are sometimes a combination or group of different data formats, as shown in Table 3.9, where it can be seen that 67% of respondents 'often' or 'sometimes' produce data in a combination of formats.

Table 3.9: How often data which researchers generate are a combination or group of different data formats

Response	Number of Respondents	Percentage (%)
Often	20	31.7
Sometimes	22	34.9
Rarely	10	15.9
Never	6	9.5
Potentially	1	1.6
Other	4	6.4
Totals	**63**	**100.0**

In addition to the types and formats of source data that respondents produce from their own research programme(s), the questionnaire also looked at why researchers might wish to access the research data generated by other research programmes and, if so, how researchers would normally access these research data. Figure 3.4 addresses the first of these issues, Figure 3.5 the second.

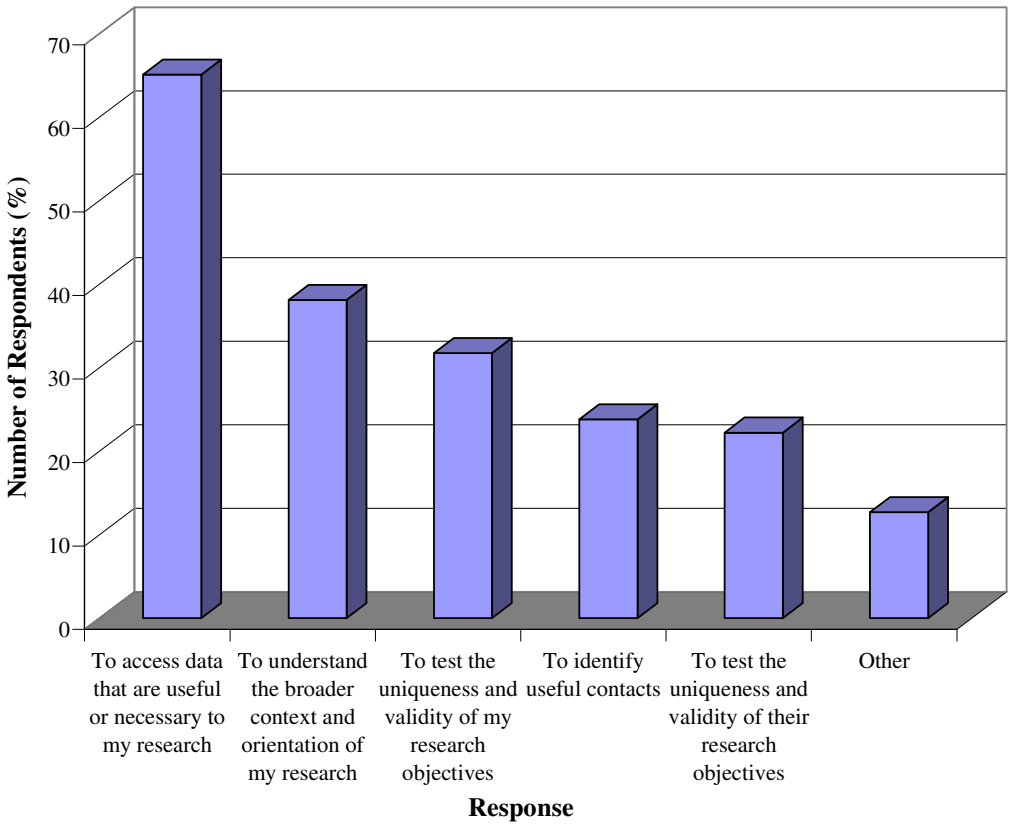

Figure 3.4: Reasons why researchers might wish to access the research data generated by other research programmes

From Figure 3.4 it can be seen that the three most selected options are all related to improving the quality of a researcher's own research with 'to access data that are useful or necessary to my research' being the most popular choice. Three of the 'other' respondents continued along this theme by saying they would wish to access others' data for comparative or cross-checking purposes. Respondents were less concerned about

accessing others' data 'to identify useful contacts' or 'to test the uniqueness and validity of their research objectives'. The remaining other responses included four respondents who don't wish to access others' data.

Figure 3.5 shows that the majority of researchers do not normally access others' research data. Methods of those that do are fairly evenly distributed. As can be seen from this figure, the 'other' option was a popular selection by respondents. The two most popular 'other' methods of access were via e-mail (which was cited by five respondents) and from publications (cited by five (different) respondents) with one respondent commenting: "Relevant data are published in articles". One researcher stated that they access research data from others by making use of personal contacts.

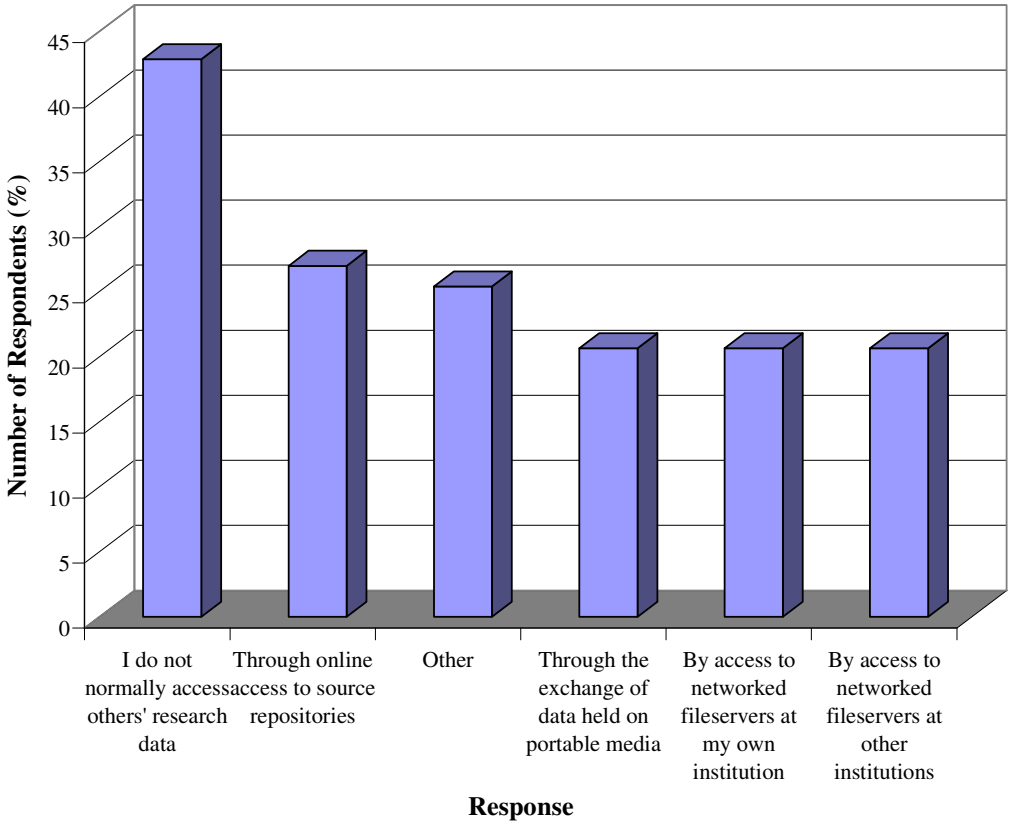

Figure 3.5: How researchers would normally access the data of other researchers

More generally, on the issue of accessing research data generated by other research programmes two Physics researchers stated in this part of the questionnaire that it is:

> "Usually impractical to access High Energy Physics experiments primary data. The expertise needed to understand it and the large size are usually prohibitive" and "the processing of data is something that can almost invariably only be done by groups very closely involved with the production of data."

3.4 Source Repositories

As mentioned already in section 3.2 the named source repositories for Physics were pre-determined as being Brookhaven National Laboratory (BNL) and the CERN laboratory. A description of each can be found in appendix B. Table 3.10 shows the number and percentages of responses to these named Physics source repositories as well as those who use 'other' or 'none' source repositories. It should be noted that the number of responses to this question is greater than the number of Physics respondents: this reflects the fact that a

number of respondents submit to more than one source repository. Thus for this question the percentages are calculated using the '71' responses rather than the '63' respondents

Table 3.10: The source repositories to which Physicists submit their data

Source Repository	Number of Responses	Percentage (%)
BNL	0	0.0
CERN	12	16.9
Other	17	23.9
None	42	59.2
Total	**71**	**100.0**

It is clear from Table 3.10 that the majority of respondents do not submit data to source repositories, with the remainder submitting to CERN and / or to 'other' source repositories. Although BNL was intended as a named Physics source repository the only four respondents to the questionnaire who had used this repository were three Biochemists and one Social Scientist. Therefore, for the purposes of this report, the BNL source repository will not be considered any further.

The 'other' source repositories used by Physics respondents were mainly those used by High Energy Physicists, with each of the following being cited by one or two respondents: Deutsches Elektronen-Synchrotron (DESY) [8], Fermi National Accelerator Laboratory (Fermilab) [10], Rutherford Appleton Laboratory (RAL) [11], Stanford Linear Accelerator Centre (SLAC) [9], The Durham HEP Database (HEPDATA) [12]. One respondent said that "Grid [13] data is distributed in different data centres (HEP groups and laboratories around the world)". Another respondent had made use of the UK Data Archive [14], whilst two theory respondents said the question was not applicable as they do not produce source data.

Asked about the frequency of submission of research data to these repositories most respondents stated 'frequent' as can be seen in Table 3.11.

Table 3.11: Frequency of submission to CERN and 'other' source repositories

	Frequently	On several occasions	Once	Never	Never, but intending to do so soon	**Totals**
CERN	7	2	0	1	2	12
Other	4	1	0	11	1	17
Totals	11	3	0	12	3	**29**

It is interesting to note that in Table 3.10, there are 29 responses saying data are submitted to CERN and / or 'other' source repositories and yet when asked about frequency of submission, Table 3.11 shows only a total of 14 responses in the 'frequent', 'on several occasions' or 'once' categories. The remaining responses were all in the 'never' or 'never, but intending to do so soon' categories, leaving some ambiguity as to how many respondents actually submit data to these named source repositories. The discrepancy may exist due to a misunderstanding in the first of these two questions where respondents may have entered details of source repositories to which their experiment data is submitted (but to which they do not personally submit) or respondents may download from (and not submit to) the specified source repositories.

3.5 Metadata

Researchers were invited to select from eleven types of generic metadata: 'what types of metadata do you consider it important to assign to your data?' They could select as many as were appropriate and were also given the opportunity to use the 'other' option to specify more discipline-specific terms. The results are shown in Figure 3.6.

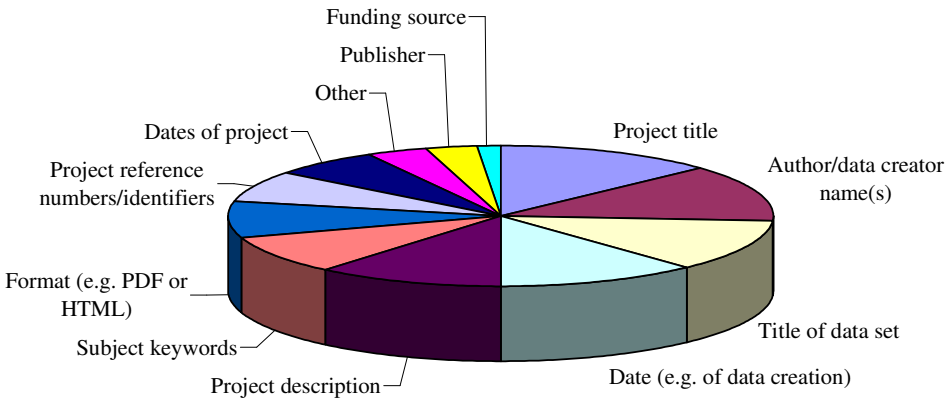

Figure 3.6: The types of metadata researchers consider important to assign to their research data

From Figure 3.6 it can be seen that there is a fairly wide recognition of the importance of all of the eleven generic terms that were offered, with between 60% and 71% of Physics respondents judging it important to assign the following to their data: project title, author / data creator name(s), title of data set, date (e.g. of data creation) and project description. The two least important terms were funding source (only selected by 8% of Physics respondents) and publisher (16%).

Many of the 21% of respondents who selected the 'other' option suggested additional metadata that was largely discipline specific. Some examples include: instrument details (including location), units (if not obvious), the type of experiment, related data sets (those taken about the same time), numerical values of selection criteria, time stamps for data points and how the data were generated. The following two suggestions were each made by two respondents: conditions under which the experiment was performed (specific examples of voltages and settings were given) and the software used (including version number). Two 'other' respondents commented on the range of data that they would need to identify and describe their data as well as the transient nature of their metadata:

> "For anything other than the journal publications derived from the data, a huge amount of additional information is required to make meaningful interpretation."

> "We have a ton of metadata which varies from experiment to experiment and from stage to stage. [Therefore] the above list is not relevant". Four 'other' respondents said the question was not applicable in their field whilst one respondent stated that they "see XML as the most appropriate format" for assigning metadata."

Table 3.12 shows the metadata requirements from Figure 3.6 arranged according to source repository communities. Using the number of respondents in each repository community (Table 3.10) the percentage of users in each community selecting each type of metadata can be calculated. From which it can be said that in all three source repository categories the percentage of respondents who selected project title, subject keywords, dates of project and dates of creation as important metadata were similar. One of the biggest differences across repository users was in project reference numbers / identifiers. CERN users valued this more than 'other' and non-repository users (67%, 41% and 31% respectively). Whilst 58% of CERN users and 59% of 'other' repository users thought author / data creator name(s) important, a higher proportion of non-repository users (76%) felt this important. Perhaps the biggest range exists in the 'other metadata' category where 10% of non-repository, 25% of CERN and 53% of 'other' repository users felt other metadata is required.

Table 3.12: Metadata requirements according to source repository communities

	CERN	None	Other	**Totals**
Project title	7	32	11	50
Project description	7	25	10	42
Project reference numbers/identifiers	8	13	7	28
Author/data creator name(s)	7	32	10	49
Title of data set	8	30	9	47
Subject keywords	5	21	8	34
Funding source	0	4	1	5
Publisher	1	7	2	10
Dates of project	4	14	5	23
Date (e.g. of data creation)	9	26	13	48
Format (e.g. PDF or HTML)	4	20	9	33
Other	3	4	9	16
Totals	63	228	94	**385**

The questionnaire then went onto ask 'at what stage are metadata assigned to your data?' Responses to this question are shown in Figure 3.7.

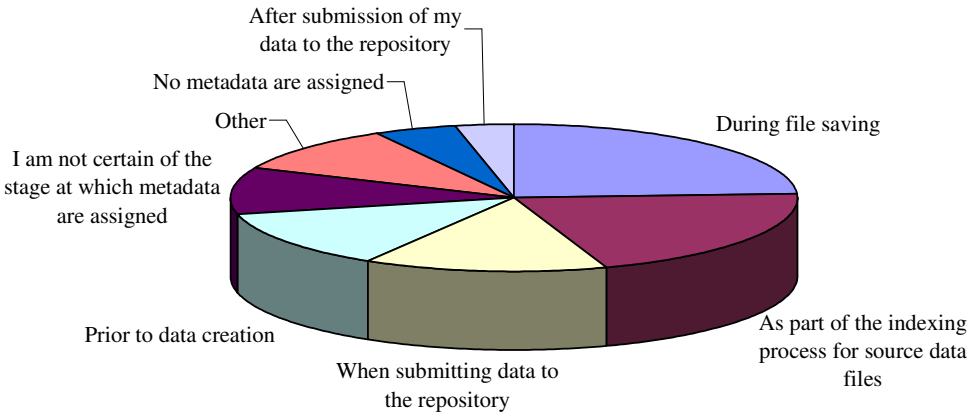

Figure 3.7: The stage at which metadata are assigned to the data of researchers

From Figure 3.7 it can be seen that metadata is most commonly assigned 'during file saving' or 'as part of the indexing process for source data files' followed by 'when submitting data to the repository' and 'prior to data creation'. 14% of the 63 respondents were not certain at what stage metadata are assigned to their data. 'Other' responses included five respondents who felt the question was not applicable to their field of research interest and one respondent who made the following comment: "metadata is assigned and refined at many stages after the raw data is taken".

Respondents were then asked 'who assigns metadata to your research data', the results for which are shown in Figure 3.8. Here it can be seen that metadata are most commonly decided by and assigned by the researchers themselves, followed by 'automatically generated' and then 'by research colleagues'. Metadata are rarely assigned by repository managers or information services / library staff. The popularity of research colleagues assigning metadata as well as the lesser selected option 'it is not known who assigns metadata', may explain the relatively high numbers of researchers included in Figure 3.7 who did not know at what stage metadata was assigned to their data. The majority of the 10 'other' respondents to this question said it was not applicable.

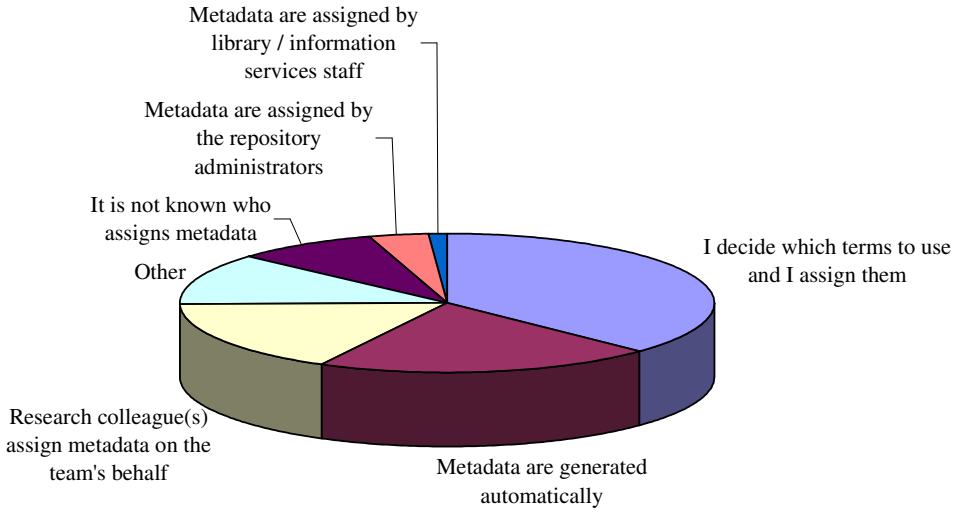

Figure 3.8: Who assigns metadata to research data?

In principle, it is instructive to consider the metadata assignment practices of 'at what stage are metadata assigned to your data' and 'who assigns metadata to your research data' relative to the level of support provided in the use of repositories (the issue of support is discussed in more detail in section 3.8) . However, in practice, due to limited statistics it is hard to draw too many meaningful conclusions. For completeness, though, these cross-tabulations are shown in Table 3.13 and Table 3.14.

As was seen from Figure 3.7, 'during file saving' and 'as part of the indexing process for source data files' were the two most selected times at which metadata are assigned to research data. In both cases the cross-tabulations of Table 3.13 show that for these two selections the most popular help provided by a librarian or other knowledge management support are in the 'provision of documentation' followed by 'online or telephone help' and 'formal training and documentation'.

From Table 3.14 it can be seen that the most help that is provided by a librarian or other knowledge management support is to the group of researchers who decide which metadata terms to use and who assign them themselves. This at first seems reasonable as those who actually do the metadata assignment themselves are more likely to require help (and therefore know what help is available) compared to those that don't. All types of help suggested in the questionnaire are used by this group of researchers with the 'provision of documentation' and 'formal training and documentation' being the most selected types of help. However, the most popular selection by those who chose and assigned metadata themselves was 'unknown' which implies that researchers do not require any help or look elsewhere for assistance. Many researchers also selected 'metadata are generated automatically' where again 'provision in documentation' and 'unknown' were popular cross-tabulations, as was 'online or telephone help'.

Table 3.13: The stage at which metadata are assigned relative to the level of support in the use of repositories

	Provision of documentation (guidance notes, fact sheets, etc.)	Formal training and documentation	Online or telephone help	Assistance with the structuring of specific searches	Assistance with the conduct of searches	Full intermediary service (e.g. the conduct of searches and organisation of results)	Unknown	Other	No Answer	**Totals**
Prior to data creation	6	3	2	2	2	0	2	1	0	18
As part of the indexing process for source data files	7	5	6	0	0	0	3	2	1	24
During file saving	7	4	4	2	1	1	5	2	1	27
When submitting data to the repository	4	0	2	1	2	0	4	2	0	15
After submission of my data to the repository	1	0	1	1	1	0	1	1	0	6
No metadata are assigned	1	1	1	0	0	0	3	0	0	6
I am not certain of the stage at which metadata are assigned	2	3	2	0	1	0	5	1	0	14
Other	0	1	2	0	0	0	3	3	0	9
Totals	28	17	20	6	7	1	26	12	2	**119**

Table 3.14: Who assigns metadata relative to the level of support in the use of repositories

	Provision of documentation (guidance notes, fact sheets, etc.)	Formal training and documentation	Online or telephone help	Assistance with the structuring of specific searches	Assistance with the conduct of searches	Full intermediary service (e.g. the conduct of searches and organisation of results)	Unknown	Other	No Answer	**Totals**
I decide which terms to use and I assign them	8	7	4	2	1	1	11	4	1	39
Research colleague(s) assign metadata on the team's behalf	5	1	1	1	2	0	5	2	0	17
Research support staff assign metadata on the team's behalf	1	1	2	0	0	0	0	0	0	4
Metadata are assigned by library/information services staff	0	0	0	0	0	0	1	0	0	1
Metadata are assigned by the repository administrators	1	0	0	1	1	0	2	0	0	5
Metadata are generated automatically	6	1	5	2	2	0	6	2	0	24
It is not known who assigns metadata	3	4	3	0	1	0	3	0	0	14
Other	0	0	2	0	0	0	5	3	0	10
Totals	24	14	17	6	7	1	33	11	1	114

3.6 Data Access and Sharing

This section is quite wide-ranging and covers the measures that respondents currently adopt to make their research data available to others as well as the formal restrictions applied to their data and the measures that are normally used to control access to their data by others. It also considers the factors that would encourage and discourage researchers to share their data.

Figure 3.9 shows the measures that the Physics respondents adopt to make their research data available. The range in responses is not vast, with the most popular measure being via e-mail (a method used by about a third of respondents), and the least popular method being via printed form (13%). 30% of respondents undertake no measures to make their research data available. There were a range of 'other' comments including 4 respondents who stated the question was not applicable (including the theorists who do not generate data) and 2 respondents who said that data is only available to members of their collaboration. A further respondent said "I will make my research data available to those who ask for it, which only usually happens inside our research group alone". Four respondents stated explicitly that publications were their

source of 'useful' research data and thus how their research data were made available. One respondent stated "journals or pre-print archives or conferences are the only realistic manner at present" whilst another stated "if someone asks for data it would be provided but most of the useful numbers are contained in publications".

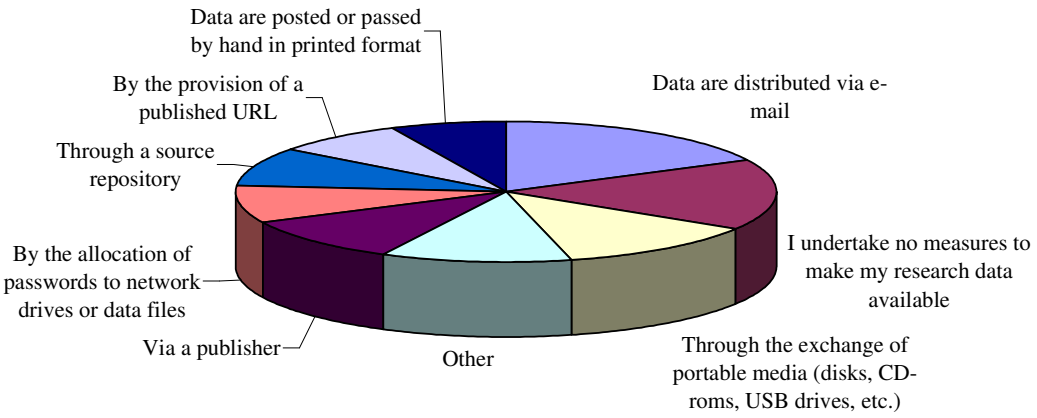

Figure 3.9: Measures currently made by researchers to make their research data available

Figure 3.10 shows the normal kinds of formal restrictions that apply to respondent's research data.

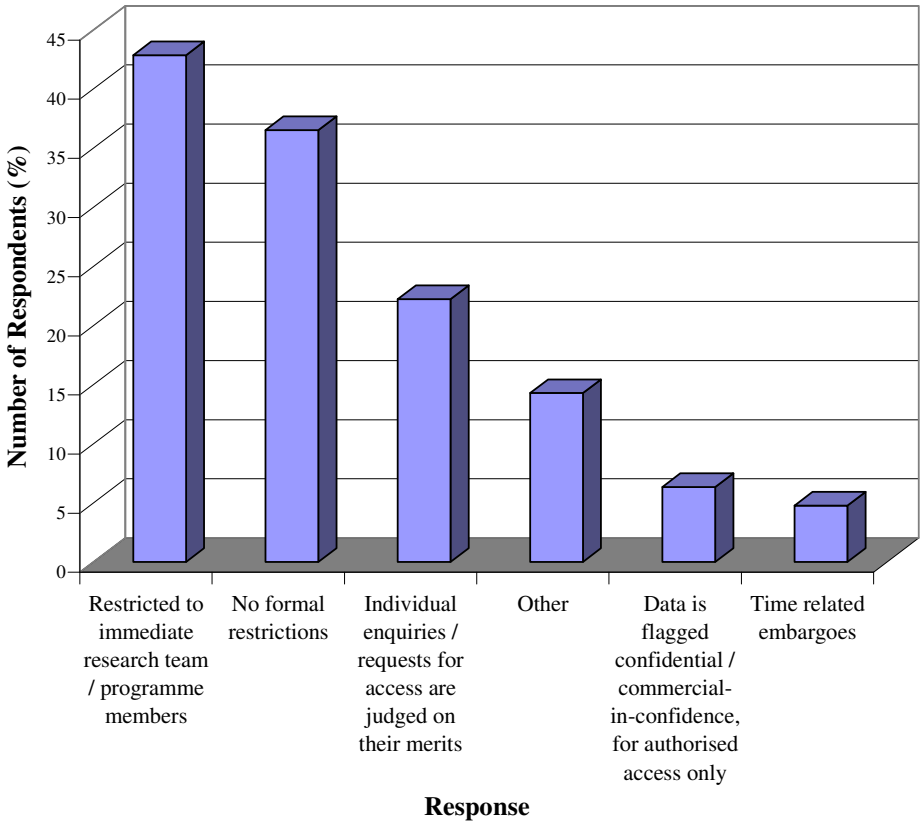

Figure 3.10: Formal restrictions that researchers normally apply to their research data

From Figure 3.10 it can be seen that most researchers either apply maximal ('restricted to immediate research team / programme members') or minimal ('no formal restrictions') access restrictions to their research data. Most of the 'other' responses were 'not applicable' (for example theorists who have no data to share). One 'other' respondent said they would "ask that [their research] data is acknowledged in publications", whilst another stated "once published, anyone can have access [to their research data] if they ask".

Figure 3.11 shows the measures that respondents normally use to control the access of data by others. The most popular response, selected by 54% of respondents, was 'stored on a private network / internet', followed by 'authorisation of ID and password for online access' and 'storage of data on standalone computers (both selected by 21% of respondents). The 'validation of data users by clicking on an e-mailed URL' was a measure not used by any Physics respondent. The majority of the 'other' comments were again 'not applicable'. Additional 'other' comments included "at present just share data with immediate members of my research group", "requests considered on a case by case basis" and "this depends on the maturity of the data".

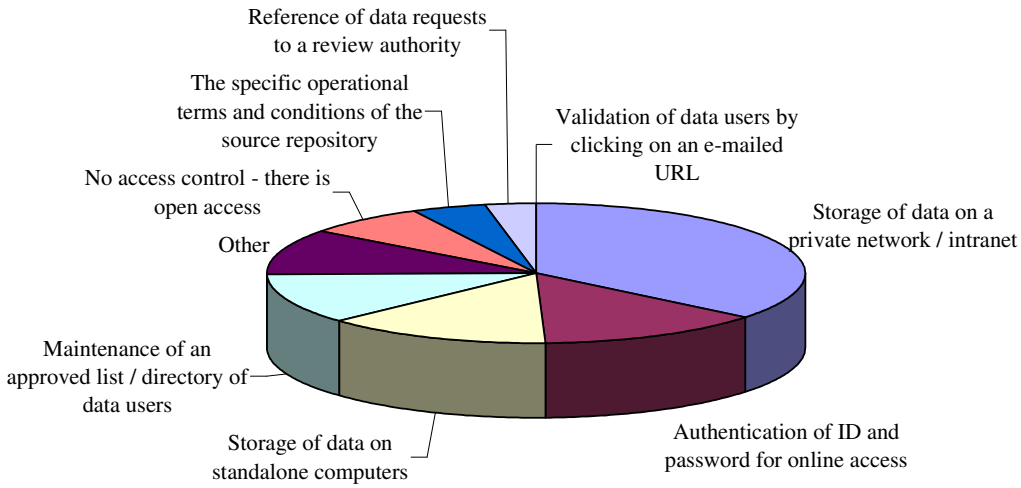

Figure 3.11: Measures normally used by researchers to control the access of data by others

Figure 3.12 shows the factors that would encourage respondents to share their data, whilst Figure 3.13 shows the factors that would discourage respondents from sharing their data.

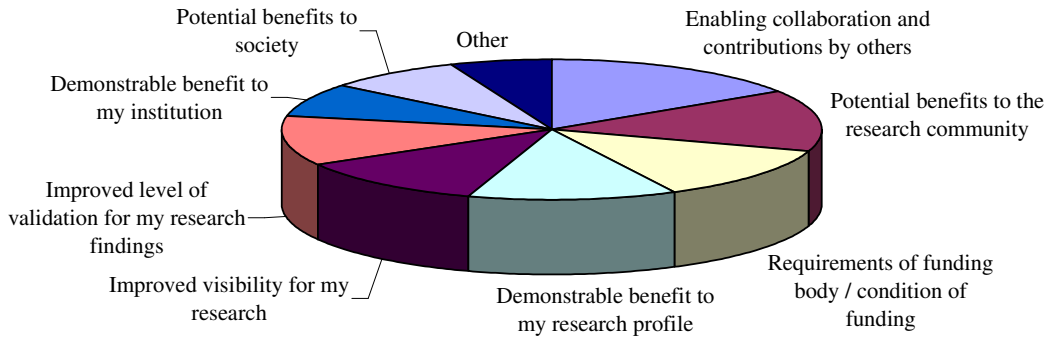

Figure 3.12: Factors that would encourage researchers to share their data

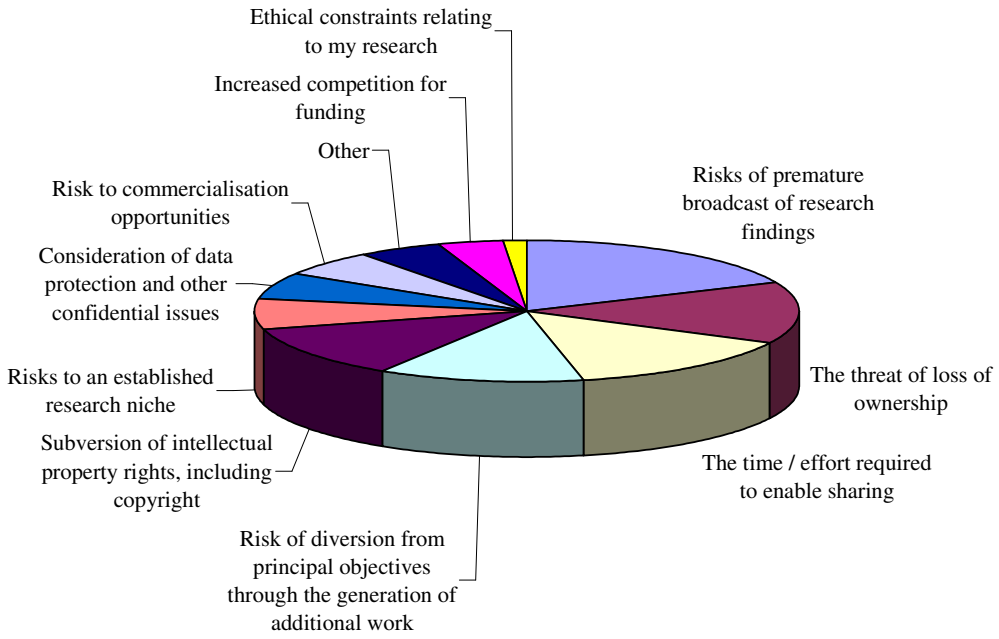

Figure 3.13: Factors that would discourage researchers from sharing their data

From Figure 3.12 it can be seen that 'enabling collaboration and contributions from others', 'potential benefits to the community', 'requirements of funding body / condition of funding', 'demonstrable benefit to my research profile' and 'improved visibility for my research' are seen as being important, whilst 'demonstrable benefit to my institution' and 'potential benefits to society' are seen as being less important. From Figure 3.13 it can be seen that 'issues of premature broadcast of research findings', 'the threat of loss of ownership', 'the time / effort required in order to enable sharing', 'risk of diversion from principal objectives through the generation of additional work' and 'subversion of intellectual property rights, including copyright' are all seen as being important, whilst 'ethical constraints relating to my research' and 'increased competition for funding seem less important.

The main issues of the 'other' comments for discouragements for sharing data were those on (or around) the subject of interpretation. The following are three quotes on this theme which are of particular interest:

> "The main concern is that data could be incorrectly interpreted without intimate knowledge of the experiment itself, and to make this information available to someone who was not already a collaborator on the project would impose a huge burden on those who were."

> "Misleading nature of uncorrected data."

> "Premature in this context would mean invalid interpretation and e.g. rushing to publish without understanding fully the data."

One 'other' respondent also cited the size of their datasets as being a discouraging factor for sharing data:

> "The research I perform includes very large datasets, in the range of hundreds of gigabytes. Without the necessary infrastructure to host this data, sharing such information in an easily accessible medium would be … well, very difficult".

Two more general comments on the 'encouragements' and 'discouragements' theme are:

> "Not encouraged or discouraged from sharing data. Just does not justify the effort."

"I am not sure who owns data taken by CERN experiments – all EU governments pay for it."

Although this latter comment was only made by one respondent the issue of who owns the data may be an important one in the context of this project.

3.7 Output Repositories

This section covers the theme of output repositories and looks at those output repositories which are used for both research and teaching as well as preferred routes to output repositories. This section also considers the methods of searching within this type of repository.

Figure 3.14 shows the kind of output repositories that Physics respondents use to find and retrieve information for use in their research and teaching activities. From this figure it can be seen that the majority of Physicists use output repositories for their research (only 3% of respondents stated 'none'), whereas their use in teaching is much less (with 46% saying they use 'none'). It is common for researchers to use more than one type of output repository, with Institutional and Publisher repositories being the most popular types for research (used by 63% and 65% of Physicists respectively), followed by discipline (44%). For teaching this distribution is more even across the three (ranging from 25-32% Physicists for each).

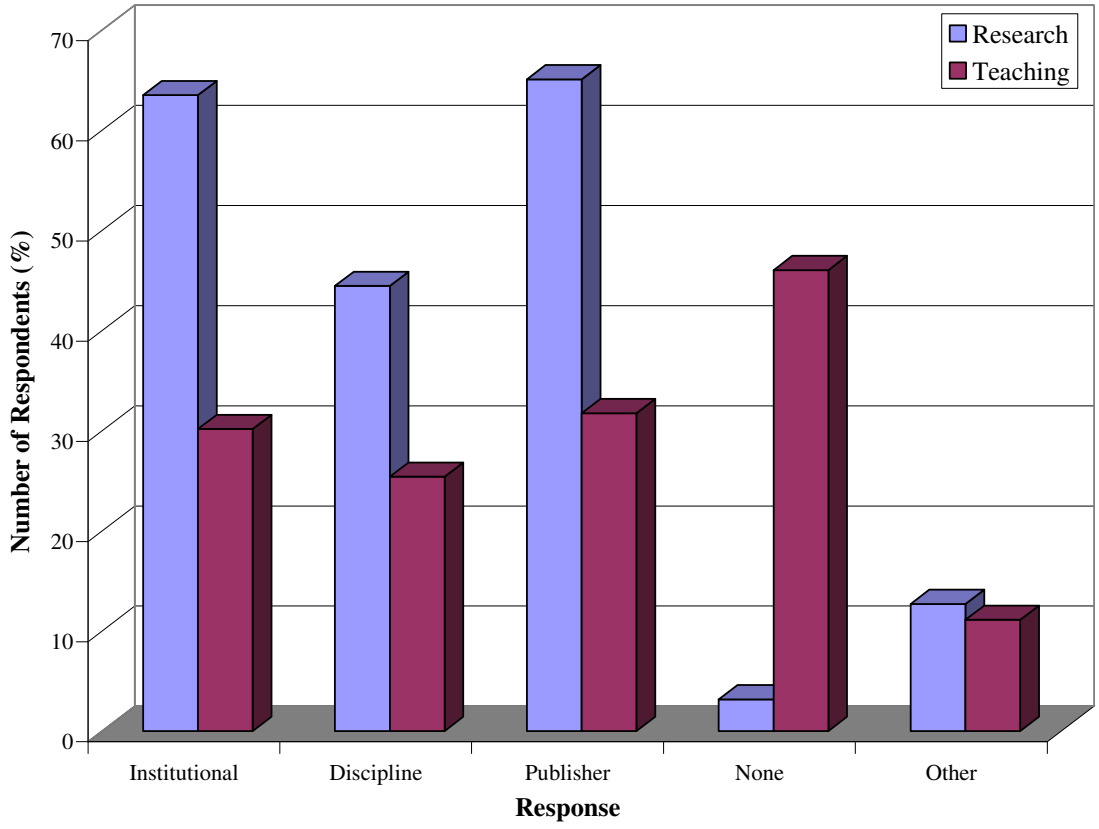

Figure 3.14: The kinds of output repository used by researchers to find and retrieve information for use in (a) research and (b) teaching

'Other' output repositories that were stated as being used by Physicists include the arXiv preprint server [15] (mentioned here by four respondents), institutional or collaboration websites (four respondents) and information found from internet searches (three respondents). It became clear from both the questionnaire 'other' responses and particularly from the interviews (discussed further in section 4.7) that there was some

ambiguity in the understanding of the term 'institutional repository' (with many selecting this option when referring to University website, University library and / or University library e-journals). The results here, therefore, for institutional repositories may be superficially high.

It is interesting here to see the cross tabulations of usefulness of output repositories, for both research and teaching, as compared to users of named source repositories. These results are shown in Table 3.15 for research use and Table 3.16 for teaching use.

Table 3.15: Usefulness of output repositories for research as compared to users of named source repositories

	CERN	None	Other	**Totals**
Institutional	8	28	10	46
Discipline	7	14	12	33
Publisher	6	32	8	46
None	0	1	1	2
Other	2	5	4	11
No answer	0	1	0	1
Totals	23	81	35	**139**

From Table 3.15 it can be seen that CERN users (58%) and 'other' source repository users (71%) makes use of discipline output repositories more than non source repository users (33%). This is compensated by 'non-users' making a greater use of publisher output repositories (76%), compared with 50% and 47% CERN and 'other' users respectively. There is more of a uniform usage across the repository communities for 'institutional' output repositories.

The trends shown for discipline and publisher repositories across the three disciplines may be influenced by the fact the High Energy Physics is currently well looked after with discipline repositories (for example, arXiv, SPIRES [16] and the CERN Document Server [17]).

Table 3.16: Usefulness of output repositories for teaching as compared to users of named source repositories

	CERN	None	Other	**Totals**
Institutional	4	11	5	20
Discipline	5	7	6	18
Publisher	5	12	7	24
None	4	21	8	33
Other	1	5	2	8
No answer	0	1	0	1
Totals	19	57	28	**104**

Table 3.16 depicts the same trend for discipline output repositories in teaching as in research. Institutional and publisher usage are fairly equally used across the repository communities but, as mentioned already, the biggest response to this question was the declared no use of output repositories for teaching.

Figure 3.15 shows the same categories of output repository as in Figure 3.14, but this time indicating where respondents deposited their research publications. The most popular option selected here are publisher repositories, which are used by 62% of physicists, followed by institutional repositories (used by 44% of respondents) - although as mentioned already, the total may be superficially high due to the ambiguity in the understanding of this term - and discipline repositories (41% of respondents). 'Other' responses included: the arXiv pre-print server and "public web-page maintained by the collaboration".

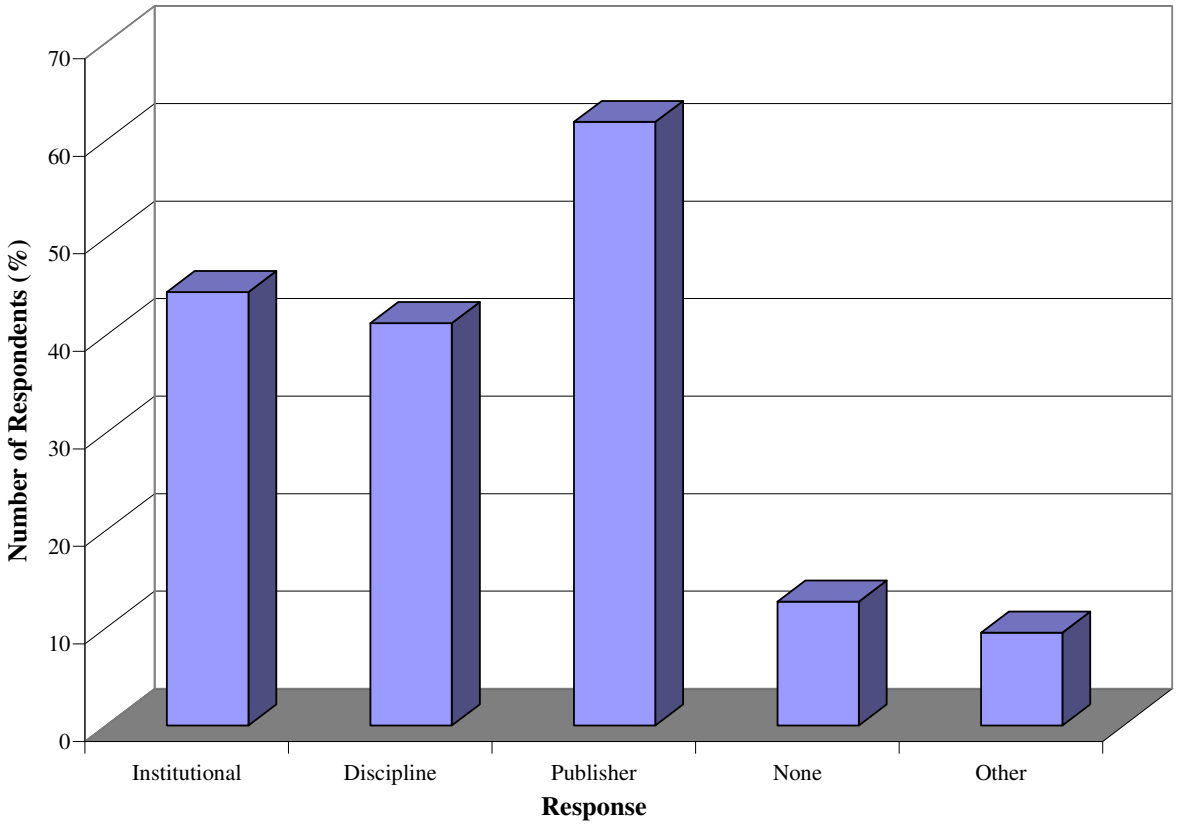

Figure 3.15: The output repositories used by researchers to deposit their research publications

Figure 3.16 shows the normal or preferred routes of Physics respondents to the contents of output repositories. From this figure it can be seen that via a known repository's URL and from an internet search engine are the two methods most used by Physicists (67% and 62% respectively), although directly to an article via a library catalogue, through a publisher's online service and through a journal's own web site are also well-used. The four 'other' respondents all said they went via the 'Web of Knowledge' [18].

Table 3.17 shows the cross-tabulation of the results from Figure 3.16 with the users of named source repositories. From this table, a number of small differences in the preferred routes to output repositories can be seen across the different source repository communities. For example, a non-source repository user was more likely to go via a library catalogue (52%) compared to a CERN (33%) or an 'other' source repository user (24%). The percentages of CERN and 'other' user routes to output repositories are very similar for 'through a publisher's online service' (both ~25%), 'directly through a specific journal's own web site' (both ~34%) and 'through an author's personal web page' (both ~18%), whereas the non-source repository users are all higher for these three methods (55%, 50% and 31% respectively). Reflecting the results of Figure 3.16, the two most popular methods across all three source repository communities are: via a known repository's URL and from an internet search engine.

Figure 3.17 shows the level of searching that Physics respondents usually find sufficient when using an output repository. Over half the respondents normally find a 'simple' search sufficient, whilst one third of the respondents usually use either an advanced or Boolean search. Six respondents expressed no preference. The one 'other' respondent stated that they "usually start off with a simple search and narrow down the search using more advance terms should too large a result be returned".

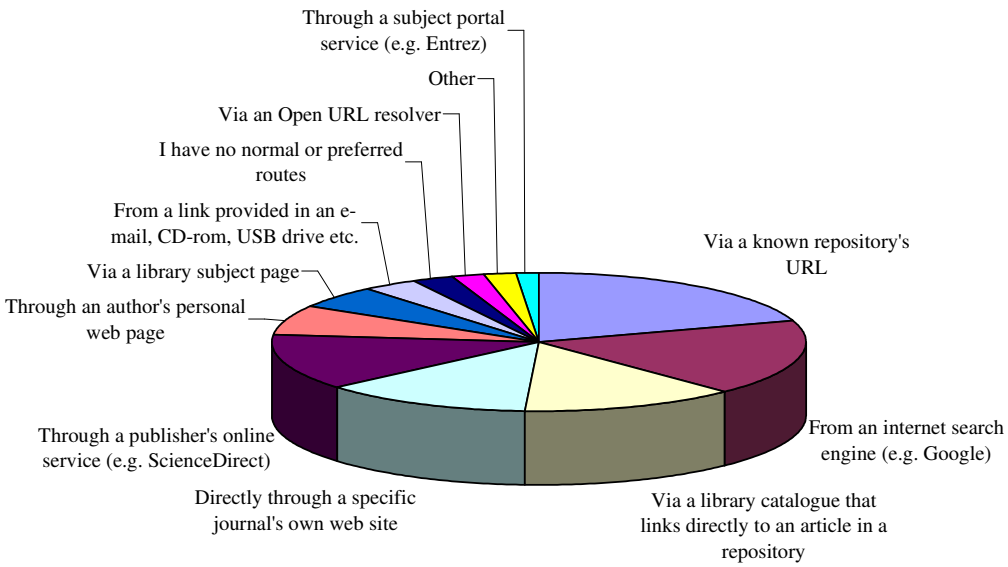

Figure 3.16: Normal or preferred routes of researchers to the contents of output repositories

Table 3.17: Preferred routes to output repositories compared by users of named source repositories

	CERN	None	Other	**Totals**
Via a known repository's URL	9	27	14	50
Via an Open URL resolver	1	3	0	4
Via a library catalogue that links directly to an article in a repository	4	22	4	30
Via a library subject page	2	8	2	12
Through a publisher's online service (e.g. ScienceDirect)	3	23	4	30
Directly through a specific journal's own web site	4	21	6	31
Through an author's personal web page	2	13	3	18
From a link provided in an e-mail, CD-rom, USB drive etc.	2	4	1	7
From an Internet search engine (e.g. Google)	8	26	8	42
Through a subject portal service (e.g. Entrez)	0	3	0	3
I have no normal or preferred routes	1	3	1	5
Other	0	4	1	5
No Answer	0	1	0	1
Totals	36	158	44	**238**

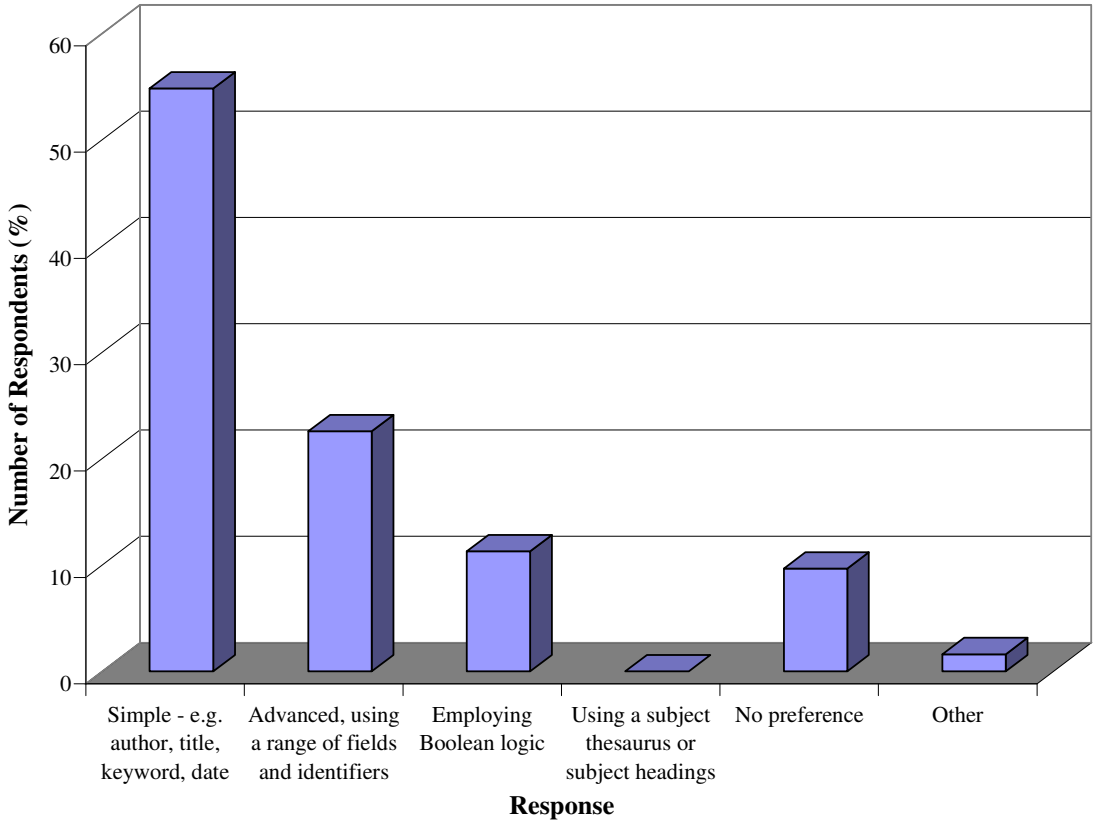

Figure 3.17: The level of searching that researchers usually find sufficient when using an output repository

Table 3.18 shows the cross tabulation of the output repositories that researchers use in the course of their research and the level of searching that they usually find sufficient when using an output repository. The results are quite uniform across the three main output repository types and echo the general picture shown in Figure 3.17: a simple search is the most popular search type in all three types of output repository (institutional, discipline and publisher). Advanced searching is the next most popular search method, although this is slightly less popular with discipline repository users. Employing Boolean logic is less popular, but its use is fairly evenly distributed across the three repository types.

Table 3.18: The level of searching that is sufficient to researchers across different types of output repository

	Institutional	Discipline	Publisher	None	Other	No Answer	**Totals**
Simple - e.g. author, title, keyword, date	20	16	24	1	4	0	65
Advanced, using a range of fields and identifiers	11	4	9	0	2	0	26
Employing Boolean logic	4	3	5	0	1	0	13
Using a subject thesaurus or subject headings	0	0	0	0	0	0	0
No preference	4	4	2	1	1	0	12
Other	1	1	1	0	0	0	3
No Answer	0	0	0	0	0	1	1
Totals	40	28	41	2	8	1	**120**

3.8 Support

The final topic addressed by the questionnaire was that of support and focused specifically on any support and / or guidance received in the use of output repositories, as well as assistance in the use of repositories (both source and output) that is provided by a librarian or other knowledge management support (KMS). The results are shown in Figure 3.18 and Figure 3.19 respectively.

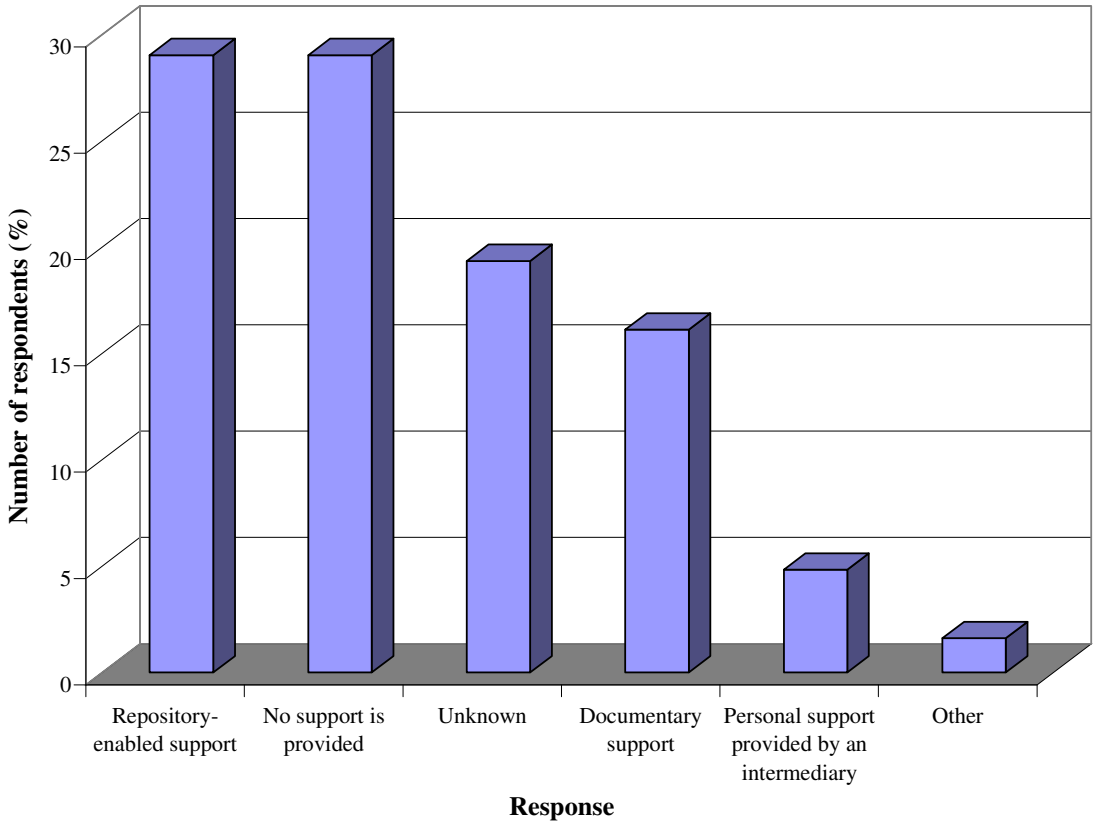

Figure 3.18: Support and / or guidance received by researchers in their use of output repositories

From Figure 3.18 it can be seen that 58% of Physics respondents have either used repository-enabled support or no support. Perhaps surprisingly, 19% of respondents didn't know if they had received support and / or guidance. The one 'other' respondent stated that "specific training was provided by their institution".

It is clear from Figure 3.19 that a large fraction of Physics respondents do not make use of assistance from librarians or other KMS in the use of repositories, as the most popular response to the question of 'what assistance in your use of repositories is provided by a librarian or other knowledge management support?' was 'unknown'. Of those who knew what assistance was offered by librarians or other KMS, the most popular choices were 'provision of documentation' and 'online or telephone help'. Continuing the theme of limited help from librarians or other KMS in assisting Physicists in their use of repositories, seven of the eight 'other' respondents stated "none" (implying no assistance is provided). The remaining other respondent stated assistance was provided in the form of "provision of passwords".

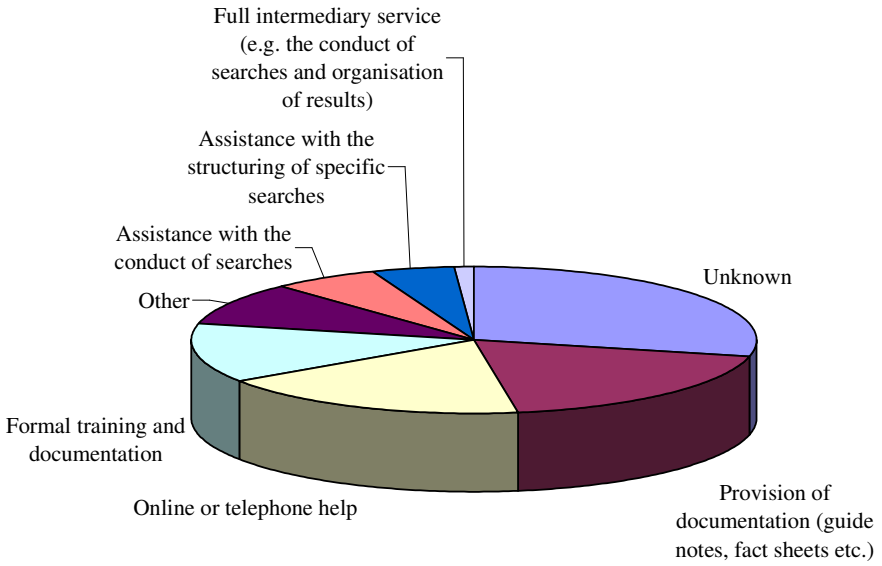

Figure 3.19: The assistance in use of repositories that is provided by a librarian or other knowledge management support

Table 3.19 presents the cross tabulation of these two questions of support, showing a distribution of results that indicate the statistics are low, thus making it difficult to draw firm conclusions. However, it is clear that 'repository enabled support' is the most popular support that is required and the assistance that is received to help with these issues is from many sources: the most popular being 'online or telephone help', the least popular being a 'full intermediary service'. The only number in double figures here is the 'no support is provided' level of support that is received versus 'unknown' level of help available.

Table 3.19: The level of support / guidance provided matched against professional intermediation

	Documentary support	Personal support provided by an intermediary	Repository-enabled support	No support is provided	Unknown	Other	No Answer	**Totals**
Provision of documentation (guidance notes, fact sheets, etc.)	5	1	4	5	1	0	0	16
Formal training and documentation	2	0	4	2	2	1	0	11
Online or telephone help	2	0	7	3	3	0	0	15
Assistance with the structuring of specific searches	0	0	4	0	0	0	0	4
Assistance with the conduct of searches	0	1	3	0	0	1	0	5
Full intermediary service (e.g. the conduct of searches and organisation of results)	0	0	1	0	0	0	0	1
Unknown	4	0	3	10	7	0	0	24
Other	1	1	2	2	2	0	0	8
No Answer	0	0	0	0	0	0	1	1
Totals	14	3	28	22	15	2	1	**85**

4 Significant Observations from the Interviews

This chapter presents the results from the Physics interviews, accompanied by a commentary. The volume of data gained from this phase of the project is very large and consequently it is not possible to include everything in a report of this nature. However, sufficient comments and quotations have been included to represent both consensus and differences in opinion on all the main topics of the project. For reference, the interview script is provided at appendix A.

4.1 Identities

Of the 20 respondents to the questionnaire who indicated that they would be happy to participate in a project interview, 13 were actually able to take part. The remaining 7 respondents had either changed their minds, not realised what the interview entailed or had other commitments (such as the exam season). However, as can be seen from Table 4.1 the 13 interviewees represented a good cross-section of academic staff, research assistants / fellows and postgraduate students from a range of Physics areas of interest. The non-intentional bias towards High Energy Physics (HEP – also known as Particle Physics) is actually very useful to this project as it is this group of Physicists that currently makes the most use of source repositories (see section 3.4). As will be seen in section 4.4, the remaining 5 interviewees do not use source repositories or use them on a smaller scale, and therefore complete a balanced sample of 'source repository' and 'non-source repository' users. The interviewees were from 5 universities across the UK.

Table 4.1: Identities of Physics interviewees

Area of Interest	Academic Staff	Research Assistant / Fellow	Postgraduate	Totals
High Energy Physics	3	2	3	8
Nanoscience	1			1
Organic Electronics	1			1
Semiconductor Physics		1		1
Solid State Physics	1			1
Surface Physics	1			1
Totals	**7**	**3**	**3**	**13**

4.2 Project Aims

A lot of information was gained on this topic during the course of the Physics interviews, therefore this section will comprise of an overview of the main findings, followed by individual subsections on source to output linkage and output to source linkage. The section will finish by considering some functionality that may be useful in achieving the proposed linkage between source and output repositories.

4.2.1 General Overview

There was a mixed response to the usefulness of linking source to output repositories (and vice versa) amongst interviewees. However, after some examples and discussion the majority of interviewees showed a greater enthusiasm and could see some benefit of the proposals to their work.

Two main points emerged: firstly, the source data of many interviewees that may be useful to others is the final Physics results produced towards the end of a particular analysis and that in a lot of cases the 'raw' or 'unprocessed' data would be of little or no use to others. Therefore, in Physics, linking output repositories to

source repositories is most likely to be of use to researchers if the emphasis on what is contained in these source repositories is centred upon final Physics results. One interviewee said:

> "To find some effective way of having more rapid and straightforward access to the data at the level at which it was ready for publication would really help."

The issues surrounding why the raw or unprocessed data are considered to be of limited use to others is explored later in this chapter.

The second main point that emerged was that almost all interviewees could see benefits in the specific output to source repository operation of being able to link to the 'source' (numerical) data that have made up a figure or a plot that appears in a publication (stored in an output repository). Many researchers commented that these details are seldom given in publications and would be very useful:

> "If this data could be part of a paper or in a source repository this would be very good."

> "In my field the numbers [making up a plot] are very useful."

> "If I was in a journal and saw a figure of interest, I would find it interesting to click on that figure and get the data that makes up that figure … However, I probably would not have a need for this often, but it would be helpful in the few instances when I need to do such an operation."

Continuing on from this latter theme, another interviewee stated:

> "I would find it very useful to get high-resolution electronic versions of individual figures from within a publication, or at least have a reference or link to where they can be found, as it is very difficult to get an enlarged good quality figure from a publication. So, for example, if you want to look at a figure from a publication in some detail, this is currently very difficult."

It follows, that if it were possible to link from output repositories to source repositories containing the data which made up the figures, this researcher would be able to re-plot the data for themselves and produce a high resolution output.

Two High Energy Physicists summed up these two main points well: both were showing their enthusiasm for linking from a figure in a publication to the numerical data making it up whilst also commenting that the most useful data they produce is the final Physics results.

> "This is probably the most useful enhancement: the numerical data points making up a plot, although only a very small part of the data is one example of where source data can be useful to other people reading or studying a publication."

> "The link between the paper and the analysed data would be very useful. So when you read a paper you click on the link that will send you … to the data that is used in the paper."

4.2.2 Source to Output Linkage

As was the impression gained from the questionnaire, it was felt that the interviewees offered support for the process of going from a source to an output repository, but were less certain how this operation (compared to going from output to source repository) could be of benefit to them.

Whilst two interviewees explicitly said that "yes, I think this [source to output operation] would be useful", a number of interviewees expanded further and gave examples of the situations in which this linkage could be useful. Three such examples include:

"If I look in a source repository for data, I would want to find where that data has been published so as I could assess the credibility of that source data."

"If authors are taking your source data and combining it with theirs to make a comparison or extend the data set, it would be useful (and beneficial to my future work) to be able to see what they have done and / or achieved"

"When work is published often the same material appears, but you are not always quite sure whether it is exactly the same data, or new data. If you could identify which papers had come from a particular data set … this would be helpful."

4.2.3 Output to Source Linkage

In corroboration of the impression gained from the questionnaire, there was a slight preference amongst interviewees for the option of going from output to source, compared with source to output, repositories and it was in this direction that the interviewees could identify more obvious benefits.

In addition to being able to navigate from a figure (or plot) in a publication to the numerical source data that was used to create it, the most popular feature amongst the majority of interviewees, there were a number of other suggestions how this link could potentially benefit Physics researchers. Two researchers suggested it could allow them to run tests or examine further interesting features that they saw in publications:

"I could perhaps see a situation where you read a publication and you see some interesting feature (for example a figure) that you don't quite understand or you believe to be wrong. It might be interesting to be able to analyse the processed data yourself to look for an error".

"I could link from publication to source data to do some tests, if there is something that I don't understand in that publication."

A second example of potential benefit was suggested by two interviewees who believed it would help in the comparison of results from other collaborations. Two further interviewees expected that such a feature would allow them to extend their data set.

A further two interviewees commented that the proposed output to source linkage may be particularly beneficial to Physics Theorists who, if they were able to access the data detailed in publications more easily, could examine different fits to that data and test out different theories. One further interviewee commented:

"If I have the data from a figure, I might wish to perform a fit to that data. Or, if a fit procedure is described in the paper, I could then try it on the data to see how it works and to understand it better."

A few comments were made about how the output to source linkage could allow additional information to be made available that was unable to be included in publications (often due to space restrictions). Thus a researcher could link from output to source repository to access further information and increase their understanding of a subject.

"Publications usually only contain a compressed version of the research carried out (due to page limits or word counts set down by most journals) and cannot include other ancillary items which are interesting, but not directly relevant to the main aims of the paper, for example simulation information, information on devices that did not work (and why) and general background information. It would therefore be interesting if this information could be made available."

"In a publication, one is very limited for space and therefore what is included must be very concise. This is especially a very real problem for high impact publications, where space may be limited to a short article with only 3-4 figures. Having some linkage between output and source repositories could

overcome this space problem. For example, if I produced a paper with 3-4 figures in, but I also had some additional figures which would be useful to the reader; it would be helpful to get a link from the publication to these additional figures."

"More information is always better, for example could provide circuit diagrams, schematics which again there would not necessarily be room for in a publication – this information could be especially useful to new postgraduate students and new researchers."

"Yes, but as this is a new idea, we would have to get used to doing it. Currently, we have to tune the article to fit the requirements of the journal. Therefore the collaboration would have to become used to providing the extra information, as the articles are usually well self-contained. We usually try and expand later in a journal that allows bigger articles in order to provide the plots that didn't fit into the restricted space ones. So this does happen and it would be useful to not have to wait (for the longer item to be published) or to remove the need for publishing the longer item, but … we would have to become used to doing it."

"Yes this would help. In HEP, however, one can find more background information and additional figures by looking in the presentations made at conferences."

"No: the author should consider and choose a more appropriate article for the story they are wishing to tell."

4.2.4 Required Functionality

All interviewees were asked to consider the proposed functionality of a dataset knowledgebase.

'A 'dataset knowledgebase' is an online service that would provide efficient two-way links between source and output repositories. This service is enhanced through the addition of features such as quality assessments or ratings, and answers to frequently asked questions (FAQs) about specific sets of data held in a repository. What is your opinion of the value of such a concept and are there specific issues you might want it to address?'

The majority of interviewees could see a need for FAQs, although two interviewees pointed out that it is not always possible to find answers to the specific question or issue that needs resolving. Two further interviewees commented that they felt a general or technical FAQs section would be useful whereas ones relating to specific analyses or subjects would be less useful. Conversely, one respondent felt that these could be very useful to tell you more specific details about the data (for example what selection criteria have been applied) which would often be omitted in a publication.

On the issue of quality assessments or ratings, again there was support by a lot of interviewees (although not quite as many as for the suggestion of FAQs). However, three interviewees were concerned that "what may be useful to one researcher may be useless to another". Further, two interviewees were concerned that quality assessments or ratings may cause important research being performed by small collaborations to be 'lost' amongst the work performed by larger ones. Two interviewees expressed the importance of making sure that any ratings were fair and that it would need to be transparent how they were determined. One researcher went further and said that "It would be good if quality assessments or ratings tied in with the RAE ratings". More positively, two interviewees thought that the concept of quality assessments or ratings could be most useful to newcomers, whilst one interviewee stated that they "would give you confidence in the data you were accessing".

A further question was asked: 'Some data repositories are open to all enquirers while others are password protected. If we are expecting to design links that will provide access from open repositories to controlled repositories, we shall need to devise some level of validation and temporary access rights. Are there any

authentication issues with regards your own source data? What degree of access protection would you expect?'

Most interviewees explained that their data has some level of access restrictions in place, although data that have been published or are ready for publication may not be so heavily restricted. Often access restrictions are an aspect of collaboration policy. A number of interviewees said that they required protection against people publishing before them and against people abusing their data, whilst a lot of interviewees considered that the same (or some) level of access restrictions would be required.

> "I can see issues where making source data available could be abused: i.e. what's to stop people pinching your data and publishing it themselves – so would need to be protected against this."

> "It could be a dangerous thing if anyone can analyse my data, as this could allow a situation in which misinterpretation of the data could arise … so some sort of access restriction would probably be required or at least restrict what data is made available."

> "No particular issues, as no access outside of the collaboration. If there was linkage between source and output repositories, we would expect the same level of access protection that we currently have. The people that generate the data have some say or rights as to whether or not to make data available …"

A number of suggestions were made that may help solve some of the access issues. For example, two researchers suggested that the items of data that were made available could be restricted, whilst a further two interviewees suggested that access is restricted to users of selected laboratories, universities and institutions perhaps by using IP addresses or laboratory passwords. Two researchers felt that one could adopt some method of contacting data owners in advance to request permission to use their data. One interviewee suggested 'Certificate Authentication' [19] as a potential solution:

> "… Each user should have a certificate which would allow them to identify themselves to source repositories or websites …"

4.3 Source Data

The types of data produced by interviewees and the file formats in which they are held were a close match with the Physics questionnaire results shown in Figure 3.2 and Figure 3.3. In the case of the HEP interviewees, the raw data (that taken by experiment) is stored mainly at the laboratories at which the data is taken (for example CERN, SLAC and DESY) and, to a smaller extent, at computing centres. Analysis is then either done remotely at that laboratory or computing centre or some fraction of the data is downloaded to an individual PC or group server at that researcher's own university. The HEP community is also participating in developing the computing *Grid* [13] which is "a complicated system with data stored around the world". The Grid enables large amounts of data storage and huge processing power (for amongst other things, HEP data storage and analyses). For most of the interviewees outside of HEP, data is stored on CDs and the hard-drives of individual PCs.

HEP interviewees indicated that they produce vast amounts of data: raw data is in the gigabytes (10^9 bytes) to petabytes (10^{15} bytes) region, processed data (from which the Physics results can be extracted) is smaller in the region of tens of megabytes (10^6 bytes) to gigabytes. Other interviewees produced much less data of the order kilobytes to megabytes.

Following on from the questionnaire's 'why you might wish to access the research data generated by other research programmes' and 'how would you access the research data of other researchers', the interviewees were asked 'do they access research data from other collaborations'. The results were evenly split with 6 respondents saying they did not access others' research data and 7 respondents saying that they did. Of those

who did, the majority accessed data via the publications produced by other collaborations. One of these researchers said:

> "The data that is accessed is processed data that has taken into account measurement uncertainties and artefacts of their apparatus. I access this through peer-reviewed journals only. I would not ask members of a collaboration for their source data (as it would be of no use to me for the reasons outlined earlier [in the interview]). I would only contact individual members of a collaboration if I required further clarification of something that had been published or appeared in a conference."

2 researchers said that they have in the past had the need to access data directly from collaborations. One of these researchers commented:

> "I access plots to get an idea of what is going on. However, if you wish to make a serious comparison it is necessary to have access to the data points that make up the plots. This is where the problem comes in as the plots are fairly freely available, but extracting the data points and their errors is much more difficult. To obtain the data points and errors from a plot, you would usually have to write to the correspondence authors of the publication and ask for these."

The most popular reason expressed by interviewees for accessing others' data, now or in the future, was for the purpose of comparison with their own research results.

4.4 Source Repositories

In the questionnaire the topic of source repositories was largely restricted to the types and frequency of submission to source repositories by respondents. In the interviews the aim was to broaden out the discussion to include a more general look at how researchers use source repositories as well as their experiences of this type of repository. Consideration was also given to the enhancements to source repositories that interviewees would like to see. Accordingly, this section is divided into two subsections which will cover these themes.

4.4.1 Use of Source Repositories

All HEP interviewees work in large collaborations (typically 100 to 3000 members) that use source repositories (for example those at CERN [4], SLAC [9], DESY [8] and the computing Grid [13]). Of the 5 interviewees who were from other branches of Physics, 2 stated that they store their data in a central collaboration area (for example on a network or in a database) whilst the remaining 3 work in smaller collaborations that do not use source repositories.

It became clear that although many interviewees work in collaborations that use source repositories, personal experience of either submitting (also referred to as 'uploading') data to, or extracting (also referred to as 'downloading') data from these repositories was more limited. Almost half of the respondents said they had no experience of either process with the remainder showing some experience of either one or both processes. 7 interviewees had experience of downloading data, although the majority stated 'occasionally' or 'on several occasions' as the frequencies. However 2 interviewees stated they 'downloaded data frequently'. The operation of uploading data was less popular with only 5 interviewees citing personal experience and that in all cases was limited to a frequencies of 'occasionally' and 'rarely'. One explanation that may explain the slight inclination towards downloading data (certainly when discussing unprocessed data) is that most researchers would need to download the unprocessed data to perform some analysis, whereas putting that unprocessed data onto the source repository in the first place may be limited to a smaller number of people (or it may even be deposited there automatically, directly from experiment). However, what is clear from the interviews is that source repository usage is not confined to just 'submitting to source repositories' as was concentrated on by the questionnaire, thus those researchers making use of source repositories may be greater than the results of the questionnaire suggest.

Interviewees who had made use of source repositories were asked to comment on their experiences. Most of the comments were on one of two issues: ease of use and documentation. On the issue of ease of use, three researchers felt they are fairly easy to use, whilst a further three researchers suggested that they could usually find what they are after when consulting other members of their collaboration who know how and where (within the source repository) the data of interest is stored. One interviewee said that they are "quite cumbersome to use". On the issue of documentation, two interviewees felt that the information (or documentation) about what the repository contains and where everything is stored within it is often limited, or badly described whilst one interviewee felt that they are "quite well documented". One researcher made the following comment on their experience of source repositories:

> "Sometimes I find that the information about what the [source] repository contains and what is available to you is limited, or badly described. One would usually have to ask someone else because the documentation is poor."

One researcher made a more general comment about source repositories:

> "They are not meant to be user friendly – but functional for members of the collaboration."

4.4.2 Enhancements to Source Repositories

Interviewees who had gained some personal experience of source repositories were asked what enhancements they would like to see made to this type of repository. Given here is a selection of these suggestions, the first two of which follow on from the experiences of source repositories detailed in the previous subsection. Three interviewees felt that better (or more logical) ways of accessing data within the source repositories would be good, with one of these researchers saying: "they can be a bit bewildering to new users simply because of the large amount of information they contain." Three interviewees felt better documentation would be a much needed enhancement.

Two interviewees made the following comments on two differing themes, which illustrate enhancements to identification of material held within source repositories as well as items, that source repositories could hold, which would assist with the dissemination of research tools:

> "Currently, the same sample can exist under three different names. It would be useful for each sample to be unique. Therefore better identification, so every sample will be unique would be an improvement."

> "It would be good if non-commercially developed software and programs could be shared more and made freely available as this would save a lot of time and repetition of work that has already been done."

A number of technical enhancements were also mentioned, including two interviewees who cited the need for source repositories to be well maintained and kept up to date, thus ensuring the usefulness of the repository and the longevity of the data stored within it. One interviewee commented on the need for faster retrieval times: "If data is too large to store on a 'primary' storage media (PSM) it would generally be stored on a 'secondary' storage media, and then loaded onto a PSM when required – which obviously takes time." Other technical suggestions included more reliable servers and technical infrastructure as well as better software.

A number of enhancements to source repositories that were suggested in the questionnaire (not necessarily by Physics respondents) were put to interviewees. It should be noted here that not all of the following suggestions were put to all interviewees, instead a selection were posed depending on the answers the researcher had already given and the time constraints of the interview.

The first suggestion was 'links between different source repositories (for example between CERN and BNL)'. Two researchers felt this could be useful, one felt it would not whilst a further two researchers stated that this was already being done, for example by the Grid. The second suggestion was 'source data organised in date order so as one can map the development of that source data'. Again there was a mixed response with three interviewees stating that this might be useful or was already being done, whilst two interviewees were unsure of the usefulness of the suggestion. One interviewee stated that they:

> "Would not want to show the full evolution / development of the data, for example publicising mistakes and all the steps taken would not really be beneficial …would only want selected things made available."

The third suggestion was that of 'including the background details to an experiment with the source data'. Interviewees were in general agreement that this would be useful: one interviewee commented that they "would hope this would come by default". The fourth suggestion of 'better searching facilities within source repositories' suggested that often searching in the sense of not knowing what data you require is unimportant as generally researchers would know in advance what data they require. However, it is the ability to be able to find quickly and efficiently, within the source repository, that data which you require that is important. One interviewee stated that this is "essential to enable one to find the data, otherwise the data is useless." The final suggestion posed to interviewees from the questionnaire was the 'ability to be able to keep a record of who uses your research data'. This enhancement was received positively by all researchers asked this question and is summed up effectively by the following two quotations:

> "This is a good point. If you produce data and then had the ability to trace who had used this data that would be a good functionality and good way to see the use of your data or work."

> "We need to know it at the level of providing assessment of the quality of research … It is important to know how often people are using our work and this is useful to us in convincing funding agencies of the quality of our research. It doesn't directly help us day-to-day, but from time-to-time, we need to provide evidence that what we are doing is of interest."

4.5 Metadata

It was clear from the interviews that the term 'metadata' is unfamiliar to many Physicists. However, the majority could see the importance of assigning data to their source data to allow its future identification. To this end, many were able to suggest terms and identifiers which they felt would be useful to assign to their data. It also became apparent during this process that the amounts and types of metadata that Physicists do (or would like to) assign to their data may vary depending on the level or stage of analysis. For example raw data may require different metadata to that of the processed data:

> "Depends on what type of data you are looking at: as metadata requirements for data at different stages (for example, raw and analysed data) would differ … For data being 'transformed' from raw data taken by the experiment to data that can be used for analysis, it was found that things like 'date' were not that helpful during these intermediate stages, whereas things such as processing specification (such as release number of code used to process it) and 'run number' were much more useful."

Researchers were asked to confirm their responses to the questionnaire of 'what metadata do you consider it important to assign to your data'. Responses were distributed fairly evenly amongst the eleven generic options given in the questionnaire, although 'project reference number' and 'funding source' were seen as being fairly unimportant.

There was a mixed response amongst researchers when asked 'do you think these suggestions are sufficient on their own to make your data meaningful to others' with one respondent saying:

> "For anything other than the journal derived from the data, a huge amount of additional information is required to make meaningful interpretation."

Of those who felt additional metadata would be needed, there was no one generic solution that was suggested, although the need to specify temperature and pressure under which the experiment was performed were cited by a number of researchers. Most other suggestions were largely specific to the various fields of Physics interest. Considering three fields of interest, the following metadata were suggested: run, event, beam conditions (High Energy Physics), details of sample such as how much material it consisted of and sample thickness (Nanoscience), the molecule, surface and x-ray energy used (Surface Physics). One researcher made the more general comment when asked about additional metadata requirements: "more specific things to the experiment".

A number of additional metadata were suggested in the questionnaire that respondents from across the seven disciplines felt would be useful. A number of these suggestions were put to the Physics interviewees. (It should be noted that interviewees were generally only asked for their thoughts on a selection of these suggestions, depending on the answers they had already given and time constraints of the interview.) The first suggestion was that of 'metadata describing how the data was generated' to which two thirds of interviewees were positive, or thought that this would be there already. The second suggestion was that of 'metadata describing the software which was used (along with version number) which was used either to generate or analyse the data'. This was less well supported with a range of answers ranging from 'yes', 'perhaps in certain cases' through to 'no'. Two other comments include:

> "Our raw data consists of a text file so it is not so important on which platform it has been created in: i.e. the file can be viewed in Windows, Linux and so on. I personally make it deliberate policy to ensure my data is stored in a platform independent way so as people can read it using any platform."

> "For HEP this would be useful, but if you were not part of the collaboration knowing which software (whatever version or release) would not be that useful, as that [specific, collaboration written] software would not be available to you."

The third suggestion was that of 'metadata detailing instrument or apparatus details'. The majority of interviewees asked about this suggestion were positive with two saying "definitely" and "essential". One further respondent explained why, in their field, these details were essential:

> "Essential. Apparatus is always being updated, revised and improved over the lifetime of an experiment, so you have to know which bits of apparatus were in and in what state (i.e. efficiency of detector, part of detector may be dead etc.) of operation they were in at any time in order to interpret the data correctly."

The final suggestion was that of 'metadata detailing other relevant data sets'. Of the few interviewees asked this question most were positive, one researcher said:

> "These links probably do exist already: perhaps though as information (either inside the data and / or in a log book) and perhaps not as a physical link … so making a physical link would be useful."

4.6 Data Access and Sharing

Under this heading the questionnaire focused on methods that researchers currently adopt to make their research data available to others, encouragements for and discouragements against sharing data as well as the formal restrictions and access controls applied to their data. It was felt that these topics were well addressed in the questionnaire and so the opportunity was used in the interviews to expand on the issue of data sharing in general and in particular on the practicalities of making data available for use by others.

In principle many interviewees were not necessarily against the idea of sharing data, but many had reservations about the practicalities of doing so in their field. Further, many interviewees felt that their 'raw', 'unprocessed' or 'primary' data would be unusable or, at best, of limited use to other researchers outside of their collaborations. However, researchers were more positive about the usefulness to others of their source data at some higher level: for example after some processing and / or analysis had been performed and the Physics results extracted. A selection of the many comments received on this subject is presented below, beginning with the ones commenting on the unprocessed data:

> "It is not generally regarded as good to make the primary data available. These are generally regarded as property of the collaboration and indeed are very difficult to interpret without all the information that goes with it (e.g. running conditions)."

> "In principle, I believe that the data should be available to all, but practically, I don't believe anyone outside of the collaboration would touch it. For example, in order to understand the raw data you need detailed knowledge of the specific analysis techniques of the experiment."

> "I could make my data available, but without access to the various simulation tools, software tools and detailed background knowledge of the experiment it would be extremely difficult to do anything useful with the data."

> "My data would be absolutely useless to other people. In principle, I am not necessarily against sharing, but my data would be useless to anyone else."

> "The data is meaningless unless one knows what is going on."

The next three comments show enthusiasm for making the processed data (from which final Physics results are extracted) available:

> "Data that has been analysed and is being prepared for publication would be useful to other people. This includes a whole series of plots and distributions showing the behaviour of the data when looked at under different variables. Data at this level will have all the corrections, that can distort the data that generally only the initial collaboration know about, taken into account and should then reflect the physics of the situation rather than the way that the apparatus worked. It is usual to try and make this information available to others in an efficient a way as possible: but this is often not that efficient and a better way of circulating this information would be good."

> "At some later stage, once some processing (analysis) had been done on the raw data and the physical quantities have been extracted, then that data could be useful to others."

> "It would be very useful to provide some level of my data (not necessarily the raw data, but something more than just the final plots) to others who were wishing to cross-check one of my results."

Besides the issues of ownership, interpretation difficulties, requirement of additional information in addition to the data, specific analysis techniques and availability of non-standard software, complicated data structures and knowledge of the apparatus or experiment that have emerged from the above quotations, a number of other issues on the theme of data sharing (and particularly unprocessed data) were cited by interviewees. Three interviewees commented on the huge amount of time that would be required by anyone trying to analyse their data (if it were made available to others) whilst two interviewees suggested that it would be essential to join the collaboration. One interviewee summed these two points up by imagining they were accessing the data of another researcher in their field:

> "In practice, for me to understand the data from another experiment would be a huge amount of work. I would have to join the collaboration to get all the information necessary to analyse the data: without doing this an outsider wouldn't know where to begin with the analysis."

When asked about the ease and amount of their source data which could be used by other researchers, rather unsurprisingly nearly half of the interviewees thought that their data could not be used easily by others. Three interviewees thought that their source data could be used by others who were familiar with the analysis techniques of their field and one interviewee was confident that their data could be used easily by others. As a percentage, most respondents who were asked felt that at most only 5-25% of their source data could be useful to others.

More generally on the subject of data sharing, concerns were expressed with regards secrecy and sensitivity issues:

> "There are secrecy issues / collaboration restrictions on the data which apply before the outcome of the research is known and the results published. It is normal policy in my field that prevent the sharing of data until 10-15 years after an experiment has finished. However, it is true that research is enriched when information is shared."

> "The data is commercially sensitive and we need to check that the 'intellectual property' is in place. This is particularly important when working with industry."

Further, a number of interviewees expressed concern about ensuring the reproducibility of results:

> "Our collaboration makes every effort to ensure the reproducibility of our data: we therefore need to select carefully what we wish to make publicly available."

> "There is an interesting issue here that if someone from outside of your collaboration was also analysing your source data – what would happen if they obtained different results to you?"

4.7 Output Repositories

From the questionnaire it was clear that Physics researchers make use of a wide range of output repositories. Whilst the questionnaire concentrated on the kinds of output repository used for research and teaching, as well as the preferred routes to their contents and the levels of searching that researchers usually undertake within them, the interviews looked to expand the discussion. The interviews aimed to gain an understanding of what makes a 'good' and 'bad' output repository as well as what factors attract a researcher to a particular output repository when looking for somewhere to deposit their research outputs. As in the case of source repositories, consideration was given to enhancements to output repositories that interviewees would like to see. Accordingly, this section is divided into two subsections which will cover these themes. It is also worth noting at this point that during the interviews there was considerable discussion about 'Open Access', details of which are included as additional observations in chapter 5.

4.7.1 Use of Output Repositories

Interviewees were asked to give examples of each type of output repository (publisher, discipline and institutional) which they used and if any stood out as being particularly good or bad. Three output repositories that were considered by interviewees as 'good' were the arXiv pre-print server [15], SPIRES [16] and PROLA (Physical Review Online Archive) [20]. These were cited for reasons of price (arXiv which is open access and PROLA which has a low cover price) meaning they are available to a large, worldwide audience and searching facilities (SPIRES):

> "It is an excellent example of a search facility – very easy to use, both for experts and novices …"

Two output repositories that were cited as 'bad' were Journal of Physics (J. Phys.) [21] and the CERN Document Server (CDS) [17]. The reasons that were stated for these choices were converse of those above, with high cover price meaning a limited audience for J. Phys. and a poor search facility on CDS:

> "The search facility on CDS is not as good as it doesn't have a plain-English interface. The data may be there, but the interface makes it difficult to find."

It became clear from the examples given for institutional repositories that there was some ambiguity in their understanding of the term, with many interviewees citing examples of university websites, university libraries and publisher online journals accessed via the websites of university libraries. It was thus suspected that the numbers of questionnaire respondents (shown in Figure 3.14 and Figure 3.15) who selected 'institutional repository' may well have been working from a different definition than that adopted by this project.

Interviewees who had personally deposited material into output repositories, or had been involved in the decision process of which repository to submit to, were asked 'what factors attract you to a particular output repository?' The two answers to this question which were received the most (6-7 interviewees stating both of these answers) were the relevance of the article to the journal it was being submitted to and the hierarchy of impact factor rankings. The following are two quotations reflecting these themes:

> "In these days of RAE, we would try and publish in the highest impact journal that is relevant to the work."

> "The impact factor is very important - RAE ratings are very important - and how respected the journal is. It is also important to consider how relevant the article is to the journal."

The circulation, or audience, of the journal was also seen as being important:

> "I like my articles to be placed where they will get a large audience."

> "Work in poorly circulated journals can get lost."

Continuing on the theme of availability, subscription and publication costs were an important factor to the decision of where to publish for some interviewees.

One interviewee cited that where to publish was down to policy of the laboratory in which the experiment is based.

Many interviewees when asked 'would you be happy for your own research outputs to be placed in an institutional (open access) repository' were positive towards the idea, although many would want some pre-conditions (mainly associated with open access) addressed before doing so. The topic of open access and deposition in open access repositories will be discussed in detail in the next chapter.

4.7.2 Enhancements to Output Repositories

As was the case for source repositories, interviewees were asked if there were any enhancements to output repositories that they would like to see. Again, this created lively discussion and a large number of suggestions were made. A few general issues included: three researchers who felt that it would be useful to make the word-processed file in which the document was written available in the output repository as well as the final portable document format (.pdf) or postscript (.ps) formats that most output repositories make available. Over two thirds of the interviewees commented on some aspect of the larger issue of 'article access' with researchers suggesting the digitisation of older or rarer articles as well as full access to all journals and, if this were not possible, a pre-print reference for all citations appearing in the 'references' sections of articles. One interviewee thought that direct links to (or the addresses of) depositors of articles would be useful. One researcher commented on the current limited functionality of publisher repositories:

> "Functionality in publisher repositories is currently limited to retrieving a specific article or going to a specific issue – and that's it! If you have a specific paper by a particular author and want to see other papers that he or she has written, this is difficult."

This last point links in well with the functionality issue cited by most interviewees as being of major importance to users of output repositories: ease of searching. Four interviewees said that a good search facility was either "important", "very important" or "critical". Nearly half of the interviewees made some comment about what constitutes a 'good' or 'bad' search facility. The following are a representative selection of quotations:

> "Knowing which keywords are most successful in returning the results you are after."

> "The pre-print output repositories are good for finding known pre-prints, but their keyword search to find relevant pre-prints, that you are not aware of, is limited."

> "Simple searches are often frustrating or insufficient – I would like a Google type search."

> "I think there is a need for an intermediate level search engine. Google is good, but if you are looking for scientific output it returns many things that you don't want. Web of knowledge tends to be very specific and only returns material in publications (no other media): so something in between these two extremes could be useful."

> "Perhaps better linkage in Google Scholar."

> "Better linkage when searching for publications. For example I currently will look in an output repository for a publication. From the publication that I have obtained, I can look at previous work (reference section). However, to look at work that has been done since, I would go to the Citations Index (to see what articles have been written since that have cited the work done in the publication I have found). I would then need to search again to find these later articles. If there was some better way of being able to search for articles both forward and backward in time, from the article that you have found then that would be helpful."

> "I would recommend that if StORe is wishing to create a search facility they look at SPIRES, which has been evolving over 20-30 yrs. There is a pro-former type search facility where you can 'click' on criteria or a very simple (almost plain English) language where you can say (for example) 'find author X, title Y and year after Z'. If the SPIRES interface were emulated, then that would be a very useful interface …"

> "…One of the problems with SPIRES and the Durham HEP database is that the way that you enter searches, is very painful / clumsy procedure and every time I use it I have to spend about half an hour looking into how to send a proper query."

Whilst the two latter quotes appear in conflict with one another when talking about SPIRES, it was felt that researcher opinion may vary according to frequency of use.

As was done for source repositories, a number of suggestions that were made from the multi-disciplinary questionnaire on the subject of 'enhancements to output repositories' were raised amongst the interviewees: again the numbers asked each question depended on earlier responses and interview time constraints. The first suggestion which asked about 'making the contents of tables in publications available in a machine readable format (i.e. not .pdf or .ps). Here, virtually all of the 13 interviewees would like this feature, although there was a wide range of suggestions for the best format. The second suggestion concerned the references at the end of an online article, and 'would hyperlinks to referenced articles be useful'. Again, virtually all of the 13 interviewees would like this feature, with 5 saying they already use such features in the output repositories that they use. Two researchers commented:

> "This would significantly reduce the amount of time that you would need to look or search for literature."

> "This would be fantastic, although I suspect there would be subscription issues. Although, the technology must be there to check whether the university had subscribed to the journal of the article that you were linking to, so this may not be such a problem."

The third suggestion asked if it would be useful if there were 'an online area where you could store links to journals or articles that you frequently use (to save having to download or continually navigating to them through online journals)'. Nine interviewees said they would like this feature, with one saying explicitly that they would not. Three interviewees said that they were happy using the bookmark facilities on the web.

Finally, on the theme of enhancements to output repositories the following question was asked: 'A number of new operations could be supported within an output repository, such as the automatic creation of links between related resources and the presentation of relationships (i.e. showing publications and their source data in adjacent windows). How do you think these could meet your needs?' Eight of the interviewees thought that this sounded interesting or could see some use in the ideas, three High Energy Physicists explicitly stated that they thought in their field this would be 'difficult' or 'not that useful' due to the issues discussed in section 4.6 whilst two interviewees thought this was 'not important' or could see no use for these things. Four researchers commented thought that these it may be technically challenging to make and implement all the relevant linkage with one interviewee saying:

> "I think this would be difficult to do generally; i.e. to find something general that would make everyone happy. Probably better to build something that would allow users to define what they wanted to show on the screen."

4.8 Support

It was clear from the interviews that Physics researchers generally seem happy with using output repositories. When assistance is required it appeared that the majority do not make use of help offered by librarians or other information professionals. All those who commented on such help said they only use it very rarely, with three researchers saying they had "never needed any formal assistance". Two respondents commented that in the past they have used it more frequently, but in more recent times they are more likely to use online assistance:

> "Not really. We used to use the librarians for all searches, but now we use 'Google' or 'Web of Knowledge' …"

> "I use it very rarely. I used help from the library in older days on several occasions for information about citations, but more recently this is available online. Over the last few years my first attempt would always be to look on the web where I usually can find all the support that I need."

This last quotation mentions online support as the method currently used when assistance is required. In general, online support is very popular, cited by many researchers as their preferred choice when support is required. Other forms of independent support that researchers cited were "help from colleagues" and "e-mail assistance". One researcher said that the need for e-mail assistance could be eliminated (or at least reduced) if the online help was better.

Other comments about the help provided by librarians and other information professionals, or help that researchers would like provided by these people include:

> "For finding things in output repositories, there is Athens help offered by the library … There is also a lot of support provided for making grant applications."

> "There is little support for submission of publications"

> "Some training or support to explicitly show you what features were available within a specific output repository may be useful."

Finally, one respondent commented on the importance of trained librarians:

> "… It is good to have trained librarians as a fall-back. On the few occasions I have required their help they have always been very helpful and it is important to have them there."

5 Additional Observations

This chapter provides information from the Physics questionnaire and interviews that has not been presented elsewhere in this report. Information presented here derives from 'free text responses' to the questionnaire and those issues addressed in the interviews that are beyond the specific remit of this project.

5.1 Project Aims

Section 5.1 covers the free text responses from the questionnaire that pertain the initial views about a bi-directional source to output link, as well as discussion about the Durham HEP Database which offers its own form of source to output linkage.

5.1.1 Free text responses of questions 2a and 3a

The free text responses from both questions 2a and 3a ('please use the following free text box if you wish to expand your answer') for source to output and output to source linkage respectively are detailed here. The comments received were in a lot of cases applicable to both directions with many explicitly saying (for example) in 3a 'see question 2a', 'see previous comment' and 'for the same reason as in 2a'. For this reason all comments from 2a and 3a have been grouped together and organised according to the main theme of their comment. Several comments cover more than one theme. Where this is the case the comment will only appear in the category considered to be most appropriate. Comments generally fitted into one of six broad categories: the first two were those generally in favour or against such a facility whilst the remaining four offer practical issues on the themes of data, access, specialist knowledge and additional information required. The results are shown in Table 5.1 where, for completeness, the actual question number for which the answer was made is given: where labelled 'both', the respondent entered some comment to 3a that refereed the reader to their answer in 2a. Many of these responses reinforce issues mentioned in the previous two chapters of this report.

Table 5.1: Free text responses to questions 2a and 3a

	Comments	Q
1.	**Can see benefits of this type of operation**	
a.	In this case I imagine that a source repository would be either access to the CFD results, or spreadsheets containing the (annotated) data of the simulation itself. Having this as a backdrop for comparison and examination would be very handy rather than having to rely on small graphs or reams of data that occasionally get printed in journal articles.	2a
b.	It would save a lot of time and hassle	3a
c.	It would be very useful in order to compare datasets	3a
d.	It'd be great but I don't think it's going to happen	3a
e.	Would be useful, especially if say graphs/figures were available as the original tables of data	3a
f.	Experimental details of extra data which was unpublished would be more beneficial. Expansion on equations via appendices would be excellent	3a
g.	Comparing experimental data with calculations or theoretical graphs or other experimental data would be very useful	3a
h.	I can only think of this feature being useful for purposes of comparison of results, in my research field	3a
i.	At best it would contain a link to the experiment/project web pages	3a
2.	**Cannot see a need for this type of operation**	
a.	High energy physics collaborations generate data, which they then analyse. They also list the publications using the data, so I have this already.	2a
b.	The specific area I work in at the moment is quite small, so a significant amount of interactions takes place between the PIs involved via email, and it is usually possible to access the raw data from experiments this way.	2a
c.	Relevant data in my field is always published in the article. There does not seem to be a need for a separate source repository.	3a
d.	I cannot use the 'raw' data from any collaboration but ones I am already on and so already have the data.	3a
3.	**Practical issues: data**	
a.	The data gathered in particle physics is usually of extremely large size (tera to peta bytes) and gets filtered and reduced through very many levels the last of which are 'private' and do not show up in repositories. It is thus impractical to follow links to primary (raw) data. Even if this could be done, access to these volumes would be prohibitively expensive and time consuming	2a
b.	Experimental data is meaningless without further processing. The same applies to data from computer simulations	2a
c.	The LHC produces many PB of data. It is refined by various processes (Reconstruction, event selection, analysis) but any particular publication will reference all this (because there are calibration constants involved) together with simulated data of equal or greater volume. Thus the notion of a 'source data repository' makes not sense to me: The source data is an enormous collection of many PBs of data.	Both
d.	Often I use repository data that have not generated a publication on their own, but have significance in combination with other data.	2a
4.	**Practical issues: access**	
a.	I don't see how this would work in my field of research as the data is generally not available to rival experiments.	2a
b.	Particle Physics collaborations don't usually make their data publicly available, so the current work strategy wouldn't benefit from it. It is hard to say how much that strategy would change, and how beneficial it'd be, if the data were to be publicly available	Both
5.	**Practical issues: specific knowledge required (e.g. that of the experiment, software)**	
a.	In experimental HEP, data themselves cannot be meaningfully used without a very large software infrastructure and detailed understanding of the apparatus.	Both
b.	For the source repository to be of use, one must know its format and how to analyse the data it contains. Without a common framework for analysis, I cannot see immediately the usefulness of linking source data with the corresponding publications.	Both
c.	All detector systems have systematic errors and bin migrations. Raw data needs to be processed by experts in each system to become usable.	Both
6.	**Practical issues: additional information required**	
a.	Difficult to say a list of numbers without details of methods could be useful but might be misleading.	Both
b.	In my field, you would need more than just the data and the paper – some technicalities of how the data were treated would need to be determined.	3a

5.1.2 The Durham HEP Database

In the interviews, researchers were asked if they knew of any services that currently offer the kind of source to output (and vice versa) linkage proposed by Project StORe. One interviewee thought that 'Nature' [22] may offer a similar service. The other service that was mentioned by four researchers, during the course of the interviews, was the Durham HEP Database. The Durham HEP Database is a web mounted service that enables members of the High Energy Physics community to 'gain easy access to a wide variety of data and information on particle physics, ranging from literature searches to actual data' [12]. The service is deemed useful for particular aspects of a programme of research, which one interviewee described as follows:

> "I use it when writing a paper or I wish to find information about what other people measured or try to find out if someone else has measured something similar to what I am measuring."

Most of the four interviewees who were aware of this service claimed only to have used it 'occasionally' or 'not at all', but a number were aware of colleagues who did use it more frequently. However, most stated that it is (or potentially is) a useful facility to have. Possible explanations for a relatively low usage may include the fact that it is only considered useful for a small part of a research programme (i.e. nearer the end of an analysis when comparing results and writing publications) and the fact that at present the data which it holds is limited. One researcher commented on this:

> "An attempt to make tabulated points (that make up a plot) available is being made by the Durham HEP Database, but this is patchy - I have only relatively rarely found what I want there. This patchiness is a problem and I would say that only ~20% of the available measurements ever make it onto the database. The Durham HEP Database could be improved by being less patchy and having a more complete record. If there was one that has the same level of coverage as the pre-print databases so as you would stand a 70-80% chance of finding what you were after that would be very very useful."

5.2 Source Repositories

The free text responses from question 26 are presented in Table 5.2: 'Having considered your current use of both source and output repositories, and the potential relationships between the two, what functionality if any do you consider is missing from the 'source repositories' that you have used?' The responses are categorised according to comments that are directly relating to functionality that is missing / enhancements researchers would like to see and those that fall outside of this remit. In addition, there were 13 other responses to this question which all stated non-use of source repositories or the question was 'not applicable'. For clarity, this latter group of comments are omitted from the table.

Table 5.2: Functionality researchers consider to be missing from the source repositories that they have used

	Comments
1.	**Missing functionality / enhancements researchers would like**
a.	More meta information
b.	The main problem is that the repositories are far from comprehensive, so the probability of a successful enquiry is rather low.
c.	'Open source' style access to the programs or detailed equations and derivations from more computationally orientated papers. These background details can be of extreme importance to peers and colleagues in the research area. They act as accelerants to research and are now only rarely available on group websites - they should be sent to the repository at the discretion of the publishing author(s).
d.	Wider cross-reference
e.	Inability to link data in different formats (e.g. raw data, paper produced, plots, and numerical entries for the plots)
f.	Uniformity of format for uniformity of data.
g.	I believe that the semantic web [23] will provide most of the functionalities I require.
h.	I can't see any
2.	**Other Comments**
a.	Fine!
b.	I am currently satisfied
c.	The only source repositories I've really used are from data within our group - data external to our group is rather hard. There doesn't appear to be a pool of data - this has to be requested from the individual authors themselves usually.
d.	They don't really yet exist in my field. I recently applied for a grant which included an aspect of setting such a thing up for our work, but this part was not funded.

5.3 Output Repositories

This section is divided into three subsections: the first two detail the results of two free text responses: the third subsection presents the issue of 'open access' output repositories, as discussed in the interviews.

5.3.1 Searching in Output Repositories

On the issue of searching in output repositories, respondents to the questionnaire were asked (question 23a) 'What further options would enhance your level of searching?' The responses made by the seven researchers who answered this question are given in Table 5.3. It can be seen that these comments are quite wide-ranging and cover some interesting points.

Table 5.3: Further options that would enhance searching

	Comment
a.	A common search scheme, e.g. Science Direct and Ingenta Connect use different keyword and Boolean logic schemes. It's a bit pedantic, but it's still irritating.
b.	I think it is very important that a standard is established to allow citation information metadata to be added to pdf academic papers - such as the ID3 tags on MP3 files. Whilst the level of conformity of citation info across academia is not as uniform as for MP3s, the addition of Authors, Title, Citation, Abstract and DOI to the paper in metadata format, would greatly enhance my ability to manage documents.
c.	Ability to search within articles quickly.
d.	Grouping results by author (where for example there is more than one physicist with the same surname)
e.	I have often been very frustrated by simple (non-Boolean) search engine
f.	Quick keyword search
g.	Full text searching, date tagging, knowledge space coordinates

5.3.2 Enhancements to Output Repositories

Table 5.4 shows the free text responses given to question 26a: 'And what functionality if any do you consider is missing from 'output repositories' that you have used?' This is the complementary question to question 26, which addressed the functionality of source repositories. These free text responses can be classified according to those who were currently satisfied and those suggesting enhancements. The enhancements proposed fit broadly into the themes of access, linkage between output repositories, metadata, linking to source repositories / source data and 'other'. It is encouraging that three respondents have made specific comments to the linkage between source and output repository in this question. There were four other free text responses that have been omitted from this table: three of which stated 'none', the fourth stated 'Don't use output repositories really'.

Table 5.4: Functionality researchers consider to be missing from output repositories that they have used

Comments
1. Currently Satisfied
a. Everything which we need in our subject can be found in the ""SPIRES"" database, see: http://www.slac.stanford.edu/spires/hep/ This just works and is very flexible, therefore for our particular subject, there is not need to develop anything else at present.
b. They are OK (2 respondents)
c. Generally fine, well maintained, easy to use.
d. I am currently satisfied.
2. Access enhancements
a. It would be useful if all historical papers were available electronically.
b. Full access to all journals
c. Limited coverage of some journals electronically
d. Some output repositories involve costs - they should be free of charge.
3. Enhancements in linkage between output repositories (perhaps using a single interface)
a. A thorough, single inter/intranet site where links to all subscribed journal publishers are kept in a straight forward, searchable, concise format. Or something like that.
b. It would be useful to be able to use a single interface/web portal.
c. Wish List: Ability to link between publications from different publishers. The Google/CrossRef effort is a good start but if references from each publications could link to others from different publishers or to the arXiv service if a preprint is available would be good.
4. Metadata enhancements
a. More meta information
b. Metadata in the pdfs
5. Enhancements by linking to source repositories / source data
a. Links to raw data in source repositories
b. They only contain papers: if you want the plots, you have to go and look somewhere else! Also, even if the plots are made available, the numerical entries for the plots are not (usually). This would be useful.
c. Always need access to raw data (i.e. graphs / plots on their own not good enough).
6. Other enhancements
a. An online area where you could store, if not the files then links to file which you are interested in.
b. Uniformity of format for uniformity of data.
c. For publishers: estimate the publication date of new articles For institutional: a tool for retrieval of bibliographic information. Example: once a set of publications have been found, I would like to have the entries in bibtex format

5.3.3 Open Access

In the interviews, two questions were asked on the issue of 'open access' output repositories: 'In general what are your feelings, if any, towards open access' and 'Would you be happy for your research outputs to be placed in an institutional (open access) repository?'

The majority of interviewees thought that open access was a good or even excellent idea in principle, although many expressed some reservations. Perhaps the most significant of these which was cited by over half of the interviewees, was the need to ensure the provision of adequate 'peer-reviewing' / 'refereeing':

> "It would be useful. The only problem is that they will not necessarily be peer-reviewed. If peer-reviewed this would be fine, as long as the peer-review mechanism is good enough for the community to trust."

> "Sounds like an excellent idea, as long as the same standard of peer-reviewing is applied: this is critical or else any old garbage could be placed on them."

> "I am not interested in this if there is no refereeing, as this is one of the main points of the publication process. Open access is fine if the same high level of refereeing was in place."

> "This seems reasonable and a step forward if articles are still refereed and the same editorial standards are maintained."

> "… whilst it is good for anyone to be able to access work, I think in science it is important to have the work peer-reviewed. I think the issue of who pays (i.e. the one who deposits or the one who accesses) is not as important as the issue of getting the work peer-reviewed …"

> "In principle open access is a great idea and I would much prefer it over subscription based services (to remove financial barriers to work). However, I appreciate the peer-reviewing has to be paid for somehow and that the peer-reviewing is essential to maintain the standards of research."

Two of the researchers who expressed a concern about peer-reviewing did, however, say that if an open access output repository were not peer-reviewed it would be important that this fact was made transparent:

> "In my opinion, open-access publications must be peer-reviewed or, if not, it must be transparent that the work is not peer-reviewed. I would also like to see the affiliations to which the depositors belonged. Need to have a standard of work so you know if the information is good or not."

Other concerns included the time it would take for a new open access output repository to become credible, as well as issues of cost and the practicality of starting new open access journals:

> "I have no experience of using these yet. The problem with them is that commercial journals have a highly valued refereeing system: which counts a lot with funding agencies when you are demonstrating the quality of your research. I am not sure how refereeing is done in open access journals, but nevertheless it will take time to establish credibility to the same level. Having said that fashions change and people change their publishing habits and publishing in some different journals over time, so new ones can come and others become less popular. If there was a 'mass' trend to move to open access so as the credibility was good and people would be tempted to go there to find things on that particular subject, then it would work."

> "Would perhaps mean a greater cost to individual departments as they would have to pay for submitting articles, but I feel this would be a reasonable cost."

> "This [Open access] is a good idea. Currently, in some journals you have to pay a publication charge to submit an article and a subscription charge to access the article, so you are paying twice!"

> "Starting new journals specifically so as they are open access is not necessarily a selling point for me, but making existing journals open access would be good."

Of course the issue of open access leads naturally onto the subject of institutional repositories, and this last quotation is one that was echoed by two further interviewees when asked about institutional repositories:

> "I cannot really see any advantages of each institution having an institutional repository …I would not automatically think to look in an institutional repository, would look in other places first"

> "I feel that every university or institution having its own output repository is not terribly profitable. I would not go to the output repository of University X, unless I knew there would be a high quantity of research outputs placed there … I am not so worried, where things are stored: the important thing is how I can get to / find the information … I Feel particle physics is well catered with as it has a fairly mature, relatively well-organised set of output repositories. Institutional repositories would probably be extremely useful in other branches of Physics, particularly those that don't collaborate on a worldwide scale. This would also increase the visibility of those researchers."

Continuing on this theme another researcher commented:

> "It would still not be quite what we want as, because of the nature of the collaborations, the amount of material coming from a particular institution would be relatively small (unless somewhere like CERN), so people wouldn't naturally look in a University repository for them – if this were the only place in which you published, work could get lost."

When asked 'Would you be happy for your research outputs to be placed in an institutional (open access) repository?' the majority of interviewees showed enthusiasm and were not necessarily against the idea. However, many of the issues and concerns about open access detailed above were mentioned again as things that they would want addressing before they would be happy to submit their own work to such a repository. A number of additional concerns were expressed, a selection of which are now presented:

> "Yes: however, I'm not sure that I would want an open-access repository to be the only place that work was published …"

> "It would be fine to put 99.9% of my output on an institutional repository: although there may be an issue with some of our projects that are funded by industry (however, with that work, if they were unhappy about it you would not be able to publish at all even through the normal routes). For things that go through the normal routes, I would be perfectly happy."

> "I would not publish in a journal just because it was new. I would need to wait and see how credible or how well used any new journal became."

> "Yes, if [the open access repository met] the criteria of (a) high impact (b) used by lots of people and (c) peer reviewed."

> "I feel everything should be made available. I am however, not sure how this will work in terms of copyright issues as the journals have the copyright (i.e. the author gives up the copyright when an article is published). They therefore, would probably not allow articles to appear as open access."

> "This might be a good backup if it were not too difficult to do. We can do something similar now … [but] we tend not to do this now as it is one more bureaucratic step. If it worked simply, with a similar protocol to arXiv that would be quite good."

The arXiv pre-print server [15] is one example of a discipline open access output repository that is widely used in the Physics domain. The impression gained from the eight interviewees who commented on arXiv is that, on the whole, Physicists like this resource. It is fairly well documented, it is well known within the Physics community and the fact that it is open access means it is adopted as a good resource for undergraduate students. However, that it is not peer-reviewed is regarded particularly negatively by some, despite the fact that a considerable amount of the material in it does appear subsequently in the peer-reviewed output repositories of publishers.

6 Summary

The results obtained from the Physics Survey of Researchers have been presented and discussed. There appears to be positive interest amongst the Physics community in the proposal for a bi-directional link between source and output repositories that can be regarded as an endorsement for the principal aim of Project StORe. However, this endorsement is not made without reservation and many researchers identified issues and problems with the development of such linkage. Perhaps the two most significant issues which need addressing are those concerning access to source data and what types (or levels) of source data should be made available to the wider community. The first issue covers both the themes of who should be able to access source data as well as the fact that a significant amount of source data in Physics currently enjoys restricted access which some researchers or collaborations may not be willing to giving up. The second issue reflects the fact that raw or unprocessed data is of limited use, in a lot of cases, to researchers outside of a particular field or collaboration whereas processed data at the point at which final Physics results can be extracted is considered to be more useful to others.

Besides gathering opinion toward the principal project aim, a considerable amount has also been learnt about the ways in which Physics researchers use both source and output repositories, and the functional enhancements they would like to see applied to both types of repository. Some of these would be easier to implement than others. For example ensuring an output repository has an excellent search facility may be more easily achieved than improving access to items held in subscription-based output repositories.

It is clear that the majority of Physicists make use of output repositories during the course of their research, whereas the use of source repositories is much more limited. High Energy Physics is a good example of one field in which source repositories are routinely used, although these tend to be private repositories restricted to a specific collaboration or laboratory membership.

It is interesting to note that there is some awareness within Physics of open-access output repositories, although knowledge of institutional repositories is more limited and a number of negative perceptions need to be addressed before many would be willing to deposit in them. An insight into the data that Physicists produce and use, the metadata assigned to these data, attitudes to data access and sharing as well as the support that Physicists use (and have available to them) has also been established.

Appendices

A Interview Questions

The interview script used for all thirteen interviews is given in this appendix. Whilst the interviews were highly structured (to allow comparison between different interviewees) there was some flexibility to omit questions where necessary.

A. Identities

A1. Discipline: Physics

A2. Employing Organisation:

A3. Role: Academic Staff, Research Assistant / Fellow, Postgraduate, Other

A4. Area(s) of Interest:

A5. Do you work as part of a collaboration? If so, approximately how big is the collaboration?

A6. Briefly describe the process of your research project / experiment from generating data → analysing data → publishing?

Further prompting to A6: How would a typical research process be for you? How is the data created, what do you do with the data you create? (How do you go about accessing data, discovering literature and how is this incorporated into your own research?)

B. Source Data

Source, or primary research, data is data that is produced during a programme of research and are used as the 'source' from which a publication is made (for example: instrument data, images, plots, spectra etc). Source data can be raw or processed and is likely to have a life cycle marked by change and evolution.

In the questionnaire you said that during the course of your research you produced the following electronic source data (*insert in advance from Q4 of questionnaire*):

In the following file formats (*insert in advance from Q5 of questionnaire*):

B1. Where do you store these data (for example: own PC, CD, source repository)?

B2. Typically how large are the files / electronic source data that you / your collaboration produce?

B3. In principle, do you think it is useful for your data to be made available to others?

B4. Could your source data be used easily by other collaborators or does it need specialist instruction / programs to turn it into something meaningful to others?

B5. Would adding some explanatory text to your source data address this problem?

B6. Approximately how much of your source data do you think would be useful to others?

B7. Do you access research data from other collaborations?

B8. If yes to B7: what type of source data generated by other experiments do you currently access (for example, tables, plots, numbers)?

B9. If yes to B7: currently, how do you access the source data generated by other research programmes?

B10. If source data from other experiments were more readily available:

(a) Could you see a use for it in your programme of research?

(b) What would be a convenient way to access it?

C. Use of Source Repositories

By source repositories, we mean places in which source, or primary research, data produced during a programme of research can be stored and maintained

C1. Does your collaboration upload data to or download data from a source repository (for example CERN, BNL)?

C2. Have you personally submitted or uploaded data to a source repository (for example CERN, BNL)?

C3. Have you personally extracted or downloaded data from a source repository (for example CERN, BNL)?

C4. If yes to C2 or C3: which ones do you use and approximately how frequently do you use these repositories?

C5. What are your experiences of source repositories (for example, is it easy to find what you are looking for? How easy are they to use?)

C6. What access restrictions do these repositories have?

C7. How could source repositories be improved?

D. Metadata

By Metadata we simply mean data about data. Metadata's principal use is to assist in the recognition, access and retrieval of data and will consist of a series of terms used to describe or identify a piece of data (for example author, title of data set).

D1. In the questionnaire you stated that you considered it important to assign the following metadata to your data (*insert in advance from Q9 of questionnaire*):

> Project reference numbers / identifiers, Author / data creator name(s), Title of data set, Subject keywords, Funding source, Publisher, Dates of Project, Date (e.g. of data creation), Format (e.g. PDF or HTML), Project description, Project title, Other suggestions (please state):

Do you think that these suggestions are sufficient on their own to make your data meaningful to others?

D2. The following were suggested in the questionnaire: do you think that these suggestions for metadata would be helpful?

(a) How data was generated
(b) What software was used (along with version)
(c) Details about instruments / apparatus
(d) Reference to related data sets

D3. Can you think of any other suggestions for metadata that would make your data meaningful to others?

D4. Do you know what types of metadata are currently assigned to your data?

D5. At what stage are metadata assigned to your research data?

D6. Who assigns the metadata?

E. Output Repositories

An output repository usually contains published articles or other texts, although it may hold other data objects that have been published. Examples of output repositories include publications at pre- or post-refereeing stage, working papers reports and PhD theses.

E1. To start this section, I would like to confirm that in the questionnaire you selected the following as output repositories that you use in your research and teaching (*insert in advance from Q20 and Q21 from questionnaire*)?

Research: Publisher, Discipline, Institutional, Other
Teaching: Publisher, Discipline, Institutional, Other

E2. Do you use arXiv? If yes, what are your thoughts (good or bad) on it?

E3. Is it your decision where your research publications are deposited, or do other members of your collaboration decide?

E4. If it is your decision: what factors attract you to a particular publication (for example: relevance of article to journal, availability of journal, ease of submitting, good website)?

E5. Considering the repositories that you have used, what are their good and bad points (good website, freely available, good search facility)?

E6. Could they be improved?

E7. Do you have any examples of good / bad output repositories?

Open access publications are those made freely available, usually via the internet. Open access publishing can be achieved either through the publication of articles in an open access journal or by depositing them in an open access repository (such as an institutional repository).

E8. In general, what are your feelings towards open access publishing?

E9. Would you be happy for your research outputs to be placed in an institutional (open access) repository?

E10. Currently, would you ever consider searching an Institutional Repository for publications?

F. Support

F1. If you have used assistance (for example provided by a librarian or some other information professional) in the use of output repositories, please would you describe it?

F2. Is there any support or guidance that you would like for using output repositories that as far as you are aware doesn't currently exist or isn't available to you?

F3. Do you think that currently you are using all the features of the output repositories that you use – or do you think other useful features exist that currently you are unfamiliar with?

G. Project Aims

The principal aim of Project StORe is to add significant value to the output repositories of research publications by enabling them to interact with source repositories of primary research data (and visa-versa)

G1. I see you answered (*insert in advance from Q2 of questionnaire*) to 'linking source to output' and (*insert in advance from Q3 of questionnaire)* 'output to source' repositories, I wonder if you can expand on why you selected these options?

G2. Are you aware of any service(s) that is already available in your discipline that links source to output (and visa-versa) repositories (for example. Durham HEP Database)?

G3. If so: (a) How does it work? (b) What are its good / bad points?

H. Reprise of Project Aims – Source Repositories

H1. Having now considered both source and output repositories, and how they might relate, what functionality do you consider to be missing from the source repositories that you have used?

For example:

H2. The following functionality have been suggested in the questionnaire: would you find any of these useful and how would they benefit your research?

1. Links between different source repositories (as well as source to output links)
2. Source data from a particular experiment organised in date order, so as one can map the development of that source data
3. Background details to an experiment included along with the data in a source repository
4. Better searching facilities in source repositories
5. The ability to keep a record of who / which collaborations have used your data

H3 We are exploring ways of providing links from repositories of source data to repositories of published papers because we believe there is a need amongst researchers to identify published (and pre-published) papers that have made use of their source data. In what way could you identify with that perceived need?

I. Reprise of Project Aims – Output Repositories

I1. What functionality is missing from the output repositories you have used?

For example:

I2. The following functionality have been suggested in the questionnaire: would you find any of these useful and how would they benefit your research?

1. Contents of tables in a machine readable format (i.e. not pdf / postscript)
2. Hyperlinks to referenced articles
3. An online area where you could store links to journals / articles which you frequently use

I3. We are considering building an interface for output repositories that would let you, as a depositor, associate newly deposited publications with the source data from which they are developed. In what way might this be of benefit to you as a researcher?

I4. A number of new operations could be supported within an output repository, such as the automatic creation of links between related resources and the presentation of relationships (i.e. showing publications and their source data in adjacent windows). How do you think these could meet your needs?

I5. What other features might you expect to be advantageous?

J. Reprise of Project Aims – Potential Solutions

J1. A 'dataset knowledgebase' is an online service that would provide efficient two-way links between source and output repositories. This service is enhanced through the addition of features such as quality assessments or ratings, and answers to frequently asked questions (FAQs) about specific sets of data held in a repository. What is your opinion of the value of such a concept and are there specific issues you might want it to address?

J2. Some data repositories are open to all enquirers while others are password protected. If we are expecting to design links that will provide access from open repositories to controlled repositories, we shall need to devise some level of validation and temporary access rights. Are there any authentication issues with regards your own source data? What degree of access protection would you expect?

B Named Repositories

Given in this appendix is a brief description of three source and output repositories. The two source repositories are the two Physics repositories specified in questions 7 and 8 of the questionnaire: Brookhaven National Laboratory and CERN, whilst the output repository is the named Physics output repository in the Project Plan [1]: University of Birmingham ePrints Service. Although the Brookhaven National Laboratory and the University of Birmingham ePrints Service repositories were not mentioned by any of the Physicists who took part in the questionnaire and interviews, their descriptions are included here for completeness.

B.1 Source Repositories

In this section the two source repositories, Brookhaven National Laboratory and CERN, are described. It should be noted that these are private source repositories where access is restricted by collaboration and laboratory membership.

B.1.1 Brookhaven National Laboratory

Brookhaven National Laboratory (BNL) [5], in New York, United States was founded in 1947 and was conceived to promote research in the biological, chemical, engineering and physical aspects of the atomic sciences. Throughout its history, BNL has had a strong focus on Physics, including the branches of High Energy, Nuclear and Condensed Matter [24].

Whilst the laboratory appears to have no single source repository that caters for all research experiments and divisions, it is believed that for some research experiments or divisions there are such source repositories. For example, the Nuclear Physics experiments at BNL's flagship Relativistic Heavy Ion Collider (RHIC) facility generates many millions of gigabytes of data which BNL manages and stores. To this end, the RHIC Computing Facility (RCF) was established in the 1990's to support the computing needs of these experiments including online recording of raw data, production reconstruction of raw data and long term archiving of all data [25]. In the mid-1990's the RCF joined forces with the ATLAS (A Toroidal LHC ApparatuS – an experiment currently being built at CERN) Computing Facility (ACF) which is also based at BNL. Combined, the RCF and ACF currently have a processing farm of over 4,000 processors, online (disk) storage of 1 petabyte, as well as a mass storage system (robotic tape system) enabling 7 petabytes of storage [25]. The RCF / ACF staff 'operates a heterogeneous, large-scale multipurpose facility, serving a geographically diverse, worldwide community of about 2,400 (and growing) users, while continually innovating and addressing ever-changing computing requirements of our user base' [25].

B.1.2 CERN

The CERN [4] laboratory in Geneva, Switzerland was founded in 1954 and is the world's largest particle physics centre. The name 'CERN' is derived from the French 'Conseil Européen pour la Recherche Nucléaire'.

The CERN source repository is a collection of disks and tapes that contain all of the Physics data collected by CERN experiments. This repository is sometimes referred to as 'CASTOR' (CERN Advanced STORage manager) [26], although more correctly, CASTOR is the mass storage system used to manage the data stored in the source repository. For example, it is CASTOR which allows transfer of data between the disks and tapes making up the source repository. The CERN Central Data Recording (CDR) system makes use of CASTOR to allow transfer of raw data from experimental areas to this central source repository [27]. The present version of CASTOR has been running since 1999 and currently handles over 51 million files totalling over 5 petabytes [26].

CASTOR is currently being upgraded to become the storage management system for CERN for the LHC (Large Hadron Collider) era (which is being built at CERN and commences operation in 2007) and, as such, is integrated with Grid technologies. When the LHC commences it will produce a huge amount of data (approximately 10 petabytes per year, which equates to one thousand times the amount of information in book form printed every year around the world) and Grid technologies (which will enable computer power and data storage to be shared) are seen as the only realistic way to access, process and store these large quantities of data [28].

Like at BNL, the Physics computing services at CERN also provide processing farms as well as tracking 'the technology, market trends and user needs. In collaboration with the users, they will introduce new services or phase out services that are no longer needed' [29].

B.2 Output Repository

Presented in this section is a description of the University of Birmingham ePrints Service.

B.2.1 The University of Birmingham ePrints Service

The University of Birmingham ePrints Service [30, 31] was founded in 2003 as part of the JISC-funded SHERPA Project [32]. It is a freely available online service developed to host the full-text of published research material produced by members of the University. The ePrints Service is part of the University of Birmingham Research Archive (UBIRA) which also comprises of two other databases: eTheses [33] ready to house PhD theses with the adoption of the recommendations of the JISC-funded EThOS Project [34]; and ePapers [35] designed to contain working papers and grey literature as and when interested parties in the University take up the opportunity to use it.

The ePrints Service is very much in its infancy and currently only contains six items, none of which are in Physics. The eTheses and ePapers databases have yet to be populated.

The advantages to researchers of such a service are many and include a wider audience for research outputs (academics, students and others, worldwide), a speedier means of research sharing and a means of helping free research output from access barriers and tolls [30]. However, a need for advocacy and promotion is clearly required in order to enable the University of Birmingham ePrints Service to become a useful facility for researchers.

C Scenarios and Use Case

In this appendix, five use scenarios and one use case are presented that attempt to capture some of the research processes and functional requirements of Physics researchers. Whilst these alone clearly cannot be comprehensive, they do consider some important aspects in the Physics research cycle.

C.1 Scenarios

C.1.1 Source Repository Scenario

Title	Downloading data from a source repository
Author	Stephen Bull
Narrative	A Physics postgraduate research student wishes to download some experimental source data from a laboratory source repository in order to begin a programme of analysis. As the source repository contains a vast amount of data, and documentation relating to the source repository is limited, the student consults collaboration colleagues in advance to find out precisely where in the source repository the data in question is located. The student uses his laboratory account details to log into the source repository and navigates to the relevant part. The student downloads the data to his local group server. With the data on his group server, the student begins his programme of analysis.

C.1.2 Output Repository Scenarios

Title	Publishing in an output repository
Author	Stephen Bull
Narrative	A Physics researcher has performed an analysis on some aspect of the experimental data, collected by his collaboration, and has written an article detailing his results. The researcher wishes to publish this article in a publisher output repository. He obtains clearance from his collaboration that the article is suitable to be published. With the help of collaboration colleagues, the researcher decides on the most appropriate publisher output repository to submit the article to. The researcher considers it important to publish in an output repository that is relevant to his article and ranks well in the ‘impact factor hierarchy’. The researcher submits his article to the chosen publisher where the article is peer-reviewed. The article is accepted by the publisher and is placed in their output repository where it can be viewed by members of institutions who have subscribed to this repository.

Title	Using a search facility to identify relevant publications in output repositories
Author	Stephen Bull
Narrative	A Physics researcher has produced some results that she wishes to compare with the results produced by other collaborations studying similar things. Using some keywords, the researcher performs an internet search which returns many matches. The researcher attempts to identify the matches that are most suitable and, using the linkage provided by the search facility, navigates to the respective output repositories. In some cases she is able to access the full-text of the article, whilst in others she is prevented due to access / subscription issues.

C.1.3 Output to Source Repository Scenarios

Title	Obtaining data from a publication
Author	Stephen Bull
Narrative	A Physics researcher has discovered a figure in a publication that contains data comparable to her own. She wishes to obtain the numerical data that make up the figure (which are not explicitly given in the publication itself) to allow comparison with her own data. The researcher prints out the publication and expands the figure a number of times on a photocopier. Using a ruler, the researcher measures the distance of each data point from both the horizontal and vertical axes of the figure. The researcher then calibrates her ruler measurements to the scales of the axes on the figure and then writes the results into a spreadsheet. The researcher compares her results to those that have been published.

Title	How a researcher would like to obtain data from a publication
Author	Stephen Bull
Narrative	A Physics researcher has discovered a figure in a publication that contains data comparable to his own. He wishes to obtain the numerical data that makes up the figure (which are not explicitly given in the publication itself) to allow comparison with his own data. The researcher follows a link from the output repository, in which the publication is held, to a source repository which contains the full numerical data which make up all of the figures in the publication. These numerical data are stored in numerous formats including text files, spreadsheets, XML and HTML formats. The researcher selects an appropriate format and extracts the numerical data. The researcher compares his results to those that have been published.

C.2 Use Case

C.2.1 Output to Source Repository Use Case

Author

Stephen Bull

Use Case Summary

A Physics researcher wishes to identify publications written by other collaborations that are relevant to their work

Primary and Other Actors

Primary Actor	Researcher
Other Actors	Search facility, output repository

Stakeholders and Interests

Researcher	To obtain the information as quickly and efficiently as possible
Output Repository Manager	To disseminate contents to educational users
Owner of Search Facility	To provide a satisfactory service to users so as they will use facility again

Precondition

None

Main Success Scenario

1	Researcher enters appropriate keywords into search facility
2	The search facility returns some matches (or results)
3	Using the information returned by the search facility, the researcher identifies the result(s) which they consider most relevant
4	The researcher follows the link, provided by the search facility, to the place within the output repository where the published article can be found
5	The researcher reviews the abstract of the article and decides they want to read the full-text version
6	The link, provided by the output repository, to the full-text version of the article is followed
7	The researcher accesses the article

Extensions

2a	Search facility returns no matches
2a1	Researcher chooses different keywords
2a2	Search facility returns matches
2b	Search facility returns too many matches
2b1	Researcher makes search more specific
2b2	Search facility returns fewer matches
3a	Researcher is unable to identify relevant items as the information provided by the search facility is too vague
4a	The link is broken
4a1	If specified, the researcher independently navigates to the output repository in question
5a	The abstract suggests the article is not what is required
5a1	Researcher returns to stage 3 of the Use Case and tries again
6a	Full text not available
6a1	Researcher backs out of User Case
7a	Authentication requirements prevent user accessing publication
7a1	Use Athens authentication

Bibliography

[1] J. MacColl and G. Pryor, (StORe Collaboration), Project Plan, *http://jiscstore.jot.com*, (November 2005).

[2] JISC Digital Repository Wiki, *http://www.ukoln.ac.uk/repositories*, (created October 2005).

[3] JISC (The Joint Information Systems Committee), *http://www.jisc.ac.uk*, (April 2003).

[4] CERN, *http://www.cern.ch*, (2006).

[5] BNL (Brookhaven National Laboratory), *http://www.bnl.gov*, (2006).

[6] R. Brun and F. Rademakers, *http://root.cern.ch*, (created 2005).

[7] Physics Analysis Workstation, *An Introductory Tutorial*, CERN Program Library, Long Writeup Q121, (1995).

[8] DESY (Deutsches Elektronen-Synchrotron), *http://www.desy.de*, (2006).

[9] SLAC (Stanford Linear Accelerator Center), *http://www.slac.stanford.edu*, (2006).

[10] Fermilab (Fermi National Accelerator Laboratory), *http://fnal.gov*, (2006).

[11] RAL (Rutherford Appleton Laboratory), *http://cclrc.ac.uk/Activity/RAL*, (2006).

[12] HEPDATA (The Durham HEP Databases), *http://durpdg.dur.ac.uk/HEPDATA*, (2006).

[13] Worldwide LHC Computing Grid, *http://lcg.web.cern.ch/LCG*, (2006).

[14] UKDA (UK Data Archive), *http://www.data-archive.ac.uk*, (created 2002).

[15] Cornell University Library, *http://www.arxiv.org*, (2006).

[16] SPIRES, *http://www.slac.stanford.edu/spires*, (2006).

[17] CERN Document Server, *http://cds.cern.ch*, (2006).

[18] ISI Web of Knowledge, *http://wok.mimas.ac.uk*, (2006).

[19] Wikipedia, *http://en.wikipedia.org/wiki/Certificate_authorty*, (2006).

[20] Physical Review Online Archive, *http://prola.aps.org*, (2006).

[21] IOP electronic journals, *http://www.iop.org/EJ*, (2006).

[22] Nature.com, *http://www.nature.com*, (2006).

[23] Semantic Web, *http://www.w3.org/2001/sw*, (2006).

[24] A History of Physics Research at Brookhaven, *http://www.bnl.gov/physics/history*, (2006).

[25] S. Misawa, *http://www.rhic.bnl.gov/RCF/GuidedTour*, (November 2005).

[26] CASTOR (CERN Advanced STORage manager), *http://castor.web.cern.ch/castor*, (2006).

[27] Central Data Recording, *http://cdr.web.cern.ch/cdr*, (2004).

[28] Grid Café, *http://gridcafe.web.cern.ch/gridcafe*, (2006).

[29] CERN IT Department, *http://it-div.web.cern.ch/it-div/what-we-do*, (2004).

[30] The EPrints Service, *http://eprints.bham.ac.uk*, (2006).

[31] G. Gilbert, Open Access – making research more available, *http://www.is.bham.ac.uk/scholcomm*, (2006).

[32] SHERPA, *http://www.sherpa.ac.uk*, (2006).

[33] The eTheses Archive, *http://etheses.bham.ac.uk*, (2006).

[34] EThOS (Electronic Theses Online Service), *http://www.ethos.ac.uk*, (2006).

[35] The ePapers Archive, *http://epapers.bham.ac.uk*, (2006).