Project StORe (Source-to-Output Repositories)
Report for Biosciences

Issued to
Graham Pryor, StORe Project Manager

Date: 15th September 2006
Version: 3

By: **Dagmar Biegon**
Project Researcher
The John Rylands University Library
The University of Manchester
Oxford Road, Manchester M13 9PP

Tel. 0161-275 8916
E-mail dagmar.biegon@manchester.ac.uk

Table of Contents

Abstract

This report was written as part of the first phase of the national higher education research project StORe (Source-to-Output Repositories). The project included a large scale analysis of repository user behaviour, with an extensive survey of the research community in seven academic disciplines. Data was acquired through a detailed on-line questionnaire and a series of individual interviews. The observations from the biosciences research community, as one of the seven survey constituencies, are presented in this report.

Most of the biosciences researchers who took part in the survey seem to be generally in favour of an improved linkage between source and output repositories. Overall, they create data as a combination of different types and formats, to which they assign the metadata themselves, during the stage of file sharing. Many researchers seem to have some reservations about sharing their research data with the potential risk of premature broadcast as the main reason given for this attitude. If and when biosciences researchers share their research data, they do so on individual request and through the exchange of portable media or as e-mail attachments. However, a large group of researchers share their data freely and have no access control mechanisms in place.

Surprisingly, only about half of the biosciences researchers stated that they submit their data to a source repository. Those who do submit do so mainly on an occasional basis. The main important source repository for the biosciences was stated to be GenBank and PubMed was named as the main important output repository. A significant proportion of researchers use an unspecific browsing strategy to keep up with new developments in their field. Many biosciences researchers prefer to develop their own, self-sufficient information research strategies, using a range of different access routes to various repositories.

In questionnaire and interviews, researchers also made various suggestions for the improvement of repository functionality and information research, ranging from more options for searching to improved help functionality. Personal communication was also mentioned as an important information access route, especially when searching across different disciplines.

List of Figures

List of Tables

1. Survey

1.1 Introduction: Purpose and key dates for the survey

Project StORe (Source-to-Output Repositories) is a UK national project, funded by the JISC (Joint Information Systems Committee) as part of their Digital Repositories Programme. The principal aim of StORe is to investigate the interactions between output repositories of research publications and source repositories of primary research data (1, 2)[1].

One outcome of the project will be the development of specific middleware. This will improve the functionality of both source and output repositories and provide better repository linkage. Another result will be a comprehensive overview about the various information research and information management strategies currently existing within the UK academic research community (3)[2].

The project runs from September 2005 to August 2007. Eight universities are taking part in the project, covering seven disciplines between them. These academic disciplines are archaeology, astronomy, biochemistry, biosciences, chemistry, social policy/political science and physics.

A detailed survey of user behaviour was conducted in order to produce the optimum requirements specification for the software. This survey was performed through an internet-based questionnaire, followed by a series of individual interviews. The outcome of the survey will be used in the StORe business analysis, on which the middleware development will be based.

Each university established the contact to selected research scientists in a different academic discipline and disseminated the project information within this community. The survey conducted from the University of Manchester covered the field of biosciences.

A generic internet-based questionnaire was set up to cover all seven disciplines. This questionnaire was designed in February 2006, tested on volunteers in different disciplines, and went live from 13 March 2006 to 21 April 2006. The results were used to identify the issues that needed further clarification, so that the researchers' opinions and information requirements could be probed better during the interviews. In Manchester, the interview phase took place in May 2006. Evaluation of the survey results was performed in June and July 2006.

1.2 Methods and tools used

1.2.1 Constituency

In Manchester, the StORe Researcher was a member of staff within the Information Resources and Academic Support Division of the John Rylands University Library (JRUL) at the University of Manchester. The project was overseen by the Faculty Librarian for Life Sciences at the same division. The primary focus was on investigating the information needs of our biosciences research community. Since today's leading edge research topics such as immunology, gene sequencing, or cancer biology involve more than just the 'classical' biosciences, we extended our survey constituency to the biological research groups within pharmacy, medicine, chemistry and computer sciences.

Our survey focused mainly on the University of Manchester, but we also included the biosciences faculties at two other universities in the North of England (Liverpool and Leeds). Commercial companies or independent research institutes were not targeted at this point. The number of biosciences researchers within these institutions within the biosciences is potentially very large and due to the timing of the questionnaire it was felt that establishing suitable selection criteria for a mailing would have been too complex for the scope of the project.

[1] see bibliography list, page 86
[2] see bibliography list, page 86

1.2.2 Questionnaire dissemination

There are various strategies for a successful on-line data collection via questionnaires. One approach is to disseminate the questionnaire information via blanket mailing lists, where a large number of potential respondents are targeted. The response rate for a blanket mailing might be low, but because of the high number of mails sent out, the absolute number of responses might be sufficient. Another strategy to increase the response rate is to address the e-mail information to named individuals and/or disseminate the information via key contacts, and send out reminders (4)[3].

All the universities participating in the StORe project chose a combined approach of the above methods, using blanket e-mailing or personalised e-mails whenever possible, and sending out reminders. The response rate was varied, ranging from about 8 to 12%. This is consistent with the general expectations for postal and on-line mail surveys (4)[4], where a response rate of less than 20% would be expected.

At the University of Manchester, we disseminated the information through key contacts and specific faculty publications. We were keen to involve the local biosciences community right from the beginning of the survey and liaised closely with key academic staff within the Faculty of Life Sciences (FLS) in order to develop a cooperative information dissemination strategy. As a direct result, our e-mails and reminders were sent to selected target groups of academic staff and research fellows within FLS.
Additionally, personalised mailings were also sent to research groups in other faculties, namely the Faculty of Medical and Human Sciences, the School of Chemistry, the School of Pharmacy and Pharmaceutical Sciences, and the School of Computer Science. The addressees were heads of research groups involved in collaborations with members of FLS. Some of these were personal contacts.

It is not possible to give the exact figures for how many recipients had access to the project information, due to the distribution mode via key contacts. Our feedback revealed that some contacts had passed the information on via e-mail to their research groups or had placed the information in various faculty publications. To give an exact figure for how many people had read these publications would be difficult, but a general estimate for how many researchers had access to the project information would be as follows:

More than **300** researchers at the University of Manchester received StORe project and questionnaire information, through various means of dissemination. Another **300** researchers were targeted via intradepartmental publications at the universities of Liverpool and Leeds. This makes the sum of both constituencies about **600**, which exceeds the project target of 500 for each discipline. We were able to meet this target by writing to three universities only because of the large number of biosciences researchers working at each university.

In order to use an alternative means of disseminating information, a project web site (2)[5] was set up in March 2006. This page was set up as a sub page of the JRUL web page, with the aim of personalising the project information and making it more accessible for our web-experienced target population. The JRUL project web page contained links to both the JISC StORe project web page and to the on-line questionnaire. It was regularly updated to reflect events such as the StORe prize draw in July.

The calculated response rate for the StORe questionnaire at the University of Manchester was **12.2%** and this is consistent within the response rate for other StORe participating universities, which ranged from about 7 to 12%. The total number of answered questionnaires within the biosciences was **40.** This is equivalent to **10.6%** of all StORe questionnaire responses, which were 377 in total.

It was noted that within the disciplines the total number of StORe questionnaire responses for biosciences, biochemistry, and chemistry was around 40 each, while the total number of questionnaire responses for archaeology, physics, astronomy and social sciences/ politics was about 60. To investigate if these figures

[3] see bibliography list, page 86
[4] see bibliography list, page 86
[5] see bibliography list, page 86

reflect a different approach to the dissemination of information or if they are arbitrary could possibly be interesting for any future data gathering exercises.

Another reason may be that we have ‘chopped up’ one large discipline, consisting of chemistry, biochemistry, and biosciences, into three sub-disciplines with much smaller constituencies. On indication for this fact may be if there is a significant overlap between areas of interest and repositories used within these three disciplines. We at Manchester have also contacted fewer universities than other disciplines, but due to the timing of the questionnaire it was felt that mailing more institutions would not have increased the responses.

1.2.3 Interviews

Interviews have the advantage that the questions can be more varied to suit the context and that the interviewer can probe interesting issues as they arise and in greater detail (5,6)[6]. The StORe interviews were used to explore two key issues of interest for this and other related research projects: the researchers’ reasons for implementing access control to their research data and the users’ opinions on improvements to repository functionality.

Interview duration: We aimed to keep the interviews to 30 minutes. It was known from experience that most people’s attention will be less focused after this time. Interviewees are also more likely to volunteer if they feel that the interview will not impose major changes to their own their time schedule (7)[7].

Interviewee population: Six responders who had answered the StORe questionnaire also volunteered for an interview. Through the use of an FLS department contact list, we were able to bring the total number of interviews up to **12**, which exceeded the project target of 10.

Interview mode: All interviews were conducted on a personal and individual basis, apart from one interview, which was conducted over the telephone. The interviews took part in the interviewees’ offices and lasted for 30 to about 45 minutes, with the interviewer taking written notes during the course of the interview and transcribing the whole interview at a later stage.

Interview questions: The interviews consisted of 15 mainly open-ended questions. The aim was to give the interviewee the opportunity to talk freely about his/her information management behaviour. The 15 interview questions, which can be found in Appendix 1, were intended as a general guideline only. In some cases, only a part of the questions were asked and preference was given to let the researchers talk relatively freely about their specific issues of interest.
Whenever an interesting issue arose or an answer needed to be clarified, follow-up questions were used, as is the general practice in conducting interviews. It was felt that an in-depth probing of the researcher’s opinions would result in best quality information.
In addition to questions about the topics which had been addressed in the StORe questionnaire, four questions were asked about cross-disciplinary search and the use of help functions.

Interviewee’s attitude: On the whole, the interviewees gave the impression to have a very or generally positive attitude towards the aims of the StORe project. However, two of the 12 interviewees showed a rather negative attitude to the project’s aims. This was explicitly stated during the interviews. Some of the other 10 interviewees, even if they were generally positive towards the project’s aims, were more negative about giving access to their data. These facts will be explored in more detail in the results section.

Interview counting: Six of the interviewees had already answered the StORe questionnaire. Their answers are marked in the results section 5 as ‘Respondents’, while the answers from the six interviewees who had not previously answered the questionnaire are marked as ‘Non-Respondents’. The sum of both numbers is given as total count.

[6] see bibliography list, page 86
[7] see bibliography list, page 86

In order to avoid double counting for identities (Question 1a), 'Respondents' were not included in the count of professional identities. However, this calculation was not feasible for all the questions, since it would have meant backtracking every answer during the interviews. An error margin of about 5-10% is therefore theoretically possible, but was judged to have no significant influence on the general observations made in both interviews and questionnaire.

Table 1: Interviews overview

Total	**Interviewees answered questionnaire**	**Interviewees not answered questionnaire**	**Personal interviews**	**Telephone interviews**	**Interviewee attitude rather positive**	**Interviewee attitude rather negative**
12	6	6	11	1	10	2

1.2.4 Evaluation tools

The StORe questionnaire web site was physically set up in Bristol Online Surveys System (BOS), a system developed by the Institute for Learning and Research Technology at the University of Bristol. Answers to the StORe questionnaire were imported into Excel 2002, which was used to perform calculations and cross-tabulations.

The BOS system does not convert multiple answers to a single question into percentages. These percentages were manually calculated whenever it seemed logical, for example when a negative answer had been given. They were judged to be appropriate to illustrate the project observations.

Scenarios were written according to the guidelines and materials provided in a special JISC training session at the Edinburgh Training Centre (8)[8].

1.3 Review of nominated repositories

1.3.1 Source repository: Universal Protein Resource (UniProt)

UniProt is currently the world's most comprehensive catalogue of information on proteins. It is run by a consortium comprising of the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SBI) and the Protein Information Resource (PIR) (9)[9]

UniProt was launched on 15 December 2003 and is mainly supported by a 3-year U.S. National Institutes of Health grant, to a much lesser extent also by E.U. contracts (through EBI), the Swiss Federal Government (through Swiss-Prot) and a National Science Foundation (NSF) grant (through PIR). It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot[10], TrEMBL[11], and PIR-PSD[12].

[8] see bibliography list, page 86

[9] see bibliography list, page 86

[10] Swiss-Prot is the manually annotated and curated protein sequence database of the SBI. This annotation is a labour-intensive process and involves assessment of information from published articles along with the use of a variety of programs/algorithms.

[11] TrEMBL (Translated EMBL) protein database, in Swiss-Prot format, is generated by computer translation of the genetic information from the EMBL (European Molecular Biology Laboratory) library. EMBL is the precursor of today's EBI.

[12] PIR-PSD (Protein Information Resource-Protein Sequence Database), run by the US National Biomedical Research Foundation since 1984, is an integrated public bioinformatics database with automated annotation.

Swiss-Prot, established in 1986 at the Swiss Institute of Bioinformatics in Geneva, is today recognised as the gold standard of protein annotation, with extensive cross-references, literature citations and computational analyses provided by expert curators (10)[13].
However, proteomics research has accelerated in recent years because of technological advances in protein science and the large amounts of genomic data generated by the Human Genome Project. Recognising that sequence data were being generated at a pace exceeding Swiss-Prot's ability to keep up with the cataloguing of detailed protein structural data, TrEMBL was created to provide automated annotations for those proteins not in Swiss-Prot. Meanwhile, PIR maintained the PIR-PSD, a database of protein sequences and curated families. In 2003, EBI, SBI and PIR decided to pool their overlapping and complementary resources, efforts and expertise. This resulted in the establishing of UniProt (10)[14]

As of September 2005, Swiss-Prot holds entries on 114,000 proteins, TrEMBL 700,000 and PIR 283,000, accessible through UniProt. As a publicly funded project, UniProt's data is freely accessible and the data is released in a timely manner.

The database provides cross-references to external data collections such as the underlying DNA sequence entries in the DDBJ/EMBL/GenBank nucleotide sequence databases, 2D and 3D protein structure databases, various protein domain and family characterisation databases, post-translational modification databases, species-specific data collections, variant databases and disease databases. As a result of this, UniProt acts a central hub for biomolecular information archived in more than 50 cross-referenced databases (9)[15].

Reference lists of keywords, journal abbreviations, Organism Identification Codes, tissues, strains and plasmids are provided. The sequences and information in UniProt are accessible via text search, BLAST similarity search (Basic Local Alignment Search Tool, for more details see Scenario 3), and FTP.

UniProt is comprised of three different components, each optimised for different uses (11)[16]:

The **UniProt Knowledgebase (UniProtKB)** is the central access point for extensive curated protein information, with accurate, consistent, and rich sequence and functional annotation, including comprehensive cross-linking. UniProtKB is the main point where the work of Swiss-Prot, TrEMBL and PIR is combined and continued.

Researchers are able to submit protein sequences directly to UniProtKB using the submission tool SPIN. UniProtKB accepts submissions of new sequences, entry updates and corrections, and annotated bibliography.

Since July 2000, entries to UniProtKB/TrEMBL have contained evidence tags. Each data item may have one or more evidence tags. Each tag consists of four elements: category, type, attributes and dates, and refers to one piece of evidence that supports the data item. If that becomes invalid, e.g. through a change to a rule for automatic annotation, the data item will be deleted if the deleted evidence tag was the last one supporting the data item. This allows for keeping automatic annotation procedures up-to-date.

The **UniProt Reference Clusters (UniRef)** clusters databases combine closely related sequences into a single record to speed searching. They are separated in UniRef100, UniRef90 and UniRef 50, collapsing all the sequences that are at least 95%, 90% or 50% identical.

The **UniProt Archive (UniParc)** is a comprehensive repository, storing the complete body of publicly available protein sequence data, reflecting the history of all protein sequences. It contains the protein sequences from Swiss-Prot, TrEMBL, PIR-PSD, EMBL, ENSEMBL, International Protein Index (IPI), PDB, RefSeq, FlyBase, WormBase and the patent offices in Europe, the US and Japan. UniParc contains the

[13] see bibliography list, page 86
[14] see bibliography list, page 86
[15] see bibliography list, page 86
[16] see bibliography list, page 86

protein sequences, sequence versions and database cross-references only. All other information concerning the sequences must be retrieved using database cross-references.

There were 5.6 million database cross-references in UniParc by October 2003. Database cross-reference is active as long as the sequence identified by the accession number remains unchanged, otherwise the cross-reference retires and can then only be used for archive searching. UniParc is available for text- and sequence-based searching.

1.3.2 Source repository: Genetic Sequence Data Bank (GenBank)

GenBank is the NIH (National Institutes of Health) genetic sequence database, an open access, annotated collection of all publicly available DNA sequences and their protein translations. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI (National Center for Biotechnology Information) (12)[17].

The precursors to GenBank, the EMBL Data Library of the European Molecular Biology Laboratory and the initial GenBank of the Los Alamos National Laboratory, were both established in the early 1980s. In February 1987 the International Nucleotide Sequence Collaboration between EMBL, GenBank and DDBJ was formalised.

The NCBI of the National Library of Medicine at NIH is responsible for the production and distribution of the NIH GenBank Sequence Database. NCBI distributes GenBank sequence data by anonymous FTP, e-mail servers and other network services.

GenBank receives sequences produced in laboratories throughout the world. It continues to grow at an exponential rate, doubling every 10 months (as of June 2006). The repository is built by direct submission from individual laboratories (for example through the submission program Sequin, see also Scenario 5), as well as from bulk submissions from large-scale sequencing centres.

Submitters to GenBank currently contribute over 3 million new DNA sequences per month to the database. In June 2006 it contained over 60,000,000 sequences, from about 200,000 organisms (13)[18].

Most biology researchers coming in contact with molecular biology are familiar with GenBank. Its use is central to modern molecular biology and to bioinformatics. Because the GenBank records are maintained by the individual scientists who discover the sequences, anybody who finds a new sequence of interest can then publish it in GenBank. The free access to this information allows scientists to study and compare the same data as their colleagues and makes collaborative research in the biomedical sciences possible.

GenBank is accessible through NCBI's retrieval system Entrez which integrates data from the major sequence databases along with other taxonomy, genome, mapping, protein structure and domain information, and the biomedical literature via PubMed.

The GenBank files comprise data from 1982 to the present. They are uploaded daily and the data bank is reloaded quarterly. There is also an automatic current-awareness search facility available, which runs weekly.

GenBank is organised as a set of libraries comprising of flat files containing many records in succession. it Is often referred to as a databank, rather than a database, because of its ASCII text file format and flat file organisation. There are currently 243 library files, most of which are over 200 MB in size. The libraries are usually distributed compressed. Uncompressed they amount to about 100 gigabytes in data (14)[19].

[17] see bibliography list, page 86

[18] see bibliography list, page 86

[19] see bibliography list, page 86

Each GenBank entry (record in a file) includes a concise description of the sequence, the scientific name and taxonomy of the source organisms, and a table of features (gene annotation) that identifies coding regions and other sites of biological significance, such as transcription units, sites of mutations or modifications, and repeats. Protein translations for coding regions are included in the feature table. Bibliographic references are included along with a link to the Medline unique identifiers for all published sequences.

1.3.3 Output repository: PubMedCentral (PMC)

PMC was developed and is operated by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the U.S. National Institutes of Health (NIH). NLM intends the archive to become a digital counterpart to the preservation and unrestricted access that it maintains for its print collection.
PMC exists for two reasons, both of which arise from NLM's congressional mandate: 1) to permanently preserve digital journal literature in the life sciences and 2) to improve access to biomedical information for health professionals, researchers and the public. It is a separate entity from PubMed, the version of Medline also provided by the NLM, but all the articles in PubMed Central have a corresponding entry in PubMed. As an archive, PMC strives to collect and preserve everything that is published in a participating journal (15)[20].

PMC was launched in February 2000, but the archive content reaches back to the earlier print issues of many of the PMC journals in order to provide online access to the complete run of these journals. In some cases the digitalised content goes back to the early 1900s.

All articles in PMC are peer-reviewed. The participating journal's own guidelines for publications apply. A journal can choose to make content available on PMC as soon as it is published, or delay for a specified time, such as one or two years after publication. PMC's business model is a reverse of the current copyright ownership practice and charges authors for publishing but not accessing articles. Copyright to all material deposited remains with the publisher or the individual authors. PMC is simply an archive.

Each publisher has password-controlled access to a web site that has usage statistics reports for that publisher‘s journal/s. These reports are updated daily.

PMC gather the content from its diverse journal sources into a single repository, where it is stored in a uniform format, the NLM Journal Archiving and Interchange XML DTD (Data Type Definition).
Submitted material must be in XML or SGML format, compliant with an acceptable journal article DTD. Supplementary material in the form of the original digital image files, video, audio or data files may also be submitted. Every time a user asks to see an article in PMC, the HTML view of this article is created on the spot, directly from a copy of the archived file.
All material of the initial HTML full text presentation in PMC is in English, but the PDF of the original journal article that is linked to the PMC entry might be in another language, which is for example the case for some Canadian journals.

The PMC OAI service (PMC-OAI) provides access to the metadata of all items in the PMC archive, as well as to the full text of a subset of these items. PMC-OAI is an implementation of the Open Archives Initiative protocol for metadata harvesting (OAI-PMH). PMC-OAI supports OAI-PMH version 2.0. It does not support earlier versions of the protocol.
In addition, PMC is developing XML tagging guidelines so that the 'administrative content (editorial board and staff lists, instructions to authors, notices and announcements, and advertisements) of a journal can also be included in the archive.

Records may be retrieved from the PMC archive in the NLM Journal Archiving and Interchange XML format, for metadata or full-text article records, or in the Dublin Core format for metadata only.

[20] see bibliography list, page 86

2. Summary of observations

This section gives a summary of the significant observations from both questionnaire and interviews. In order to get the best overview about the results, it should be read in conjunction with those sections of the report that contain the relevant tables and figures.

2.1 Identities

About 70% of the StORe questionnaire responses came from university academic staff, with significantly fewer responses from university research assistants, postgraduate students and other researchers. Similar observations were made in many of the other seven different disciplines.
It should be noted that out of 12 interviewees, six had previously answered the questionnaire (all of them university academic staff) while six had not previously answered the StORe questionnaire. This latter group consisted of three university academic members of staff and three university research assistants, respectively.

Many researchers stated more than one field of interest. It was judged that a narrower categorizing than the RAE system would reflect the diversity of interests in the biosciences best. The categories were therefore set as a combination of the various discipline areas at the University of Manchester and fields of interest, as stated by the researchers.

Within the field of biosciences research, molecular biology has experienced a significant growth within the last decades. This makes it an important field of interest for many biologists pursuing a career in research.
Indeed, molecular biology as field of interest was stated by about 30% of the biosciences researchers in general and 33% of the interviewees in particular. This is more than double of how often other biosciences fields were given as interest. Bioinformatics was stated by even more (42%) interviewees. Research work in molecular biology and related fields often requires a high level of IT competence, which possibly has some consequences for the handling of scientific source data and the usage of output repositories.

It should be noted that some more specialised fields of interest, such as cancer research, often border onto different disciplines, such as medicine. This reflects an overall trend of quite fluid boundaries between modern disciplines. It can be assumed that this fact has a positive effect on both cross-disciplinary co-operation and cross-disciplinary information research.

2.2 Project aims

Most (i.e. more than 80%) of the biosciences researchers seem to support the StORe project aims. In their answers, they stated that they would find improved links from both source-to-output and output-to-source repositories a significant advantage to their work, or at least useful.
Interviewees seemed to fall within two roughly equal, distinctive groups with three being positive about improved repository links (even already using such a feature), and three being uninterested or negative, with the remaining six not expressing any opinion. However, this number is too small to give more than an indication of an underlying trend.

The current functionality of both source and output repositories seems to be adequate for many researchers, as indicated by the number of researchers who actually took the trouble to state this opinion in a questionnaire free text answer (5 and 9, respectively).
However, many of the other responses seem to point to a need for better standardisation within source repositories and possibly also indicate a lack of familiarity with output repository search strategies. A potential improvement to meet these needs could be the introduction of fully self-explanatory features and/or improved help functions (see also additional topic 4.8: Help functions). Library administrative issues, such as missing links to full text articles were also among the missing functionalities, as stated by the researchers.

2.3 Source data

Biosciences researchers seem to produce a wide range of different data, ranging for example from spectrograms to videos. The main data types within this range are images, drawings/plots, raw data and gene/protein sequences. A preference for these four data types was noted especially for the interviewees. One explanation could be that the research output in important fields such as molecular biology consists of a large amount of electrophoresis and microarray images, as well as gene/protein sequences.

Data formats are mostly consistent with the aforementioned data types. It should be noted that gene/protein sequences are usually written in plain text or FastA, which is another text format. This fact, together with the writing of experimental protocols, could account for the significant amount of text produced. Electrophoresis and microarray images, but also medical images such as CAT scans, are most often stored as JPG files, which may account for the high occurrence of (JPG, TIF, BMP, GIF).
In addition, more unusual data formats were mentioned during the interviews, such as Multiple Spanning Tree Protocol (MST). It should be kept in mind that one important field of interest for interviewees was given as bioinformatics and that bioinformatics researchers would possibly work with specialised data formats.

In general, the creation of plots, databases, text data and statistical data, with their corresponding formats, would certainly be expected as output for biosciences research since they reflect key processes in the scientific evaluation of experimental data.

Most researchers (70%) stated that they often or sometimes generate a combination or groups of different data, making this prevalent behaviour for the biosciences.

More than half of the researchers stated that their motivation for accessing research data was to gain knowledge through accessing data (75%) and understanding data in context (50%). This shows a clear focus on the advancement of one's individual knowledge, rather than being interested primarily in evaluating other researchers' output. Indeed, testing others' objectives and identifying useful contacts seems to be of much lesser importance (13% each).

Online access to the data of other researchers was stated by 35% of the questionnaire respondents to be the most important route for accessing biosciences research source data, followed by portable media (25%) and own networked fileservers (18%). In contrast, 75% of the interviewees favoured the exchange of portable media, followed by networked fileservers at the own institution (50%). Online access was used by 30% of the interviewees and 25% also stated that they use personal communication routes to access other researchers' data. Other institutions' networked fileservers and e-mail attachments do not seem to be of much importance for neither questionnaire respondents nor interviewees (15% and 10%, respectively). This question allowed for multiple answers.

A significant number of biosciences researchers also stated that they do not normally access other researchers' data (21%).

It should be noted that during the interviews two researchers working with microarray data stated that they would not normally access other researcher's data because of the unmanageably large file size. They would be interested in processed information only and avoid accessing the raw data.

Generally speaking, the observations for the access routes to biosciences research source data point to a high level of existing data exchange with much of the research data not being in the public domain.

2.4 Source repositories

Surprisingly, about 50% of the researchers stated that they do not submit their source data to any source repository. Most researchers who submit do so to GenBank (about 25% of all answers). This may be due to the fact that submission to GenBank (or PDB or EBI)[21] is mandatory prior to the submission of a publication in a scientific journal, as a requirement by the publisher.

The names of preferred repositories, as stated in interviews and questionnaire, might benefit from some clarification. It is possible that for interviewees, a submission to the unspecific 'Internet' (3 answers) could include one of the named questionnaire 'online access repositories' and that submission to 'EBI' (2 answers) includes UniProt[22] (1 answer). None of the interviewees stated GenBank as source repository of choice.

Questionnaire respondents who had stated a named repository also gave the frequency of their submissions and the numbers are consistent. Most questionnaire respondents stated that they submit to GenBank on an occasional basis (33% of all researchers who submit).
This may be due to the inherent nature of research processes in molecular biology: Researchers (especially postgraduate students) may only generate a single or a few genetic sequences during the whole course of their complex research work and thus would mainly submit on an occasional basis. A higher sequence output would only be expected from a highly specialised individual, a commercial sequencing lab, or a rather large molecular biology research group. See also Scenario 5 for details (Depositing a genetic sequence in a public repository).

2.5 Metadata

There seems to be a marked difference in the choice of metadata between questionnaire respondents and interviewees. However, during the interviews the impression was gained that at least some researchers were not familiar with the term metadata and seemed to be translating it to a possible equivalent in their field, such as annotation or program documentation. Some seemed to be relying on the explanation and examples given by the interviewer (and this is theoretically possible for the questionnaire) and the results should be thus interpreted with care.

The main types of metadata assigned by all biosciences researchers to their data are the author's name, date and project title. Questionnaire respondents also named subject keywords as an important type of metadata. Interviewees however seem to put greater importance on links, database accession numbers and the description of experimental conditions. This is especially true for interviewees who had not previously answered the StORe questionnaire. Subject keywords did not seem to be of the same importance for interviewees and only three named this type of metadata.

One explanation may be that many interviewees work with emerging fields, such as microarrays, and thus deal with inconsistent standardisation, such as the MIAME[23] standard not being fully implemented. This makes extensive caption of data, for example the detailed description of the experimental conditions, absolutely essential. A researcher working in molecular biology or bioinformatics would also submit data in the form of gene or protein sequences. These require different metadata from e.g. scientific publications and would not necessarily require subject keywords. Instead, different metadata may be required by the source repository administration, for example the allocation of database accession numbers and links.

Most researchers (44%) seem to assign metadata during the stage of file saving, with quite a significant proportion not being certain at what stage metadata is assigned (17%) or not assigning any metadata at all (10%), again possibly indicating a lack of familiarity with the term metadata.
Assigning metadata to research data seems to be a task which is mainly carried out by the individual researcher (i.e. by 71%). To a much lesser extent, this task is the responsibility of a team colleague, repository administrator or support staff.

[21] PDB (Protein Data Bank) is managed by the Research Collaboratory for Structural Bioinformatics; EBI runs several sequence and microarray databases.
[22] The EBI protein database is a part of UniProt.
[23] MIAME describes the Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. It is a preliminary standard, set up by the Microarray Gene Expression Data Society.

2.6 Data access and sharing

The main method for sharing data seems to be via e-mail attachments (38% of questionnaire respondents and 33% of interviewees). In the case of questionnaire respondents, 'making data available via a publisher' (40%) and the 'exchange of portable media' (28%) also seem to be important. However, a surprising 25% of the questionnaire respondents stated that they do not have any measures in place to make their research data available.

For the interviewees, the 'use of a publicised URL' was also important, stated by about 50%. None of the interviewees mentioned data exchange via a publisher, compared to 30% of the questionnaire respondents. Interviewees also stated a variety of additional routes of data sharing. It could be possible that questionnaire respondents had scientific publications in mind when answering this question and that the interviewees were possibly thinking of sharing their data through a public domain web site.

An interesting statement was given by one interviewee who said that once his research group's data was submitted to the public domain, they would not keep a copy of it. This could possibly indicate a policy of outsourcing very large source data sets in order to save computer storage capacity. The same interviewee makes use of virtual shared web drives to exchange information with international collaborators.

Most biosciences (63%) researchers state individual reasons, such as improved visibility, as the main factor that would encourage them to share their data. However, practical reasons such as a requirement of the funding body, or altruistic reasons, such as benefits to the research community, also feature highly, with nearly the same number of researchers stating those factors as encouraging. The least popular reason seems to be a benefit for the researcher's institution (38%). These figures might be of interest for investigations prior to the establishment of institutional repositories.

The main factor that would discourage researchers from sharing their research data was given as the potential risk of premature broadcast (68%). Possible risks to commercialisation opportunities or ethical constraints do not seem to feature very highly as factors that would discourage biosciences researchers from sharing their research data. This could be explained by the fact that these parameters are only applicable for rather few biosciences research projects. It is worth noting that one interviewee stated that the amount of data in his field was too large to be shared at all and that he would not want 'to spend hours and hours of sifting through other people's data'.

Two interviewees stated that they did not share any of their data at all (and that they were negative towards the StORe project aims). The reason given by one interviewee was that he considered his findings to belong to his employer only (the university) and thus would only share data within the university's groups. Another researcher stated that he feared increased competition for funding and 'stealing of results' by other researchers. This same reason was given by a third researcher who would only freely share teaching materials and would otherwise give access to his research data to known individuals not from 'the outside world'.

Access to research data seems to be given mainly on individual request (30% of all answers). Some interviewees stated relatively specific restrictions, such as the GNU license terms[24] for free software or even the inherent complexity of their research data. However, about 37% of the biosciences researchers stated that they do not employ any formal restrictions for access to their research data, making this the highest percentage of all answers.

Most researchers in the biosciences give limited access to their research findings only since they control access through storage on standalone devices (39%) or storage on a private network/intranet (30%). This is consistent for both questionnaire respondents and interviewees. Procedures which are more formalised, such as the maintenance of an approved user list (2 answers), or the reference of data requests to a review

[24] The GNU Public License is the most commonly used free software license, granting the recipients of a computer program the rights to run, copy, modify and redistribute a program, provided that the source code is always made available and that no further restrictions are imposed on any copies.

authority (1 answer), seem to be important for few researchers only. About 25% of researchers stated that they do not have any access control to their data in place.

2.7 Output repositories

The main output repositories for retrieving research information and finding teaching information or depositing publications seem to be publisher repositories, followed by discipline repositories. The use of institutional repositories was stated by fewer researchers. It is possible that the Manchester biosciences research community, as focus of the survey for the biosciences, is not yet used to an institutional repository due to a current lack of such a facility at Manchester. However, throughout all seven disciplines, institutional repositories did not seem to feature very highly in the choice for output repositories. This could possibly indicate either a lack of awareness or a potential scope for repository improvement.

Most bioscience researchers gave PubMed as their main choice of output repository (23% of all who named output repositories). PubMed does indeed provide conveniently free access and excellent usability, as stated by many interviewees. It features a well-developed functionality, for example comprehensive cross-linking to other widely used repositories such as GenBank, which could account for making it the repository of choice for many researchers.

More than 50% of the interviewees stated a habit of browsing as part of their general information search strategy during their research work.
Three interviewees (25%) stated that they prefer to read research publications in paper format. One said that he reads those printouts on the bus on his way home, which would make the electronic links within the appendices of no use to him. Two interviewees stated explicitly that they read everything on-line and never in paper format.
One interviewee explained the collaborative searching method within his research group. His group performs about 50 keyword searches per week, and they also scan the table of contents of all important journals. The research group have split the task of searching between them, with all of them being aware of the fields of interest of their colleagues.

The main access route for accessing the contents of output repositories differed slightly for questionnaire responders and interviewees, with the use of an internet search engine being preferred by questionnaire responders, closely followed by journal web sites and via a known repository URL (48%, 45% and 43% of all questionnaire answers, respectively). For interviewees, the main access routes were subject portals or a known repository's URL (both 75% of all interviewees). This question allowed for multiple answers.

The group of interviewees did not seem to be quite homogenous. Within this group, the use of subject portals was the most important route for researchers who had not previously answered the questionnaire (Non-Responders). The use of an internet search engine or a known repository's URL were the most important routes for researchers who had also answered the questionnaire (Responders). This may point to a higher independency from internet search engines for bioinformaticians/molecular biologists and/or a greater preference for direct access routes by this group. However, in both questionnaire and interviews other options also featured relatively highly, making the access routes to output repositories within the biosciences quite diverse.

60% of all bioscience researchers prefer simple searching when using an output repository. Nearly a quarter (23%) use advanced search modes and only 10% employ Boolean logic. None of the researchers uses subject thesaurus/subject headings. This observation was made for both questionnaire responders and interviewees.

2.8 Support

Personal support provided by an intermediary seems to be important for many researchers, with more than a third (39%) of all biosciences researchers stating that they receive such support. However, nearly a quarter (24%) stated that they receive no support at all, thus giving a diverse picture of the support provided. The main means of support provided by library or knowledge management staff were stated as on line/telephone help and the provision of formal training/documentation.

As half of the interviewees do not receive any support in using output repositories (50%), they may rely on they presumably high degree of computer literacy and/or self-sufficiency. Judging from the individual interviews, in those cases where remote support was given (41%), it was given through online help and not by telephone.

Possible improvement of standards, functionality and terminology were all suggested by questionnaire respondents and interviewees as factors that would enhance their searching.

2.9 Cross-disciplinary information research (additional topic)

One of the main outcomes of the StORe project will be an improved functionality of source and output repositories with subsequently improved information searching across disciplines. It was therefore thought to be of interest to explore the biosciences researchers' views on cross-disciplinary information search and common access standards.

Two thirds of the interviewees stated that they access information from different disciplines on a regular basis, with the different disciplines being in most, but not in all cases related to the biosciences. The specific disciplines depended on the individual researcher's field of research, so were for example mathematics (as source of algorithms useful for bioinformatics) and geology (as determining soil structure which in turn determines plant ecology) given as interests for cross-disciplinary information research.

Some researchers seem to favour personal communication over digital access. One reason stated for this attitude was that the different terminology could be clarified best in a conversation. Differing terminology was also given as reason for unsuccessful information research, together with inherently unreliable information, insufficient metadata capture, and various other reasons. It is interesting that one researcher would not want to access cross-disciplinary source data (spectroscopic data) because of the potentially large file size.

Altogether, the interviewees' opinion on a common access standard for different disciplines was rather varied, with some researchers being doubtful about practicalities.

2.10 Help functions (additional topic)

This topic was explored because it became clear during the interviews that although many interviewees did not receive any repository support on a regular basis, most of them had some experience in the occasional use of an on-line help function.

Many researchers could see a potential benefit of technically improved help functions, but their views on the usefulness of these functions was somewhat mixed. This may be due to the researchers' varying experience with on-line help functions, ranging from useful FAQ lists to unhelpfully rigid keyword options.

When asked about alternatives to help functions, personal communication routes, on-line encyclopaedias such as KEGG[25] and step-by-step guides were named. One researcher stated that he would prefer to 'try things out', since he considers this to be the most informative procedure. Again, this might reflect a high level of IT competence and self-sufficient information searching.

[25] Kyoto Encyclopedia of Genes and Genomes

3. Data from questionnaire and interviews

3.1 Identities

▪ Question 1a: In order that we might better understand the way you generate and use data, please would you identify your own role by selecting from the following list: *University academic staff, University research assistant, Postgraduate student, Undergraduate student, Contracting researcher, Independent researcher, Other.*

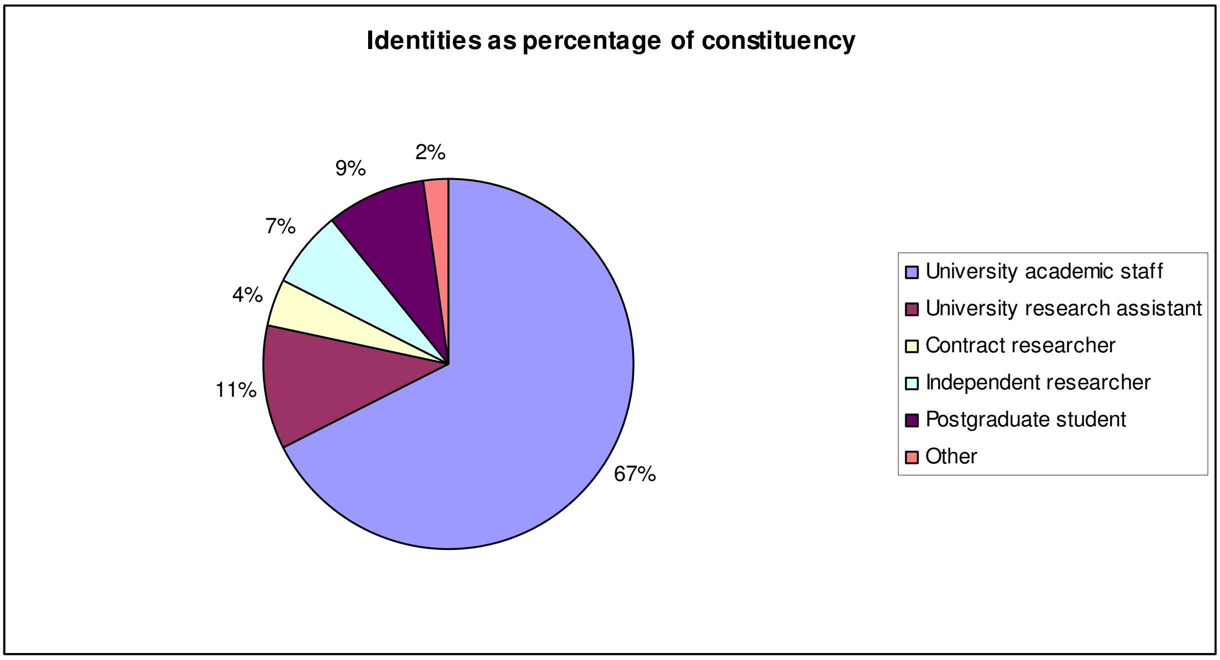
▪ 67% of the StORe questionnaire responses came from university academic staff, with significantly fewer responses from university research assistants, postgraduate students and other researchers. Similar observations were made in many of the other seven different disciplines.
It should be noted that out of 12 interviewees, six had previously answered the questionnaire (all of them university academic staff) while six had not previously answered the StORe questionnaire. This latter group consisted of three university academic members of staff and three university research assistants, respectively.

Table 2: Identities

Category	**University academic staff**	**University research assistant**	**Contract researcher**	**Independent researcher**	**Postgraduate student**	**Other**
Questionnaire	28	2	2	3	4	1
Interviews	9	3	0	0	0	0
Total numbers	*31*	*5*	*2*	*3*	*4*	*1*
Total percentage	*67%*	*11 %*	*4%*	*7%*	*9%*	*2%*

Other = Clinical Scientist (researcher in the NHS)

Figure 1: Identities



Other = Clinical Scientist (researcher in the NHS)

▪ **Question 1b: What are your main fields of interest?**

▪ Many researchers stated more than one field of interest. It was judged that a narrower categorizing than the RAE system would reflect the diversity of interests in the biosciences best. The categories were therefore set as a combination of the various discipline areas at the University of Manchester and fields of interest, as stated by the researchers.

The fields of interest represent both the focus of research at the University of Manchester (as the main part of the biosciences constituency) and the current focus within biosciences in general. Within the field of biosciences research, molecular biology has experienced a significant growth within the last decades, which makes it an important field of current interest for many researchers.
Most interviewees gave bioinformatics or molecular biology as their field of interest. This probably has implications for their handling of data and their level of IT expertise.

It is worth noting that some more specialised interests, such as cancer research, often border onto different disciplines, such as medicine. This reflects an overall trend of more fluid borders in modern biosciences, which has presumably an effect on cross-disciplinary co-operation as well as information research across disciplines.

Table 3: Fields of interest

Interest field category	**Questionnaire**	**Interview**	***Total***
Molecular biology	10	4	*14*
Biological sciences general	9		*9*
Bioinformatics	2	5	*7*
Ecology	6		*6*
Biochemistry	3	2	*5*
Neuroscience	5		*5*
Pharmacology	4	1	*5*
Physiology	5		*5*
Cancer research	3	1	*4*
Plant biology	2	2	*4*
Behavioural science	3		*3*
Evolution	3		*3*
Microbiology	2	1	*3*
Developmental biology	1		*1*
Immunology	1		*1*
Medical Education	1		*1*

3.2 Project aims

▪ **Question 2: 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?**

▪ Most (i.e. more than 80%) of the biosciences researchers seem to support the StORe project aims. In their answers, they stated that they would find improved links from both source-to-output and output-to-source repositories a significant advantage to their work, or at least useful.

Interviewees seemed to fall within two roughly equal, distinctive groups with three being positive about improved repository links (even already using such a feature), and three being rather uninterested or negative, with the remaining six not expressing any opinion. However, the numbers are too small to give more than an indication of an underlying trend.

Table 4: Advantage of source-to-output repository links

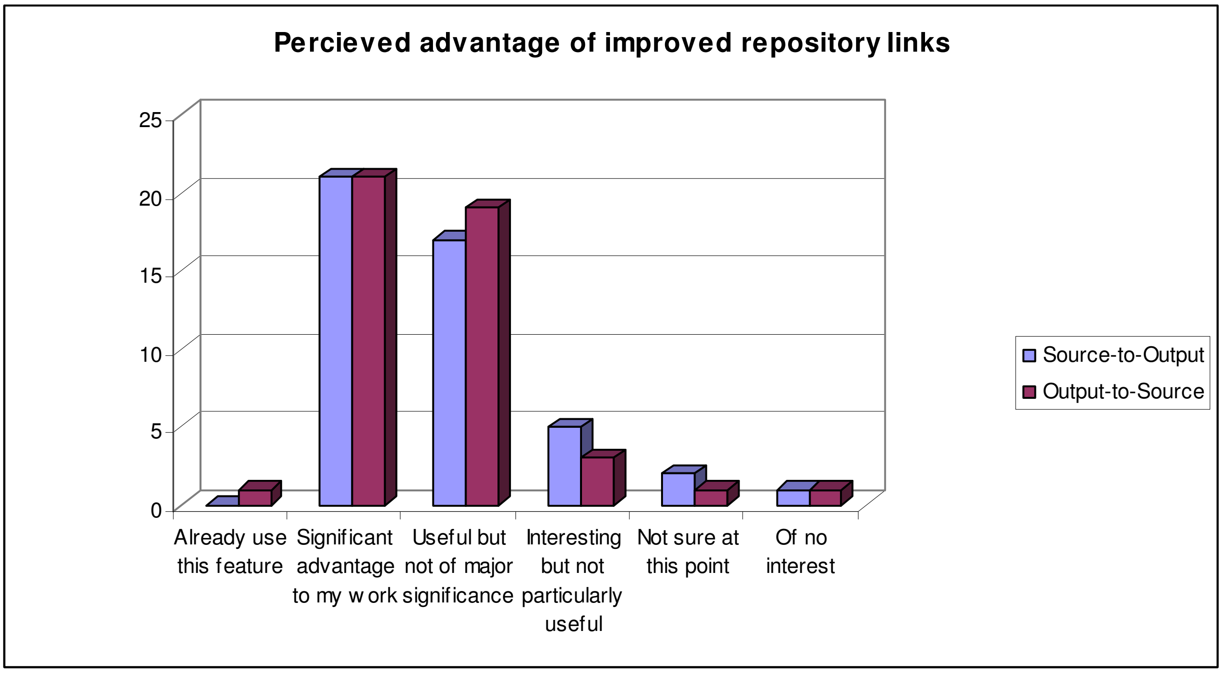
Category	**Significant advantage to my work**	**Useful but not of major significance**	**Interesting but not particularly useful**	**Not sure at this point**	**Of no interest**
Questionnaire	18	17	4	1	0
Interviews	3	0	1	1	1
Total	*21*	*17*	*5*	*2*	*1*

▪ **Question 3: How advantageous to you would it be if it were possible to go directly from within an online publication (electronic article or other text) to the primary source data from which that publication was developed?**

Table 5: Advantage of output-to-source repository links

Category	**Already use this feature**	**Significant advantage to my work**	**Useful but not of major significance**	**Interesting but not particularly useful**	**Not sure at this point**	**Of no interest**
Questionnaire	0	19	19	2	0	0
Interviews	1	2	0	1	1	1
Total	*1*	*21*	*19*	*3*	*1*	*1*

Figure 2: Perceived advantage of improved repository links



The bars represent the total numbers from the above tables.

▪ Question 26: 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 existing functionality seems to be adequate for a number of researchers. However, many answers reflect a need for better standardisation within source repositories and probably also indicate a lack of expertise in output repository search strategies. This could make better self-explanatory features or even improved help functions for output repositories desirable.

Please see next pages for a detailed overview about the individual answers.

Table 6: Missing functionalities from source repositories

Standards, Formats:
No clear data standards (3 entries)
Formats inconsistent (2 entries)
No clear metadata standards
Functionality and usability:
Generally OK, but it seems that often the description of how the data were generated has to fit in prescribed fields and it might help if submitters could add a set of 'notes' on anything peculiar in the methods or findings, or any caveats that might be considered when analysing the data.
Data hidden behind unhelpful titles – not knowing what is there
Not rigorously updated
Search facilities could be improved
Perhaps I would feel better about posting my data if a log was kept of who had access to the data and that some kind of agreement of proper use was associated with downloading it.
Other:
Often, personal contact is better in finding the right information.
Registration – giving away too many details
Would like to have a discussion forum
Nothing missing / not applicable:
None (5 entries)
Not applicable (2 entries)
Never used (1 entry)

Table 7: Missing functionalities from output repositories

Search strategy:
Problems with choice of keywords
Not enough/no good search assistance (2 entries)
Library and databases:
Several databases have to be scanned to ensure good coverage of the field
It would be helpful if there was one database that would have EVERYTHING on. As it is, I need to check at least 5-6 different databases as each comes up with different results.
Literature selection not broad enough
Linked cross referencing is OK at the moment, but could be more extensive to provide full text for all references. Literature selection can be a bit elective (PubMed)
More links to actual papers. I have to use Google Scholar for that.
Limited access due to high subscription costs (2 entries)
Missing compatibility with Endnote
Other:
Links through to web sites and articles do not always work
No consistent layout between pages
No consistent search facilities
Would appreciate a discussion forum
No site for practical information in my field
Registration or plug-in. It creates an instant barrier when you have to do something in order to get through to the information.
Difficulties in navigating the web site (2 entries)
More and different metadata should be captured
Nothing missing / not applicable
No problems and I like the layout and handling of outputs in PDF.
None/ works very well (8 entries)
Not applicable (1 entry)

3.3 Source data

▪ **Question 4: What kinds of electronic source data do you produce?**

▪ The biosciences researchers stated that they produce a wide range of different data, ranging for example from spectra to videos. The main data types within this range are images, drawings/plots, raw data and gene/protein sequences. A preference for these four data types was noted especially for the interviewees.

Table 8: Data types produced

Data Type	**Questionnaire**	**Interviews**	***Total***
Images	24	2	*26*
Drawings, Plots	21	1	*22*
Raw data	16	5	*21*
Gene/protein sequences	19		*19*
Databases	16	1	*17*
Text-based files	11	5	*16*
Statistical Data	14		*14*
Derived data	10		*10*
Instrument data	7	1	*8*
Photographs	8		*8*
Spectra	3	2	*5*
Plans, Maps	3		*3*
Qualitative questionnaire data	3		*3*
Quantitative questionnaire data	3		*3*
Video	3		*3*
Audio	1		*1*
Radiographic data	1		*1*
Spatial coordinate files	1		*1*
Synthetic data	1		*1*
Topographical data	1		*1*

▪ **Question 5: In what formats are these source data held?**

▪ Data formats are mostly consistent with the aforementioned data types. It should be noted that gene/protein sequences are usually written in plain text or FastA, which is another text format. This fact, together with the writing of experimental protocols, could account for the significant amount of text produced. Electrophoresis and microarray images, but also medical images such as CAT scans, are most often stored as JPG files, which may account for the high occurrence of (JPG, TIF, BMP, GIF).

Table 9: Data formats produced

Data Format	**Questionnaire**	**Interviews**	**Total**
Spreadsheets (e.g. Excel)	30	3	33
Image files (e.g. JPG, TIF, BMP, GIF)	27	4	31
DOC	28		28
Plain text (TXT)	16	4	20
PDF	16		16
Database files (e.g. Access, MySQL)	13		13
Statistical software	11		11
Tables/catalogues	8		8
RTF	5		5
HTML	4		4
ASCII		3	3
CAD/GIS	2		2
XML	2		2
Flat files (e.g. FITS)	1	1	2
MIAME		2	2
Endnote	1		1
Quick Time	1		1
Binary files		1	1
MST		1	1

▪ **Question 6: Are the data you generate sometimes a combination or groups of different data?**

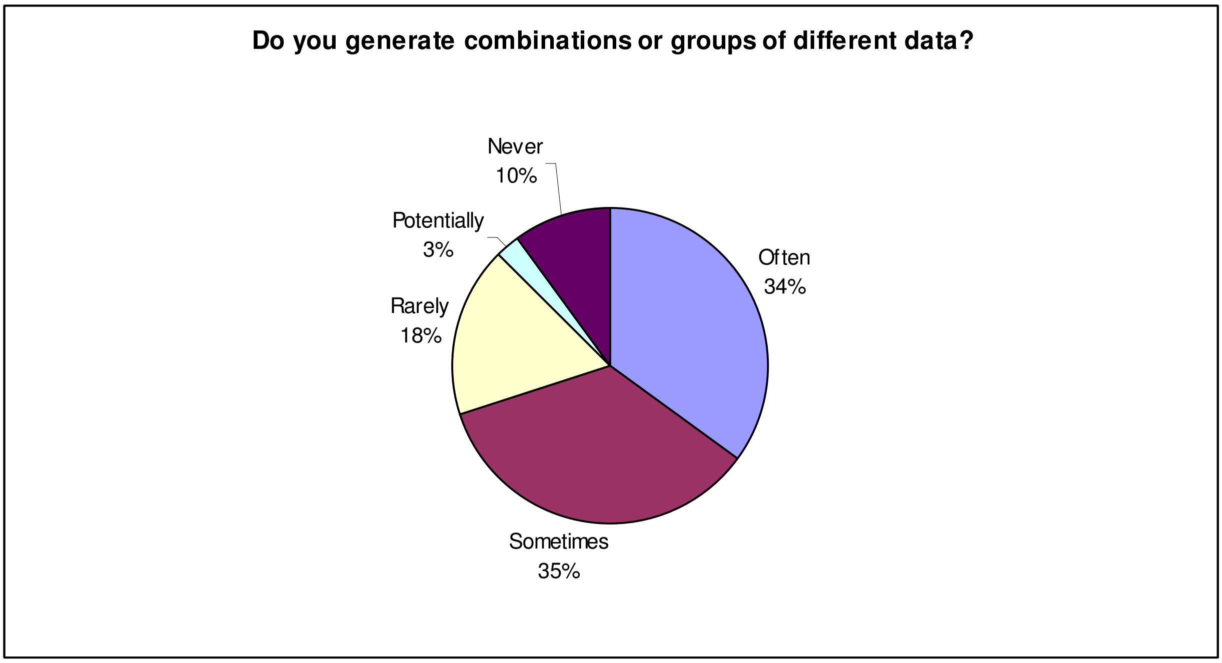
▪ Most researchers (70%) stated that they often or sometimes generate a combination or groups of different data, making this prevalent behaviour for the biosciences.

Table 10: Combination data

Category	Often	Sometimes	Rarely	Potentially	Never
Answer count	14	14	7	1	4
Answer percentage	35%	35%	17.5%	2.5%	10%

Counts are from the questionnaire only

Figure 3: Combination data



Counts are from the questionnaire only.

▪ **Question 12: Why might you wish to access the research data generated by other research programmes?**

▪ More than half of the researchers stated their motivation for accessing research data to be the gaining of own knowledge through accessing data (75%) and understanding data in context (50%). This shows a clear focus on the advancement of one's individual knowledge, rather than being primarily interested in evaluating other researchers' output. Indeed, testing others' objectives and identifying useful contacts seems to be of lesser importance (13% each).

Table 11: Reasons for accessing research data

Category	To access data	To understand context	To test my objectives	To test their objectives	To identify useful contacts
Answer count	30	20	15	11	11

Counts are from the questionnaire only

The reasons were:
To access data that are useful or necessary to my research
To understand the broader context and orientation of my research
To test the uniqueness and validity of my research objectives
To test the uniqueness and validity of their research objectives
To identify useful contacts

▪ Question 13: How would you normally access the research data of other researchers? (more than one answer possible)

▪ Online access seems to be the most important route for accessing the data from other researchers. However, a significant number of biosciences researchers also stated that they do not normally access other researchers' data.

Table 12: Research data access routes

Category	Online access	Portable media	Own networked fileservers	Other networked fileservers	E-mail attachments	Do not access
Questionnaire	22	10	7	6	4	12
Interviews	2	2	2			2
Total count	*24*	*12*	*9*	*6*	*4*	*14*

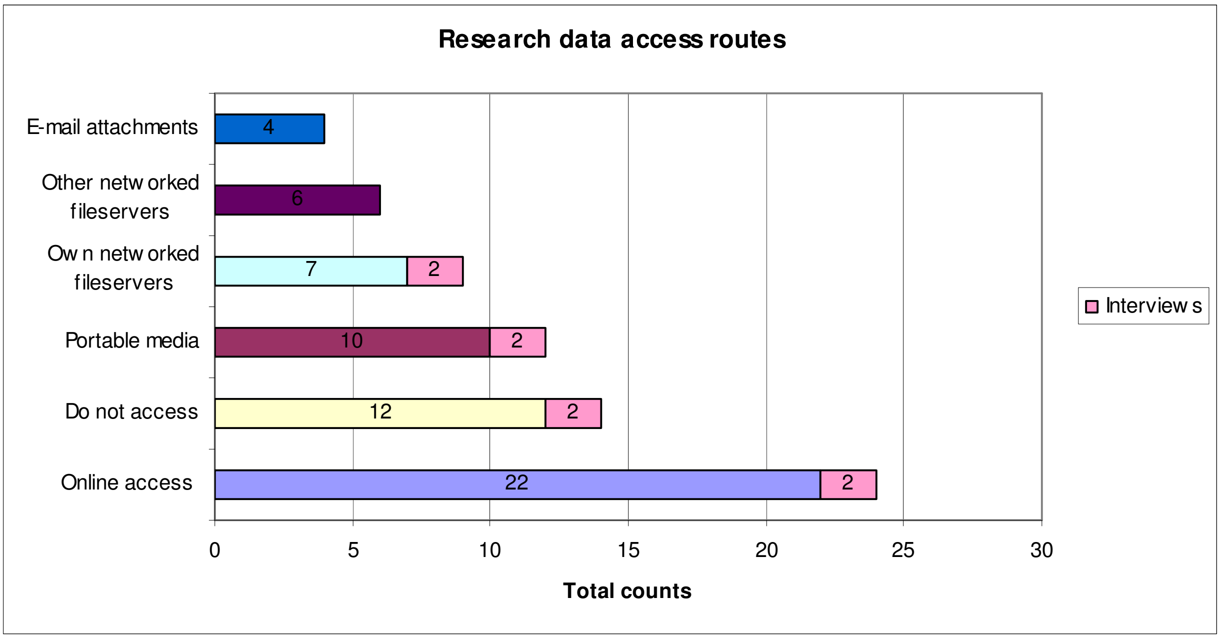
Counts are from the questionnaire only.

The access routes were:
Through online access to source repositories
I do not normally access others' research data
Through the exchange of data held on portable media (disks, CD-ROMs, USB drives, etc.)
By access to networked fileservers at my own institution
By access to networked fileservers at other institutions
Other: Please specify (all respondents wrote 'e-mail attachments' here)

Other options mentioned in the interviews were:

Personal communication:
By personal contact – the scientific community in my field is very small and I know most of my fellow researchers in the field.
We apply for grants together with other people and are fully aware of what they are doing.
We get quite a lot of information through collaborations and word of mouth.
Other:
Through news alerts containing clickable links but they only link through to very broad overviews.

Figure 4: Research data access routes



3.4 Source repositories

▪ Question 7: To which source repositories do you submit your data?

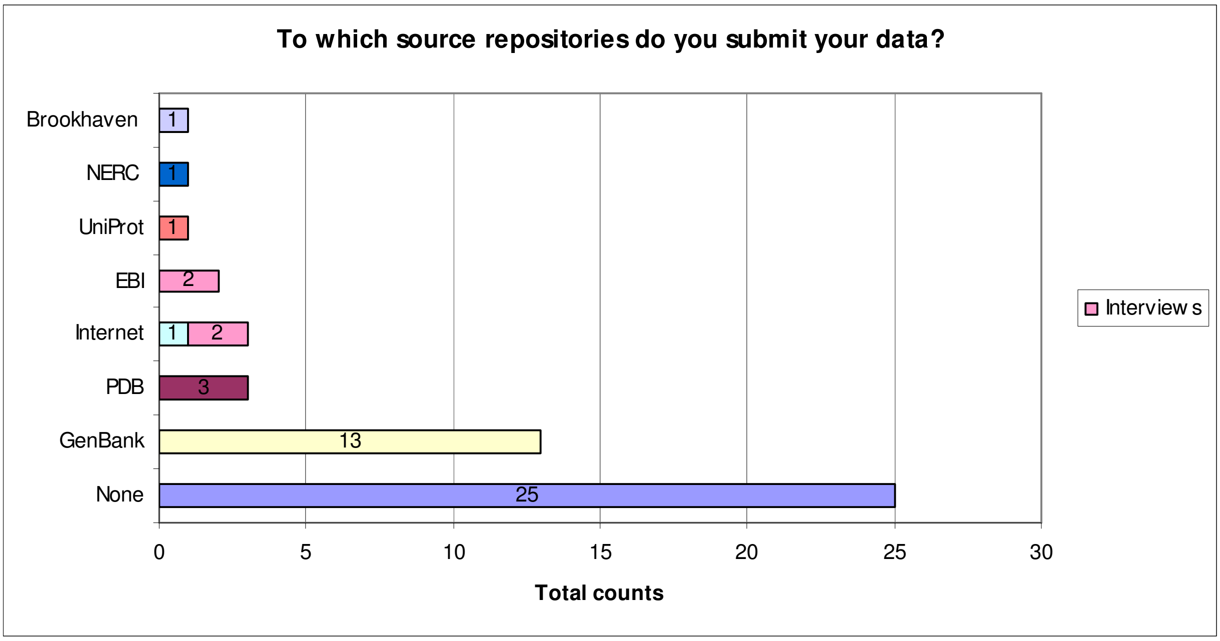
▪ Surprisingly, about half of the researchers stated that they do not submit their source data to any source repository. Most researchers who submit do so to GenBank (about 25% of all answers).

It should be noted that the submission to the unnamed 'Internet' could possibly include one of the named online access repositories and that EBI may possibly UniProt.

Table 13: Source repositories

Name	None	GenBank	PDB	Internet	EBI	UniProt	NERC	Brookhaven
Questionnaire	25	13	3	1	0	1	1	1
Interviews	0	0	0	2	2	0	0	0
Total count	*25*	*13*	*3*	*3*	*2*	*1*	*1*	*1*

Figure 5: Source repositories



▪ **Question 8: How often have you submitted data to any of these source repositories?**

▪ The numbers are consistent with question 7, i.e. questionnaire respondents who have stated a named repository also give the frequency of their submissions. Most questionnaire respondents stated that they submit to GenBank on an occasional basis.

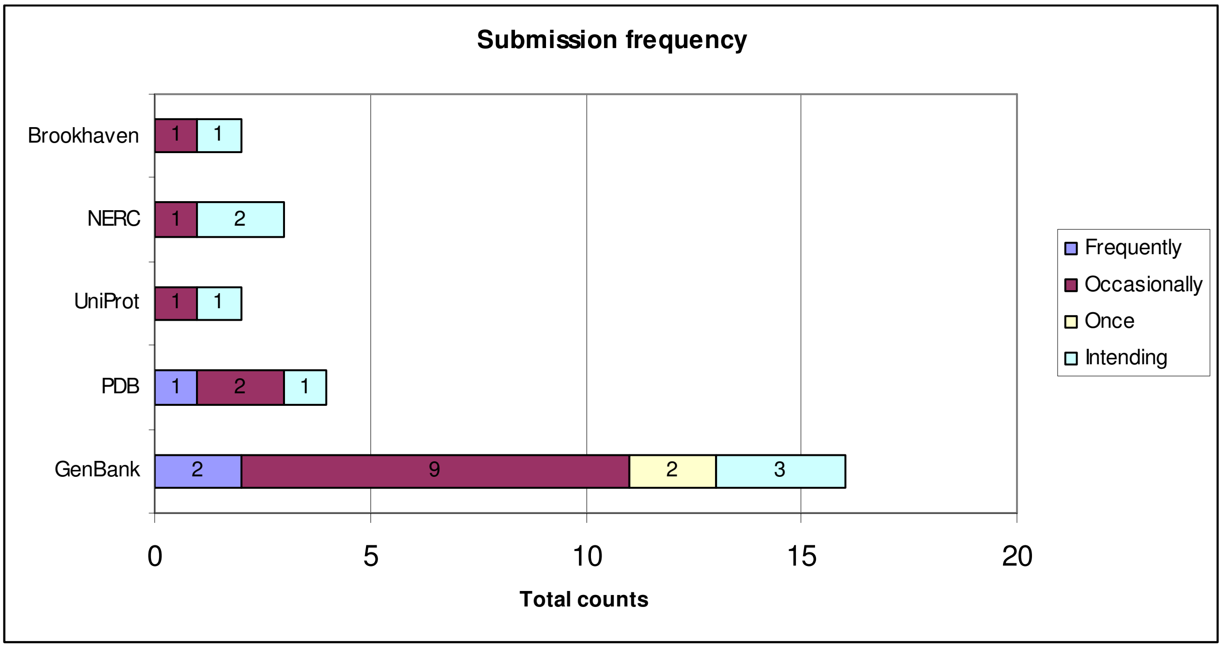
Table 14: Submission frequency

Submission frequency	GenBank	PDB	UniProt	NERC	Brookhaven
Frequently	2	1			
Occasionally	9	2	1	1	1
Once	2				
Intending	3	1	1	2	1
Never	17	26	15	26	26

As percentage of all answers:

Submission frequency	GenBank	PDB	UniProt	NERC	Brookhaven
Frequently	1.5%	0.8%			
Occasionally	6.6%	1.5%	0.7%	0.7%	0.7%
Once	1.5%				
Intending	2.2%	0.7%	0.7%	1.5%	0.7%
Never	12.4%	19%	11%	18.9%	18.9%

Figure 6: Submission frequency



Submission frequency 'Never' is not depicted in this figure. Counts are from the questionnaire only.

3.5 Metadata

▪ **Question 9: By selecting from the following options, please would you indicate what types of metadata you consider it important to assign to your data.**

▪ The main types of metadata assigned to biosciences research data are the name of the author, the date, subject keywords and project title. Interviewees seem to put greater importance to links and the description of experimental conditions than questionnaire responders.

Table 15: Metadata type

Metadata Type	**Questionnaire**	**Interviews**	***Total***
Author or data creator name	33	5	*38*
Date (e.g. of data creation)	31	3	*34*
Subject keywords	30	3	*33*
Project title	28	5	*33*
Project description	23	3	*26*
Data format	24		*24*
Title of data set	23		*23*
Project reference number or identifier	15	2	*17*
Dates of project	15		*15*
Funding source	8		*8*
Publisher	8		*8*
Description of experimental conditions	1	6	*7*
Database accession numbers		5	*5*

Links to websites, networked files		5	*5*
Description of data variables	1		*1*
Don’t understand the term metadata	1		*1*

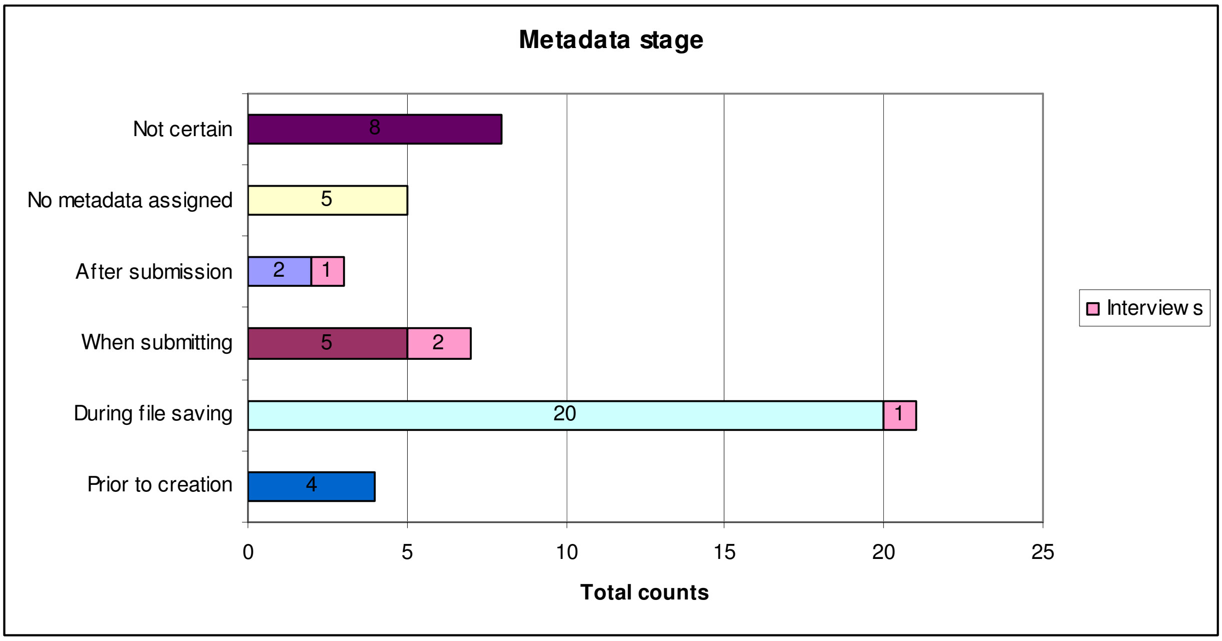
▪ **Question 10: At what stage are metadata assigned to your data?**

▪ Most researchers seem to assign metadata during the stage of file saving, with a significant proportion not being certain at what stage metadata is assigned.

Table 16: Metadata stage

Category	**Prior to creation**	**During file saving**	**When submitting**	**After submission**	**Not certain**	**No metadata assigned**
Questionnaire	4	20	5	2	8	5
Interviews		1	2	1		
Total count	*4*	*21*	*7*	*3*	*8*	*5*

Figure 7: Metadata stage



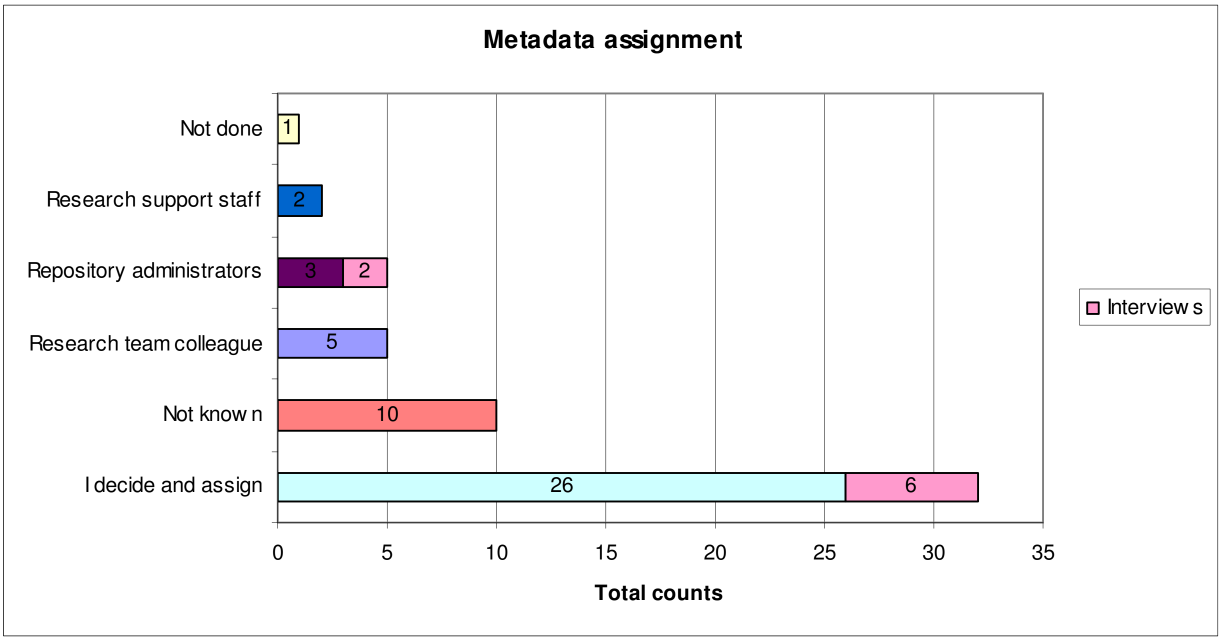
▪ **Question 11: Who assigns metadata to your research data?**

▪ Assigning metadata to research data seems to be a task which is mainly carried out by the individual researcher. To a lesser extent, a team colleague or repository administrator is responsible for assigning metadata.

Table 17: Metadata assignment

Category	I decide and assign	Research team colleague	Repository administrators	Research support staff	Not done
Questionnaire	26	5	3	2	1
Interviews	6		2		
Total	*32*	*5*	*5*	*2*	*1*

Figure 8: Metadata assignment



3.6 Data access and sharing

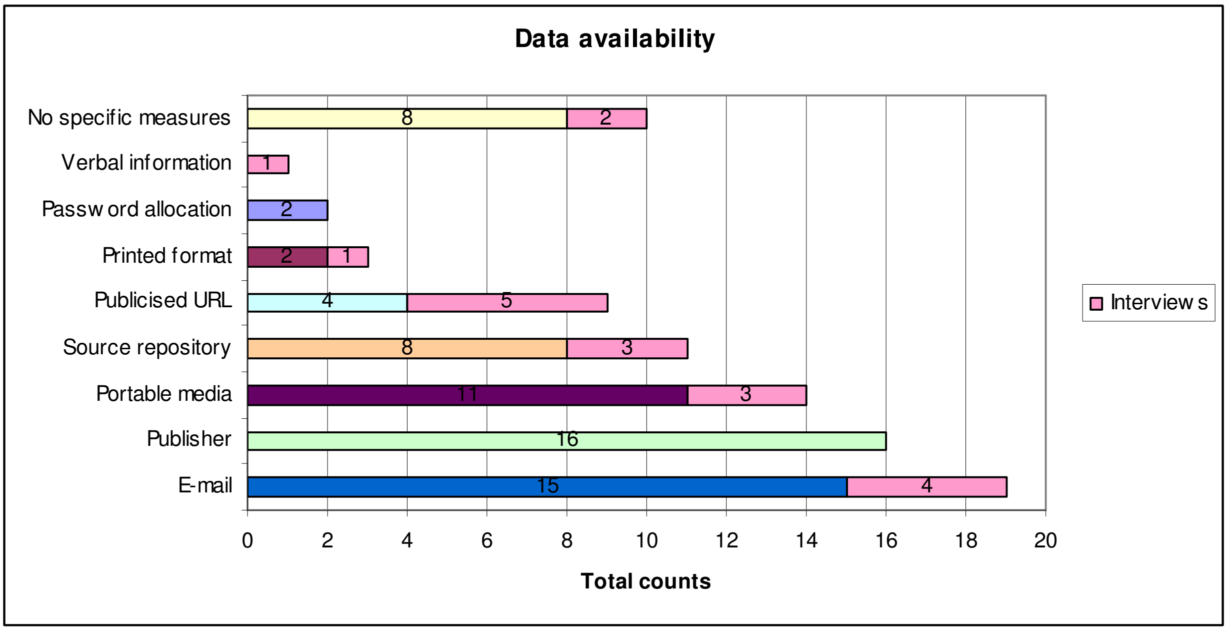
▪ Question 14: What measures do you use to make your research data available?

▪ Researchers in the biosciences field seem to exchange data mainly through e-mail or via a publisher, followed by the exchange of portable media, through a source repository and by the provision of a publicised URL. However, a substantial amount of bioscientists do not undertake any measures to make their research data available.

Table 18: Data availability

Data exchange through …	Questionnaire	Interviews	Total
Data are distributed via e-mail	15	4	*19*
Via a publisher	16		*16*
Through the exchange of portable media	11	3	*14*
Through a source repository	8	3	*11*
By the provision of a publicised URL	4	5	*9*
Data are posted or passed by hand in printed format	2	1	*3*
By the allocation of passwords to networked files at my institution	2		*2*
Verbal information exchange only		1	*1*
I undertake no measures to make my research data available	8	2	*10*

Figure 9: Data availability



For additional information, see also 5.4. **Data access and sharing**, page 70

▪ **Question 15: What factors would encourage you to share your research data?**

▪ Most biosciences researchers state individual reasons, such as improved visibility, as a factor that would encourage them to share their data. However, practical reasons such as a requirement of the funding body, or altruistic reasons, such as benefits to the research community, also feature highly. The least popular reason seems to be a benefit for the researcher's institution, even though the difference in response counts here is not very large.

Table 19: Factors in favour of sharing data

Factors	Answer Count
Improved visibility	25
Benefit to my profile	24
Benefits to the research community	24
Funding body requirement	24
Benefits for society	23
Improved validation	21
Enabling collaboration	19
Benefits to my institution	15

Counts are from the questionnaire only.

The factors were:
Improved visibility for my research
Demonstrable benefit to my research profile (this may be improved status, future funding or new research prospects)
Potential benefits to the research community
Requirement of funding body/condition of funding
Potential benefits to society
Improved level of validation for my research findings
Enabling collaboration and contribution by others
Demonstrable benefit to my institution (research presence, income, etc.)

For additional information, see also 5.4. **Data access and sharing**, pages 70/71

▪ **Question 16: What factors would discourage you from sharing your research data?**

▪ Most researchers seem to be concerned about a risk of premature broadcast as the main factor that would discourage them from sharing their research data. Possible risks to commercialisation opportunities or ethical constraints do not seem to feature very highly as a discouraging factor within biosciences.

Table 20: Factors against sharing data

Factors	Answer count
Risk of premature broadcast	27
Loss of ownership	22
Time/effort required to enable sharing	23
Risk of diversion from objectives	17
Intellectual property rights	17
Established research niche	14
Funding competition	8
Data protection	8
Ethical constraints	6
Risk to commercialisation	4

Counts are from the questionnaire only.

The factors against sharing data were:
Risk of premature broadcast of research findings
The threat of loss of ownership
The time/effort required to enable sharing
Risk of diversion from principal objectives through the generation of additional work
Subversion of intellectual property rights, including copyright
Risks to an established research niche
Increased competition for funding
Consideration of data protection and other confidentiality issues
Ethical constraints relating to my research
Risk to commercialisation opportunities

▪ **Question 17: Normally, what kind of formal restrictions do you apply to your research data?**

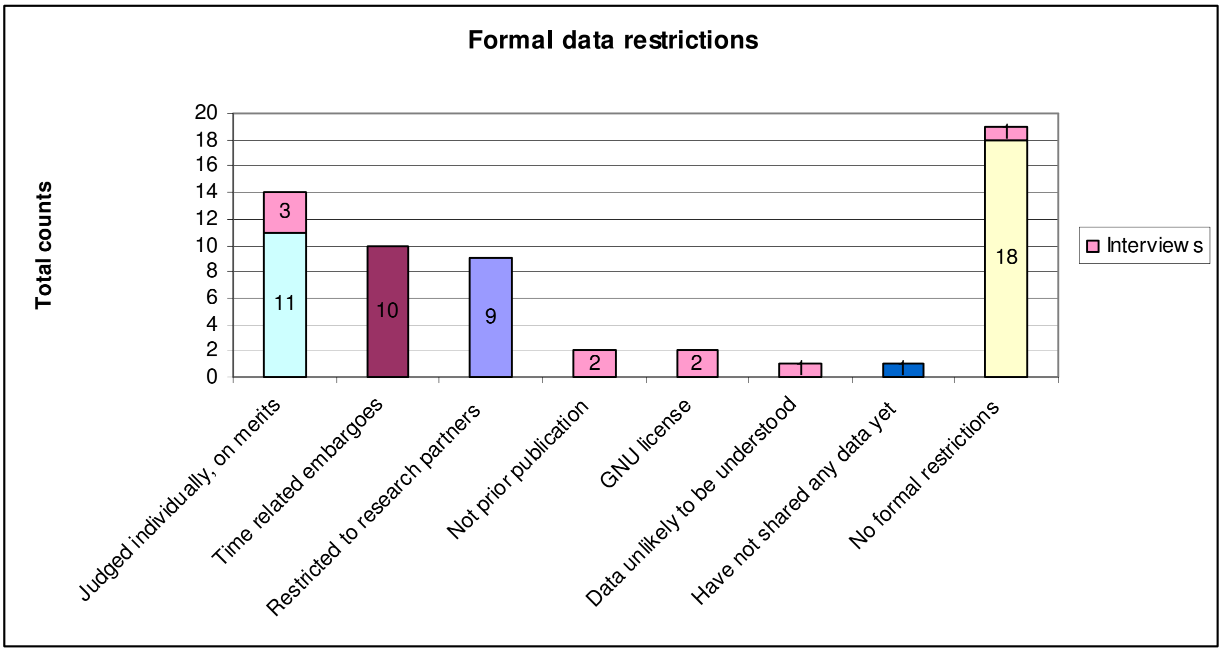
▪ Access to research data seems to be given mainly on individual request. Some interviewees use rather specific restrictions. However, most researchers in the biosciences stated that they do not employ any formal restrictions for access to their research data.

Table 21: Formal data restrictions

Restriction Type	Questionnaire	Interviews	*Total*
Judged individually, on merits	11	3	*14*
Time related embargoes	10		*10*
Restricted to research partners	9		*9*
Not prior publication		2	*2*
GNU license		2	*2*
Data unlikely to be understood		1	*1*
Have not shared any data yet	1		*1*
No formal restrictions	18	1	*19*

The restrictions were:
Individual enquiries/requests for access are judged on their merits
Time related embargoes
Restricted to immediate research team/programme members
Downloads are covered by GNU license terms
Our data is extremely complex and is unlikely to be understood by many people, and this serves as an access control
Have not shared any data yet
No formal restrictions

Figure 10: Formal data restrictions



▪ **Question 18: What measures do you normally use to control access to your data by others?**

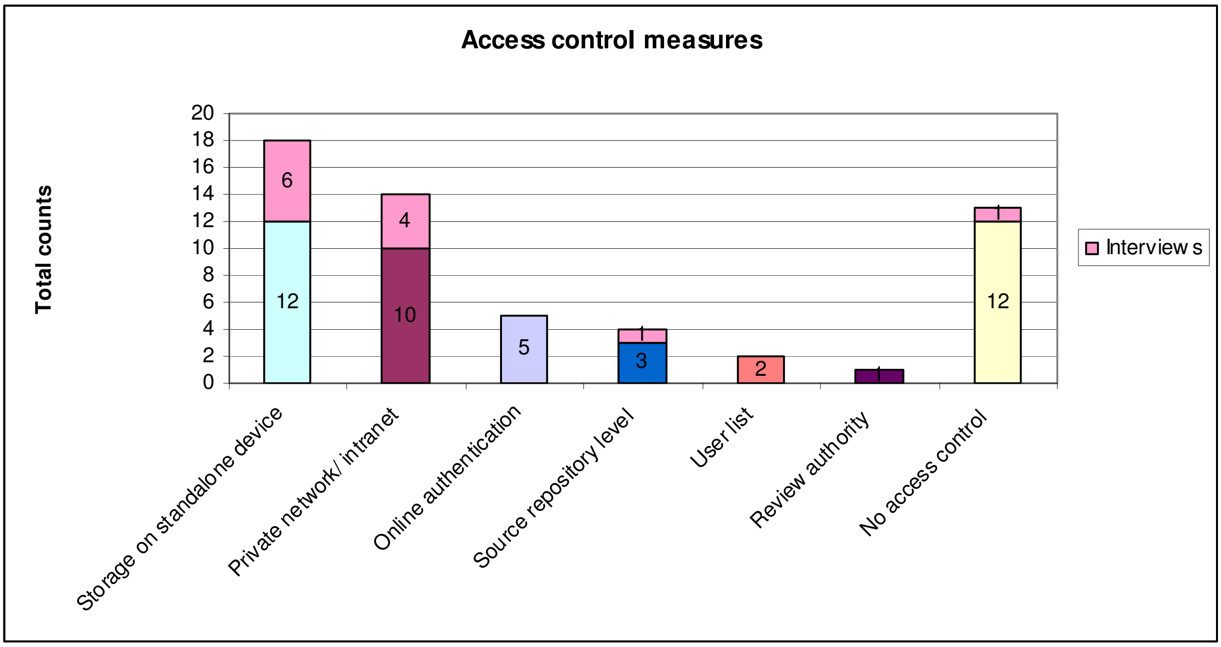
▪ Most researchers in the biosciences do not seem to be fully networked since they control access to their data through storage on standalone devices. This is closely followed by access control through data storage on a private network/intranet. More formal procedures, such as the reference of data requests to a review authority, do not seem to feature very highly as access control measures.

Table 22: Access control measures

Access control	**Questionnaire**	**Interviews**	***Total***
Storage on standalone device	12	6	*18*
Private network/ intranet	10	4	*14*
Online authentication	5		*5*
Source repository level	3	1	*4*
User list	2		*2*
Review authority	1		*1*
No access control	12	1	*13*

The measures were:
Storage of data on standalone computers (or experimental devices)
Storage of data on a private network/intranet
Authentication of ID and password for online access
The specific operational terms and conditions of the source repository
Maintenance of an approved list/directory of data users
Reference of data requests to a review authority
No access control – there is open access

Figure 11: Access control measures



3.7 Output repositories

▪ Question 19: Which kind of output repositories do you use to find and retrieve information for use in your research?

▪ Publisher and discipline repositories are important for retrieving research information, finding teaching information and depositing publications. Institutional repositories seem to be of less importance.

Table 23: Research output repositories

Category	Publisher	Discipline	Institutional	None
Questionnaire	28	24	15	2
Interviews	6	6		
Total	*34*	*30*	*15*	*2*

▪ Most bioscience researchers gave PubMed as their main choice of output repository (23% of all who named output repositories). PubMed does indeed provide conveniently free access and excellent usability, as stated by many interviewees. It features a well-developed functionality, for example comprehensive cross-linking to other widely used repositories such as GenBank, which could account for making it the repository of choice for many researchers.

Table 24: Research output repositories in detail

Name	User count
PubMed	10
Google, Froogle, Google Scholar	7
EBI	6
ISI WoK, WoS	4
NCBI	4
University library web site	4
Pub Med Central	2
SciFinder Scolar	1
Structure Databases	1
Promoter Databases	1
ATCC	1

The users were counted from both interviews and questionnaire

▪ **Question 20: Which kind of output repositories do you use to find and retrieve information for use in teaching?**

Table 25: Teaching output repositories

Category	Publisher	Discipline	Institutional	None
Questionnaire	20	20	11	8

Counts are from the questionnaire only.

▪ **Question 21: To which output repositories do you deposit your research publications?**

Table 26: Depositing in output repositories

Category	Publisher	Discipline	Institutional	None
Questionnaire	23	13	7	9

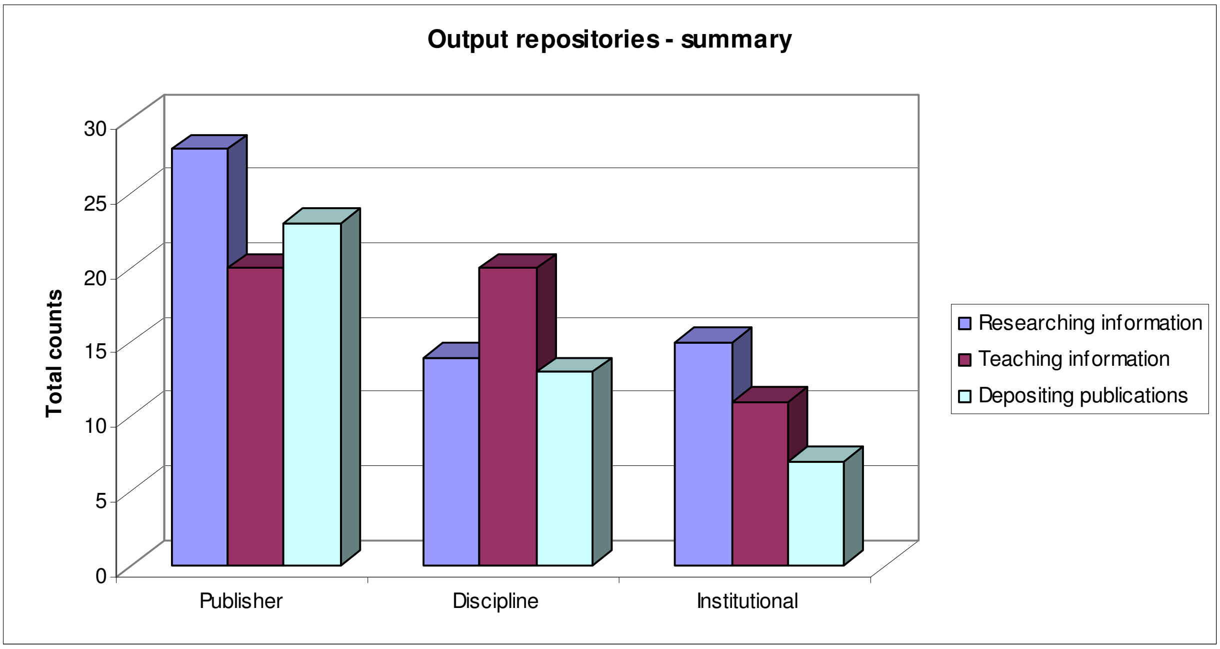
Counts are from the questionnaire only.

Table 27: Output repositories summary

Category	Publisher	Discipline	Institutional
Research information	28	14	15
Teaching information	20	20	11
Depositing publications	23	13	7

Counts are from the questionnaire only.

Figure 12: Output repositories summary



Counts are from the questionnaire only.

▪ **Question 22: What are your normal or preferred routes to the contents of output repositories?**

▪ The main access route for accessing the contents of output repositories differed slightly for questionnaire responders and interviewees, with the use of an internet search engine being preferred by questionnaire responders and subject portals or a known repository's URL by the interviewees.

Table 28: Routes to output repositories

Access route	**Questionnaire**	**Interviews**	***Total***
Repository URL	17	9	*28*
Journal web site	18	7	*25*
Internet search engine	19	6	*25*
Subject portal	11	9	*19*
Publisher's service	11	2	*13*
Library catalogue	11	2	*13*
Author's web page	7		*7*
E-mail link	5	1	*6*
Open URL resolver	1		*1*
Through an EndNote library	1		*1*
I have no normal or preferred routes	4		*4*

The routes were:
Directly through a specific journal's own web site
Via a known repository's URL
From an Internet search engine (e.g. Google)
Through a subject portal service (e.g. Entrez)
Through a publisher's online service (e.g. Science Direct)
Via a library catalogue that links directly to an article in a repository
Through an author's personal web page
From a link provided in an e-mail, CD-ROM, USB drive etc.
Via an Open URL resolver
Through an EndNote library
I have no normal or preferred routes

▪ **Question 23: What level of searching do you normally find sufficient when using an output repository?**

▪ Most bioscience researchers prefer simple searching when using an output repository. Nearly a quarter use advanced search modes and only a few employ Boolean logic. None of the researchers uses a subject thesaurus or subject headings. The interview and questionnaire answers showed both a similar trend.

Table 29: Levels of searching

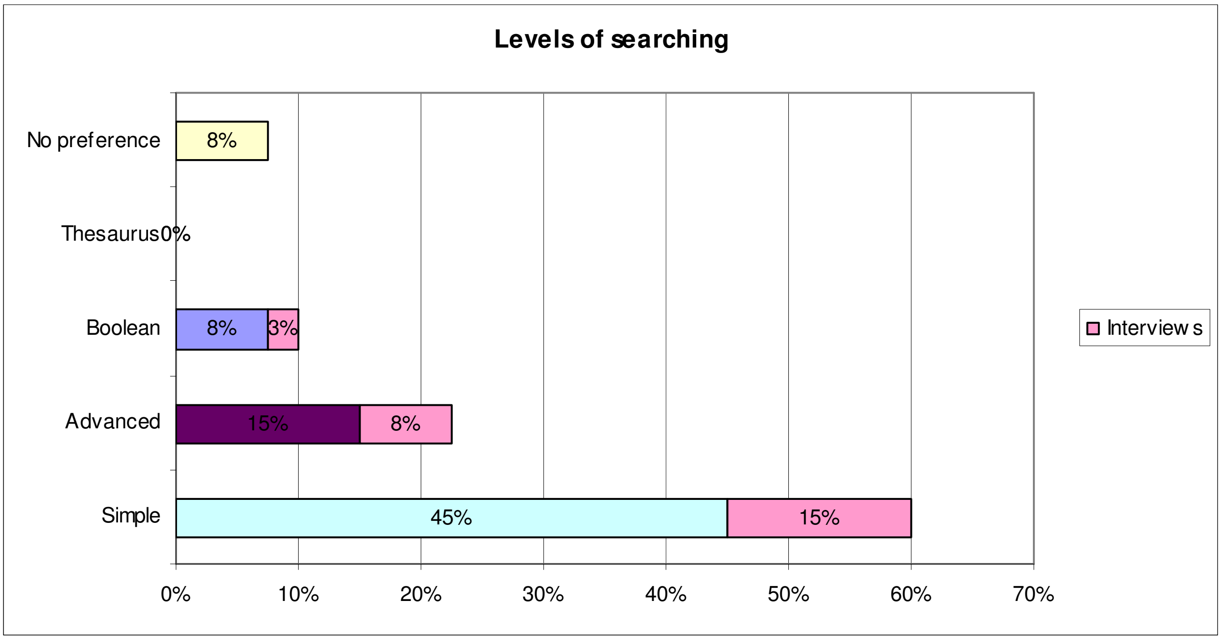
Category	Simple	Advanced	Boolean	Thesaurus	No preference
Questionnaire	18	6	3	0	3
Interviews	6	3	1	0	0
Total count	*24*	*9*	*4*	*0*	*3*

As percentage of searching:

Category	Simple	Advanced	Boolean	Thesaurus	No preference
Questionnaire	45%	15%	7.5%	0%	7.5%
Interviews	15%	7.5%	2.5%	0%	0%
Total percentage	*60%*	*22.5%*	*10%*	*0%*	*7.5%*

The levels were:
Simple – e.g. author, title, keyword, date
Advanced, using a range of fields and identifiers
Employing Boolean logic
Using a subject thesaurus or subject headings
No preference

Figure 13: Levels of searching



▪ **Question 23 a: What further options, features, or functionality would enhance your level of searching?**

Table 30: Search enhancing features

Standards:
Similar layout between journal web sites or subject portals (3 entries)
Functionality:
Being able to search full text rather than only title or abstract (as on Google Scholar)
Limit to one organism only
Alternative spellings (American/English) and abbreviations (e.g. beta and B)
Citation index important, and should be together with publication
Links from output repository to literature or web sites should be working and updated
Terminology:
I often don't seem to use the right keywords or terminology (2 entries)
Other:
University/Faculty should set up their own on-line journal, with a peer review system for the publications
Information units could be subject to small charge (like i-tunes). Index of how many downloads should be together with publication.
Revised format for articles would help searching: keep the title, authors, and abstracts. Skip the introduction and replace with links to reviews. The material & methods section is very important and can be in great detail. In the results section should be no explanations but simply a list of tables and figures. There can be active links in here. Discussion should be very brief.
Unambiguous gene names (2 entries), suggestion: list the gene names and show the counts on how often this particular name is being used in the repository. Or a good web site that lists gene orthologues.
Older maps should be available in digital format

3.8 Support

▪ Question 24: Do you receive support and/or guidance in your use of output repositories?

▪ Personal support provided by an intermediary seems to be important for many biosciences researchers, with more than a third stating that they receive such support. However, nearly a quarter stated that they receive no support at all, thus giving a diverse picture of the support provided. Many interviewees do not seem to receive any support, possibly preferring a self-sufficient search strategy.

Table 31: Output repositories support

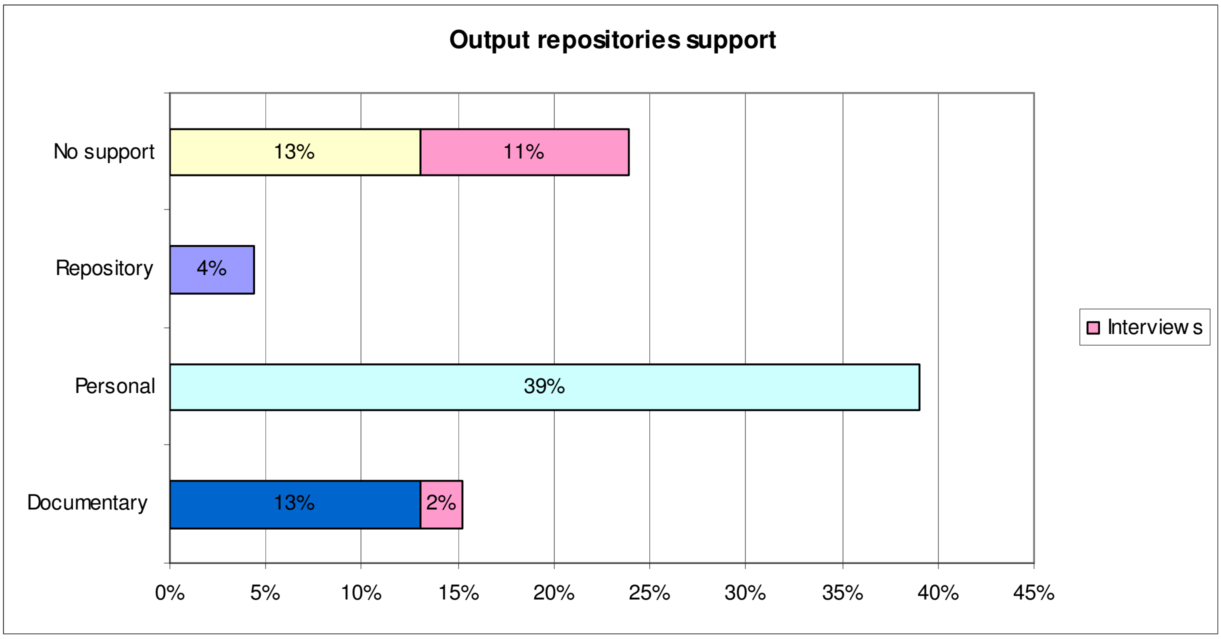
Support	Documentary	Personal	Repository	No support
Questionnaire	6	18	2	6
Interviews	1			5
Total count	*7*	*18*	*2*	*11*

As percentage of support received:

Support	Documentary	Personal	Repository	No support
Questionnaire	13.0%	39.1%	4.4%	13.0%
Interviews	2.2%			10.9%
Total percentage	*15.2%*	*39.1%*	*4.4%*	*23.9%*

The support options were:
Documentary support
Personal support provided by an intermediary
Repository-enabled support
No support is provided

Figure 14: Output repositories support



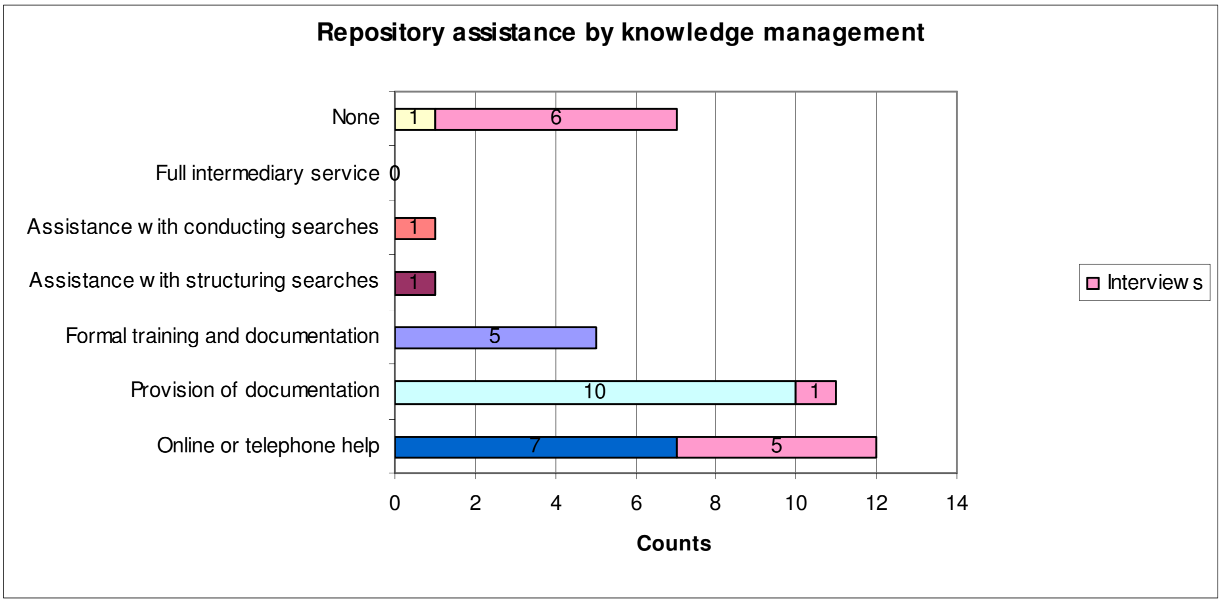
▪ **Question 25: What assistance in your use of repositories is provided by a librarian or other knowledge management support?**

▪ The main means of providing support by library or knowledge management staff seem to be on line/telephone help and the provision of formal training/documentation, with half of the interviewees stating that they do not receive any assistance through a librarian or knowledge management support.

Table 32: Repository assistance by knowledge management

Assistance	**Questionnaire**	**Interviews**	***Total***
Online or telephone help	7	5	*13*
Provision of documentation (guidance notes, fact sheets, etc.)	10	1	*11*
Formal training and documentation	5		*5*
Assistance with the structuring of specific searches	1		*1*
Assistance with the conduct of searches	1		*1*
Full intermediary service (e.g. the conduct of searches and organisation of results)	0	0	*0*
None	1	6	*7*

Figure 15: Repository assistance by knowledge management



4. Selected themes presentations with relationships (cross-tabulations) from the questionnaire

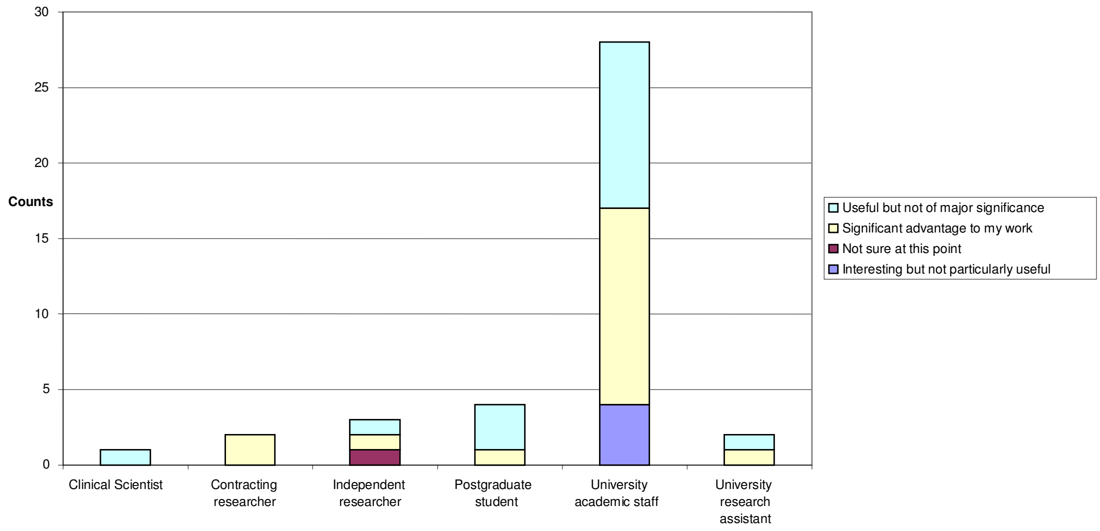
4.1 Value of source-output repository links relative to professional roles

▪ Answers to question 1 were depicted against the answers to question 2, showing the majority of university academic staff being in favour of improved source-to-output repository links.

Table 33: Value of improved source-to-output repositories links, by professional role

Professional role	Value of source-to-output repository links	Count of professional role
University academic staff		**28**
	Significant advantage to my work	13
	Interesting but not particularly useful	4
	Useful but not of major significance	11
University research assistant		**2**
	Significant advantage to my work	1
	Useful but not of major significance	1
Contracting researcher		**2**
	Significant advantage to my work	2
Independent researcher		**3**
	Significant advantage to my work	1
	Useful but not of major significance	1
	Not sure at this point	1
Postgraduate student		**4**
	Significant advantage to my work	1
	Useful but not of major significance	3
Clinical Scientist		**1**
	Useful but not of major significance	1
Grand Total		**40**

Figure 16: Value of improved source-to-output repository links, by professional role

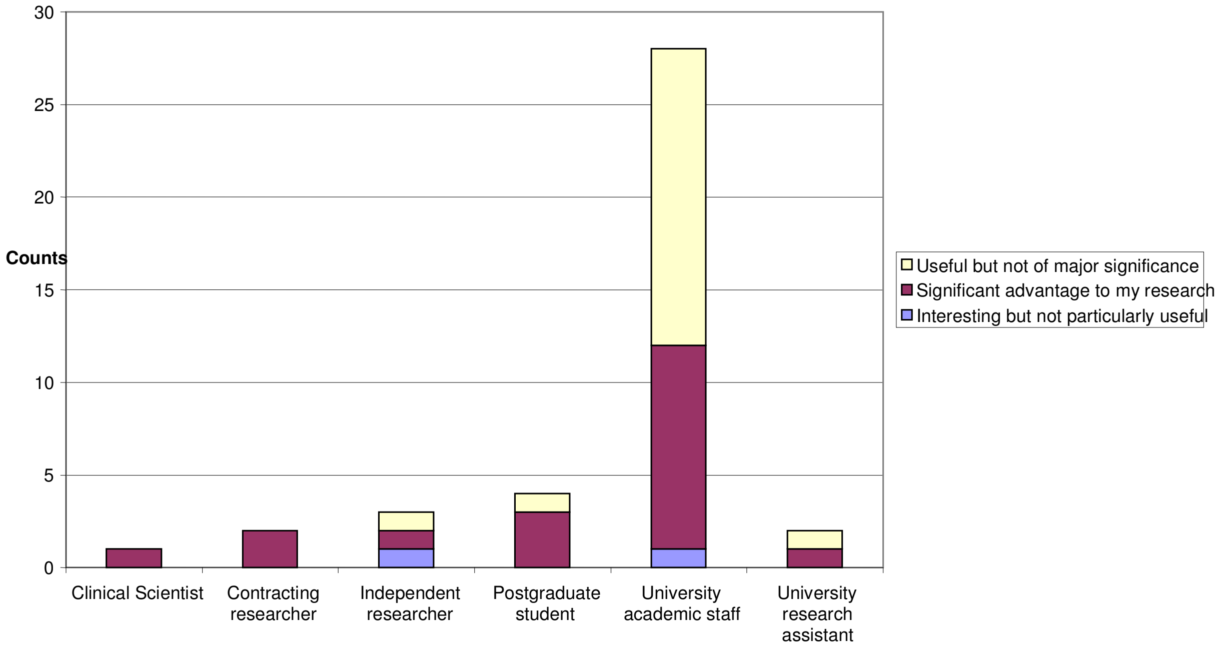


- Answers to question 1 were depicted against the answers to question 3, showing the majority of university academic staff being in favour of improved output-to-source repository links.

Table 34: Value of improved output-to-source repositories links, by professional role

Professional role	Value of output-to-source repository linkage	Count of Professional role
Clinical Scientist		**1**
	Significant advantage to my research	1
Contracting researcher		**2**
	Significant advantage to my research	2
Independent researcher		**3**
	Significant advantage to my research	1
	Useful but not of major significance	1
	Interesting but not particularly useful	1
Postgraduate student		**4**
	Significant advantage to my research	3
	Useful but not of major significance	1
University academic staff		**28**
	Significant advantage to my research	11
	Useful but not of major significance	16
	Interesting but not particularly useful	1
University research assistant		**2**
	Significant advantage to my research	1
	Useful but not of major significance	1
Grand Total		**40**

Figure 17: Value of improved output-to-source repository links, by professional role



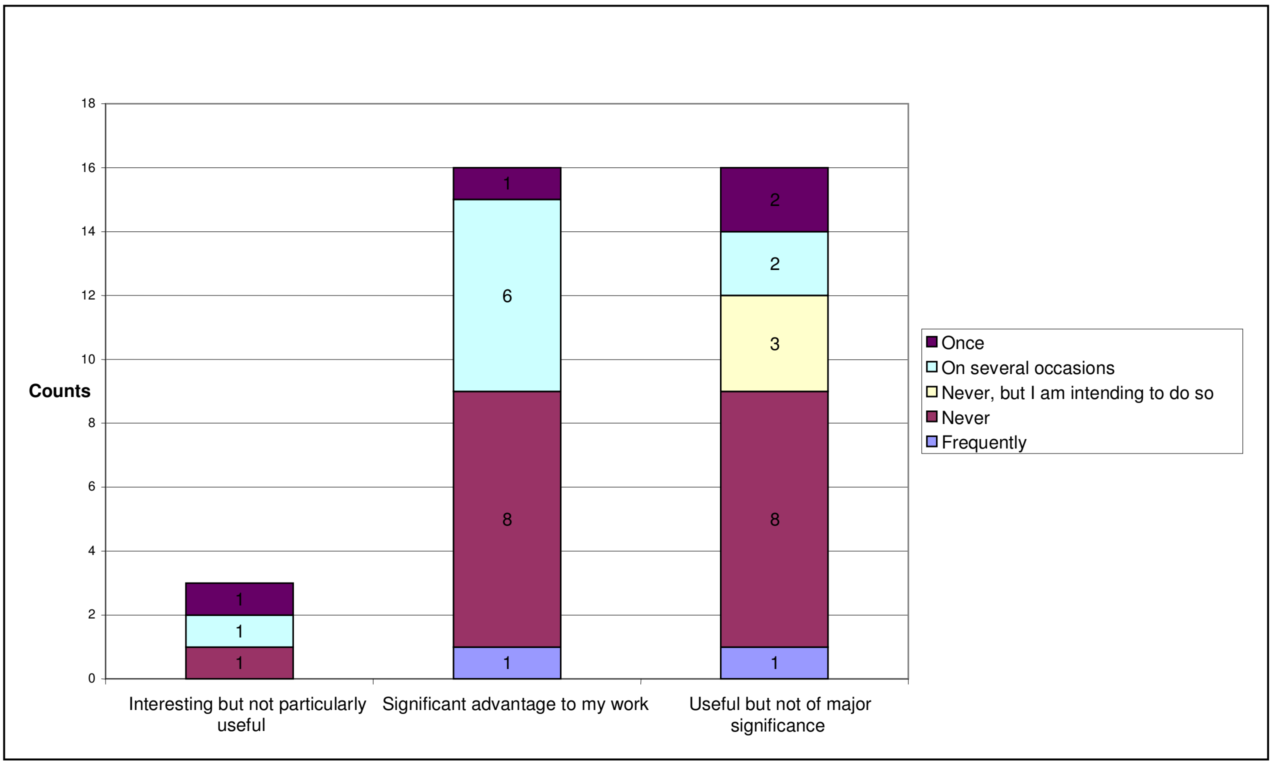
4.2 Value of source-output repository links relative to different repository communities

▪ Answers to question 2 were depicted against answers to question 8, showing GenBank users generally being in favour of linkage between source and output repositories. GenBank was chosen because it is the main important source repository in the biosciences (see also 3.4 Source repositories)

Table 35: Perceived value of source-to-output repository links, relative to GenBank users

Perceived value of source-to-output repository links	Submission frequency to GenBank	Count of submission frequency
Significant advantage to my work		**16**
	Frequently	1
	On several occasions	6
	Once	1
	Never	8
Useful but not of major significance		**16**
	Frequently	1
	On several occasions	2
	Once	2
	Never, but I am intending to do so soon	3
	Never	8
Interesting but not particularly useful		**3**
	On several occasions	1
	Once	1
	Never	1
Grand Total		**35**

Figure 18: Perceived value of source-to-output repository links, relative to GenBank users

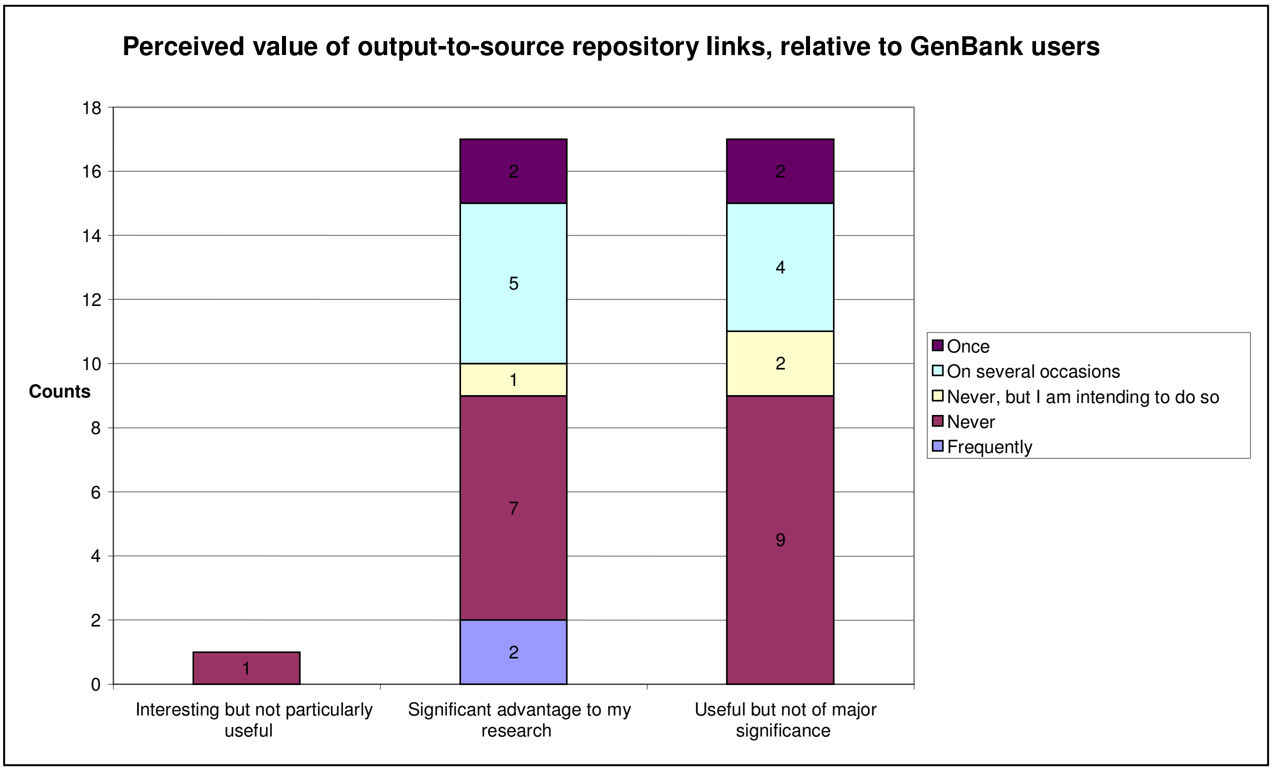


▪ Answers to question 3 were depicted against answers to question 8, showing GenBank users generally being in favour of linkage between source and output repositories.

Table 36: Perceived value of output-to-source repository links, relative to GenBank users

Link output-to-source repositories	Submission frequency to GenBank	Count of submission frequency
Significant advantage to my research		**17**
	Frequently	2
	On several occasions	5
	Once	2
	Never, but I am intending to do so soon	1
	Never	7
Useful but not of major significance		**17**
	On several occasions	4
	Once	2
	Never, but I am intending to do so soon	2
	Never	9
Interesting but not particularly useful		**1**
	Never	1
Grand Total		**35**

Figure 19: Perceived value of output-to-source repository links, relative to GenBank users



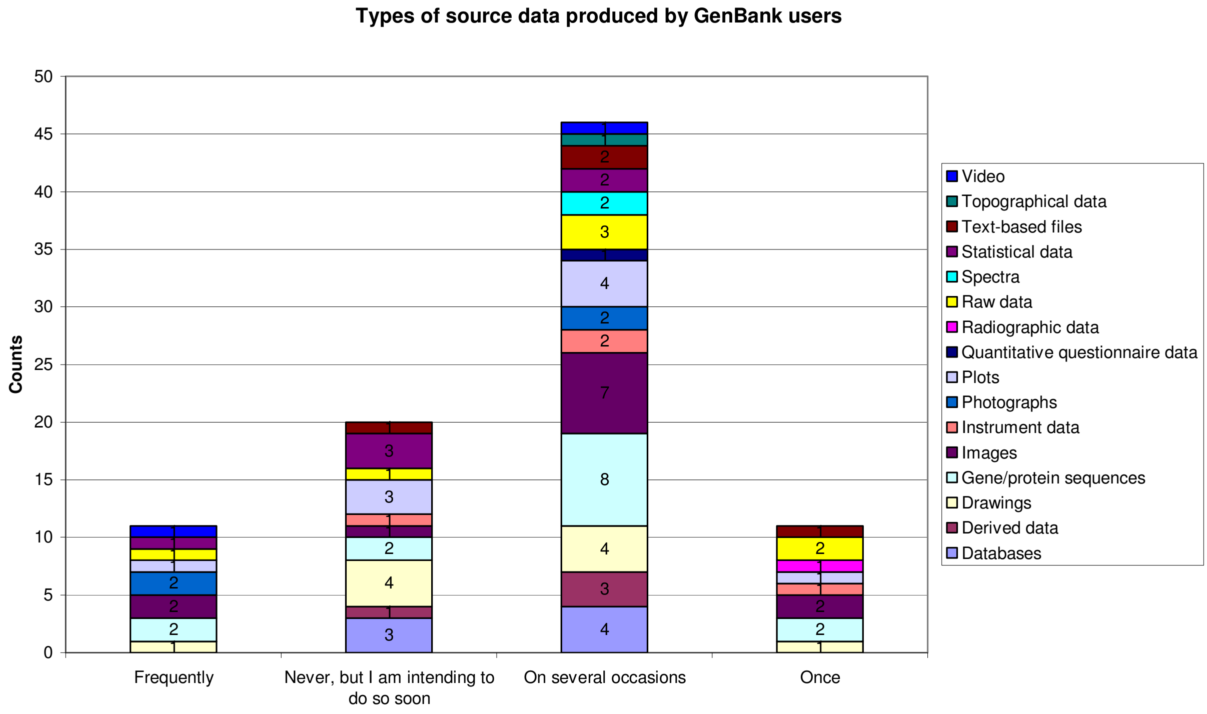
4.3 Different types of source data according to different repository communities

▪ Answers to question 4 were depicted against answers to question 8, showing GenBank users producing a variety of different source data types, especially when submitting only occasionally to GenBank. However, the submissions to GenBank follow a relatively strict format. See also Scenario 5 for further details.

Table 37: Source data produced by GenBank users

GenBank user	Source data produced	Count of source data produced
Frequently		**11**
	Gene/protein sequences	2
	Images	2
	Drawings	1
	Plots	1
	Photographs	2
	Raw data	1
	Statistical data	1
	Video	1
On several occasions		**46**
	Gene/protein sequences	8
	Images	7
	Databases	4
	Derived data	3
	Drawings	4
	Plots	4
	Instrument data	2
	Photographs	2
	Raw data	3
	Spectra	2
	Statistical data	2
	Text-based files	2
	Topographical data	1
	Video	1
	Quantitative questionnaire data	1
Once		**11**
	Gene/protein sequences	2
	Images	2
	Drawings	1
	Plots	1
	Instrument data	1
	Radiographic data	1
	Raw data	2
	Text-based files	1
Never, but I am intending to do so soon		**20**
	Gene/protein sequences	2
	Images	1
	Databases	3
	Derived data	1
	Drawings	4
	Plots	3
	Instrument data	1
	Raw data	1
	Statistical data	3
	Text-based files	1
Grand Total		**88**

Figure 20: Types of source data produced by GenBank users



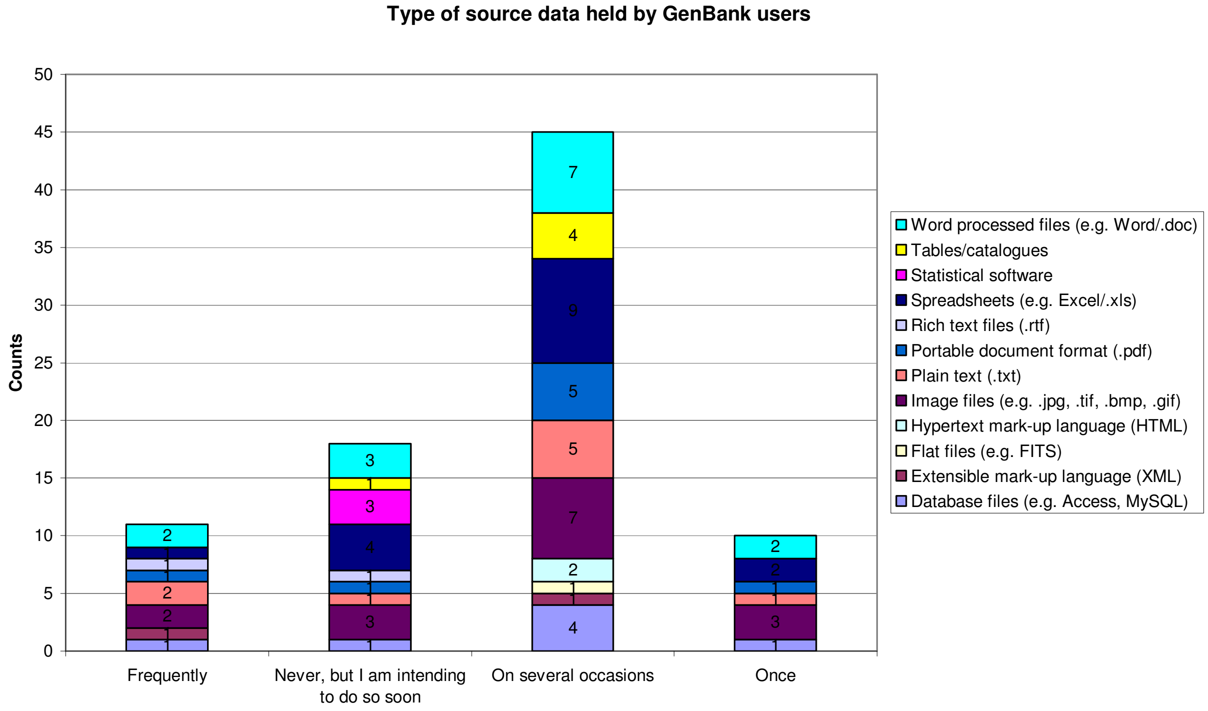
Different types of source data according to different repository communities

▪ Answers to question 5 were depicted against answers to question 8, showing that GenBank users hold source data in a variety of different formats, with text and images formats being predominant especially among users who submit on several occasions.

Table 38: Source data held by GenBank users

GenBank user	Source data held	Count of source data held
Frequently		**11**
	Spreadsheets (e.g. Excel/.xls)	1
	Word processed files (e.g. Word/.doc)	2
	Image files (e.g. .jpg, .tif, .bmp, .gif)	2
	Plain text (.txt)	2
	Portable document format (.pdf)	1
	Database files (e.g. Access, MySQL)	1
	Rich text files (.rtf)	1
	Extensible mark-up language (XML)	1
On several occasions		**45**
	Spreadsheets (e.g. Excel/.xls)	9
	Word processed files (e.g. Word/.doc)	7
	Image files (e.g. .jpg, .tif, .bmp, .gif)	7
	Plain text (.txt)	5
	Portable document format (.pdf)	5
	Database files (e.g. Access, MySQL)	4
	Tables/catalogues	4
	Hypertext mark-up language (HTML)	2
	Extensible mark-up language (XML)	1
	Flat files (e.g. FITS)	1
Once		**10**
	Spreadsheets (e.g. Excel/.xls)	2
	Word processed files (e.g. Word/.doc)	2
	Image files (e.g. .jpg, .tif, .bmp, .gif)	3
	Plain text (.txt)	1
	Portable document format (.pdf)	1
	Database files (e.g. Access, MySQL)	1
Never, but I am intending to do so soon		**18**
	Spreadsheets (e.g. Excel/.xls)	4
	Word processed files (e.g. Word/.doc)	3
	Image files (e.g. .jpg, .tif, .bmp, .gif)	3
	Statistical software	3
	Plain text (.txt)	1
	Portable document format (.pdf)	1
	Database files (e.g. Access, MySQL)	1
	Tables/catalogues	1
	Rich text files (.rtf)	1
Grand Total		**84**

Figure 21: Type of source data held by GenBank users



4.4 Metadata requirements against repository used

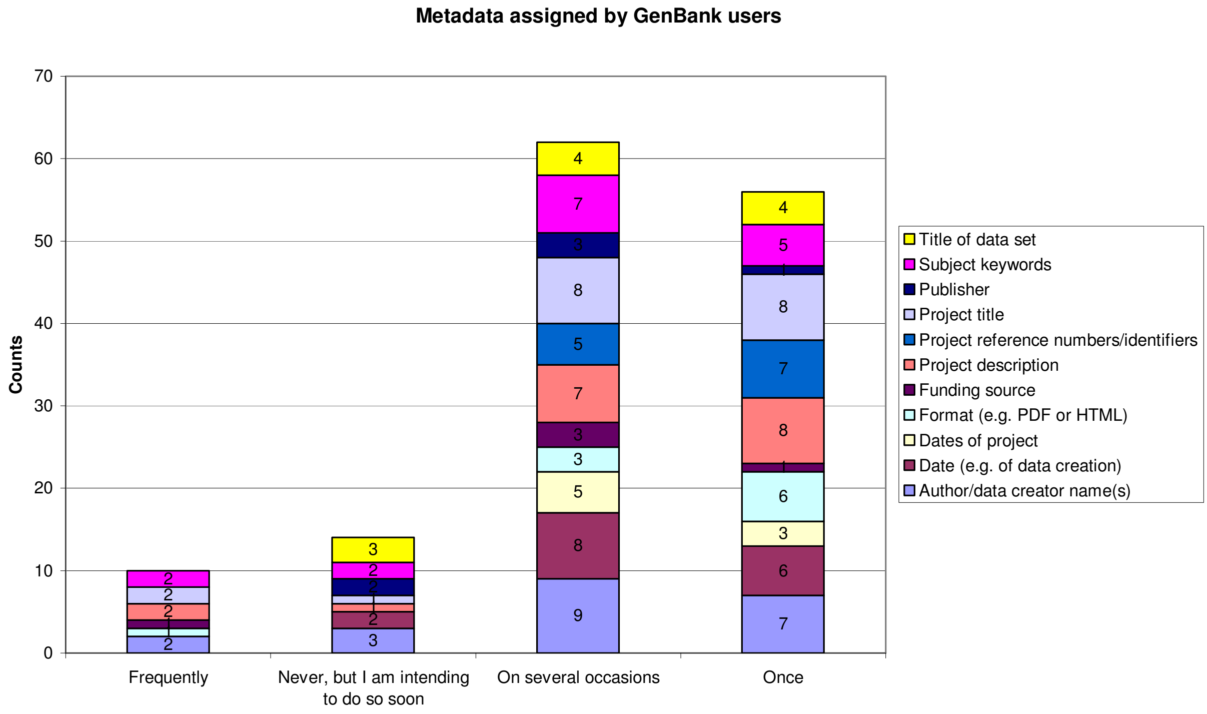
▪ Answers to question 9 were depicted against answers to question 8, showing title, author, date and project description being important for GenBank users, especially when submitting on several occasions.

Table 39: Metadata assigned by GenBank users

GenBank user	Metadata	Count of metadata
Frequently		**10**
	Author/data creator name(s)	2
	Project title	2
	Project description	2
	Subject keywords	2
	Format (e.g. PDF or HTML)	1
	Funding source	1
On several occasions		**62**
	Author/data creator name(s)	9
	Date (e.g. of data creation)	8
	Project title	8
	Project description	7
	Subject keywords	7
	Dates of project	5
	Project reference numbers/identifiers	5
	Title of data set	4
	Format (e.g. PDF or HTML)	3
	Funding source	3
	Publisher	3

Once		**56**
	Author/data creator name(s)	7
	Date (e.g. of data creation)	6
	Project title	8
	Project description	8
	Subject keywords	5
	Dates of project	3
	Project reference numbers/identifiers	7
	Title of data set	4
	Format (e.g. PDF or HTML)	6
	Funding source	1
	Publisher	1
Never, but I am intending to do so soon		**14**
	Author/data creator name(s)	3
	Date (e.g. of data creation)	2
	Project title	1
	Project description	1
	Subject keywords	2
	Title of data set	3
	Publisher	2
Grand Total		**142**

Figure 22: Metadata assigned by GenBank users



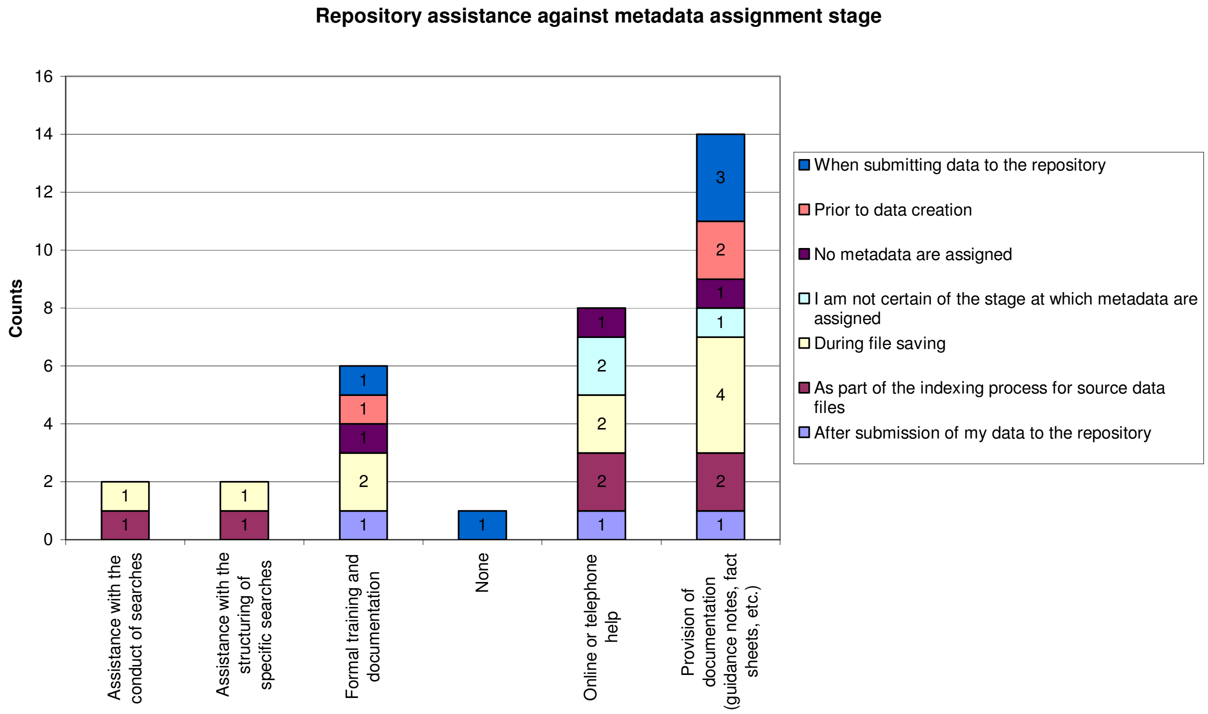
4.5 Metadata requirements and practice against library support

- Answers to question 10 were depicted against answers to question 25, showing a general preference of metadata assignment during file saving, especially in the case of researchers who would consult the documentation when accessing a repository.

Table 40: Metadata assignment stage compared to support level provided

Repository assistance	Metadata stage	Count of metadata stage
Provision of documentation (guidance notes, fact sheets, etc.)		**14**
	During file saving	4
	When submitting data to the repository	3
	As part of the indexing process for source data files	2
	Prior to data creation	2
	I am not certain of the stage at which metadata are assigned	1
	After submission of my data to the repository	1
	No metadata are assigned	1
Online or telephone help		**8**
	During file saving	2
	As part of the indexing process for source data files	2
	I am not certain of the stage at which metadata are assigned	2
	After submission of my data to the repository	1
	No metadata are assigned	1
Formal training and documentation		**6**
	During file saving	2
	When submitting data to the repository	1
	Prior to data creation	1
	After submission of my data to the repository	1
	No metadata are assigned	1
Assistance with the conduct of searches		**2**
	During file saving	1
	As part of the indexing process for source data files	1
Assistance with the structuring of specific searches		**2**
	During file saving	1
	As part of the indexing process for source data files	1
None		**1**
	When submitting data to the repository	1
Grand Total		**33**

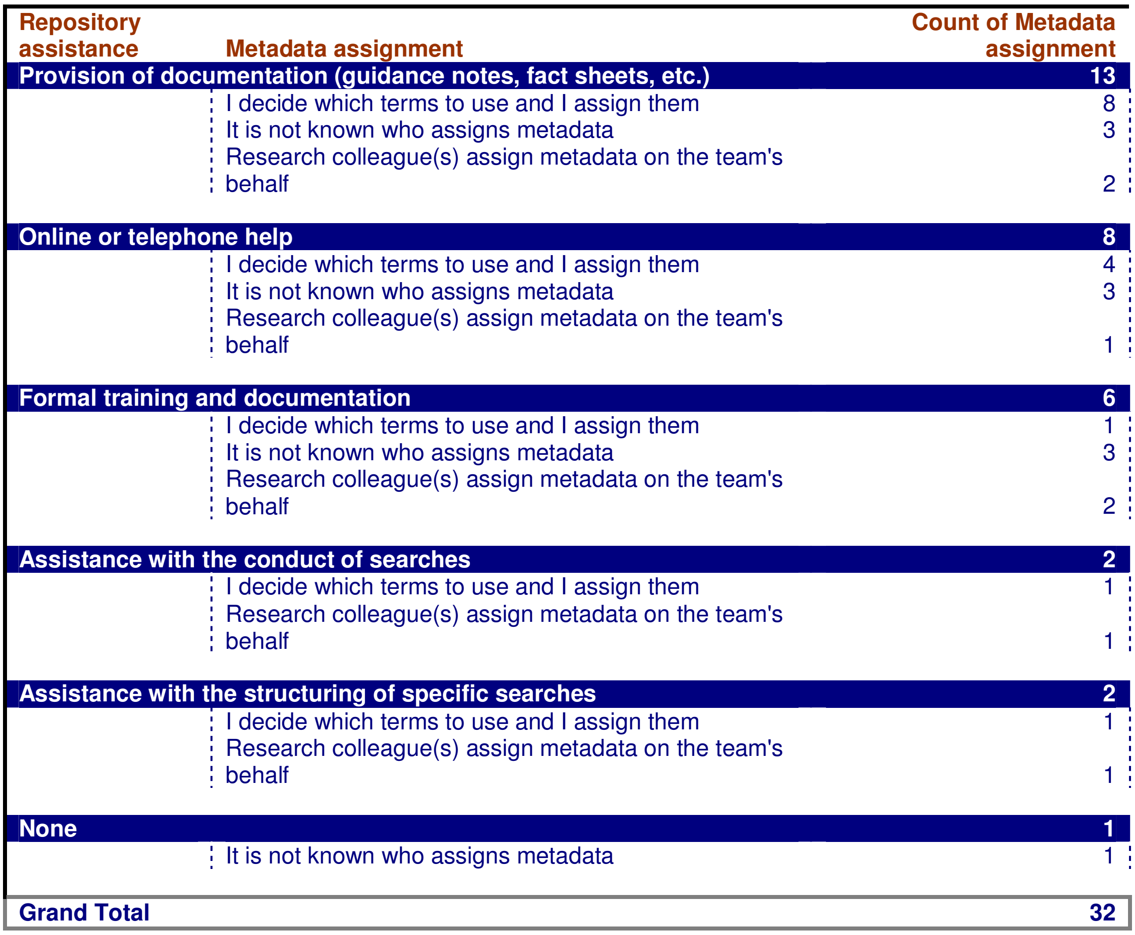
Figure 23: Repository assistance against metadata assignment stage



Metadata requirements and practice against library support

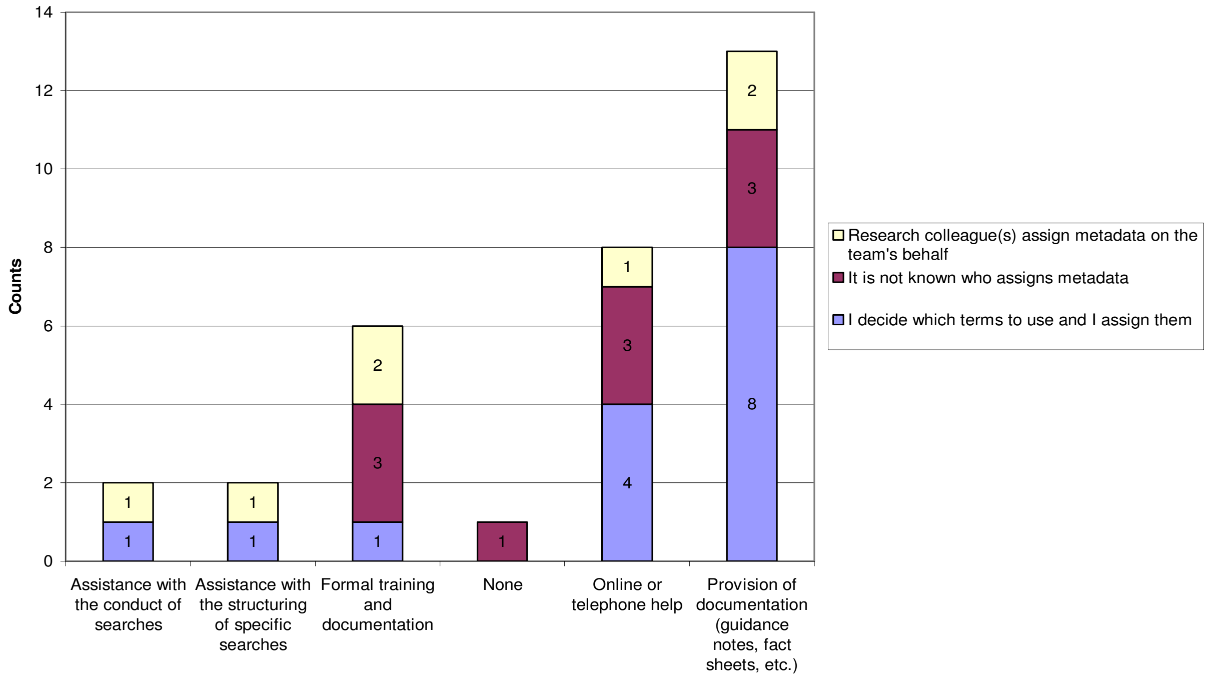
▪ Answers to question 11 were depicted against answers to question 25, showing a clear preference of researchers for assigning metadata themselves, especially in the case of researchers who would consult the documentation when accessing a repository.

Table 41: Metadata assignment compared to support level provided



Repository assistance	Metadata assignment	Count of Metadata assignment
Provision of documentation (guidance notes, fact sheets, etc.)		**13**
	I decide which terms to use and I assign them	8
	It is not known who assigns metadata	3
	Research colleague(s) assign metadata on the team's behalf	2
Online or telephone help		**8**
	I decide which terms to use and I assign them	4
	It is not known who assigns metadata	3
	Research colleague(s) assign metadata on the team's behalf	1
Formal training and documentation		**6**
	I decide which terms to use and I assign them	1
	It is not known who assigns metadata	3
	Research colleague(s) assign metadata on the team's behalf	2
Assistance with the conduct of searches		**2**
	I decide which terms to use and I assign them	1
	Research colleague(s) assign metadata on the team's behalf	1
Assistance with the structuring of specific searches		**2**
	I decide which terms to use and I assign them	1
	Research colleague(s) assign metadata on the team's behalf	1
None		**1**
	It is not known who assigns metadata	1
Grand Total		**32**

Figure 24: Metadata assigned compared to support level provided



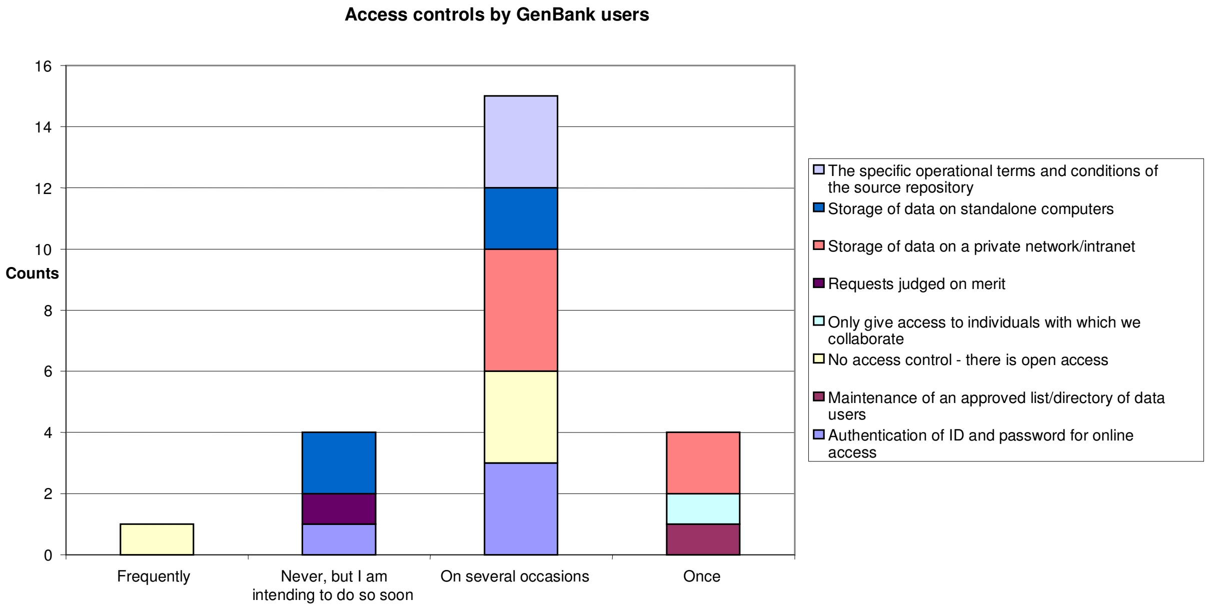
4.6 Usefulness of access controls by users of named source repositories

▪ Answers to question 18 were depicted against answers to question 8, showing preference for storage of data on a private intranet especially for users who submit to GenBank on an occasional basis.

Table 42: Usefulness of output repositories named by GenBank users

GenBank user	Access control measures	Count of access control measures
Frequently		**1**
	No access control - there is open access	1
On several occasions		**15**
	Storage of data on a private network/intranet	4
	Authentication of ID and password for online access	3
	No access control - there is open access	3
	The specific operational terms and conditions of the source repository	3
	Storage of data on standalone computers	2
Once		**4**
	Storage of data on a private network/intranet	2
	Maintenance of an approved list/directory of data users	1
	Only give access to individuals with which we collaborate	1
Never, but I am intending to do so soon		**4**
	Authentication of ID and password for online access	1
	Requests judged on merit	1
	Storage of data on standalone computers	2
Grand Total		**24**

Figure 25: Access controls by GenBank users



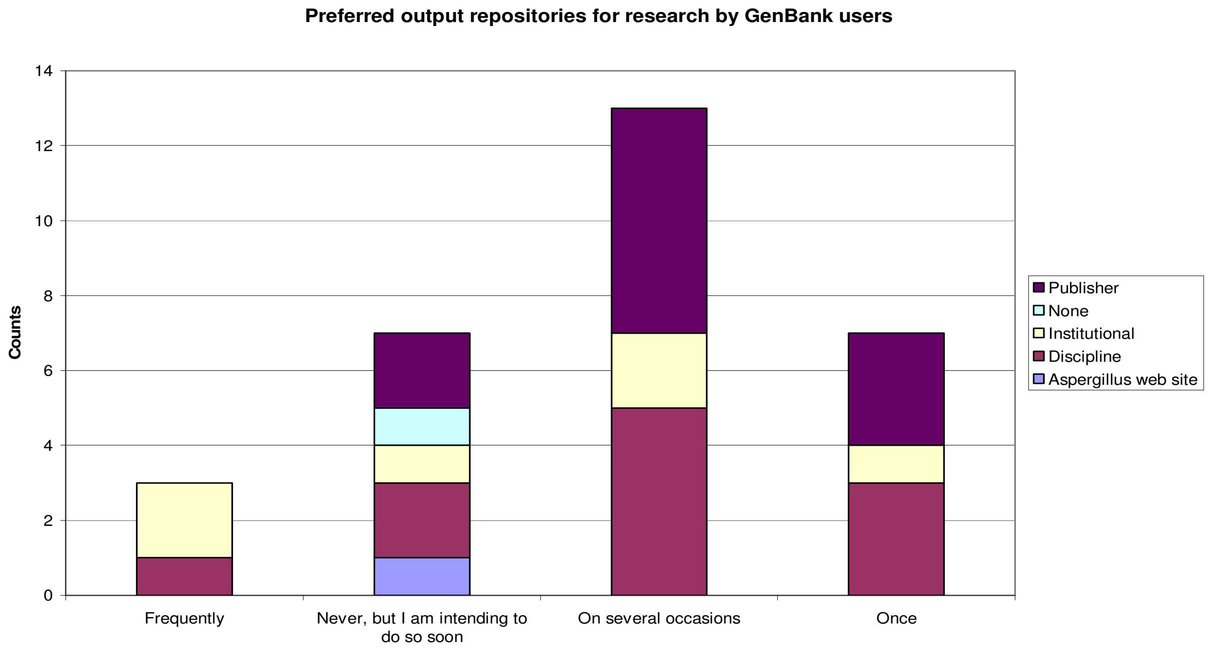
Output repositories for research by users of named repositories

- Answers to question 19 were depicted against answers to question 8, showing a preference for publisher and discipline repositories among GenBank users, as for biosciences researchers in general.

Table 43: Output repositories for research by GenBank users

GenBank user	Research output repository	Count of research output repository
Frequently		**3**
	Discipline	1
	Institutional	2
On several occasions		**13**
	Publisher	6
	Discipline	5
	Institutional	2
Once		**7**
	Publisher	3
	Discipline	3
	Institutional	1
Never, but I am intending to do so soon		**7**
	Publisher	2
	Discipline	2
	Institutional	1
	Aspergillus web site	1
	None	1
Grand Total		**30**

Figure 26: Preferred output repositories for research by GenBank users



4.7 Searching across different types of output repository, with enhancements

▪ Answers to question 19 were depicted against answers to question 23, showing a clear preference of simple searching across publisher and discipline repositories. No graph was produced for these figures.

Table 44: Level of searching depending on different types of research output repositories used

Research output repository	Level of searching	Count of level of searching
Publisher		**29**
	Simple - e.g. author, title, keyword, date	21
	Advanced, using a range of fields and identifiers	5
	Employing Boolean logic	3
Discipline		**24**
	Simple - e.g. author, title, keyword, date	18
	Advanced, using a range of fields and identifiers	3
	Employing Boolean logic	3
Institutional		**14**
	Simple - e.g. author, title, keyword, date	9
	Advanced, using a range of fields and identifiers	3
	Employing Boolean logic	1
	No preference	1
Aspergillus web site		**1**
	Simple - e.g. author, title, keyword, date	1
None		**2**
	No preference	2
Grand Total		**70**

Searching across different types of output repository, with enhancements

▪ Answers to question 19 were depicted against answers to question 23 and 23 a, giving a varied picture of possible enhancements suggested by the researchers. The numbers are possible too small to describe a certain trend.

Table 45: Free text suggestions for enhancements by research output repositories and their levels of searching

Research output repositories	Level of searching	Enhancements
Publisher	Advanced, using a range of fields and identifiers	Being able to search full text rather than only title or abstract (as on GoogleScholar)
Publisher	Simple - e.g. author, title, keyword, date	Extensive database *
Publisher	Simple - e.g. author, title, keyword, date	Alternative spellings (American/English) and abbreviations (e.g. beta and B)
Institutional	Simple - e.g. author, title, keyword, date	Extensive database *
Discipline	Simple - e.g. author, title, keyword, date	e.g. limit to one organism
Discipline	Advanced, using a range of fields and identifiers	Being able to search full text rather than only title or abstract (as on GoogleScholar)
Discipline	Simple - e.g. author, title, keyword, date	Extensive database *
Discipline	Simple - e.g. author, title, keyword, date	Alternative spellings (American/English) and abbreviations (e.g. beta and B)
Aspergillus web site	Simple - e.g. author, title, keyword, date	Alternative spellings (American/English) and abbreviations (e.g. beta and B)

* The free text answer was:
'It would be helpful if there was one database that would have EVERYTHING on. As it is, I need to check at least 5-6 different databases as each comes up with different results. Also, links to the actual papers are v. v. useful. I normally have to go through Google Scholar for that type of service though.'

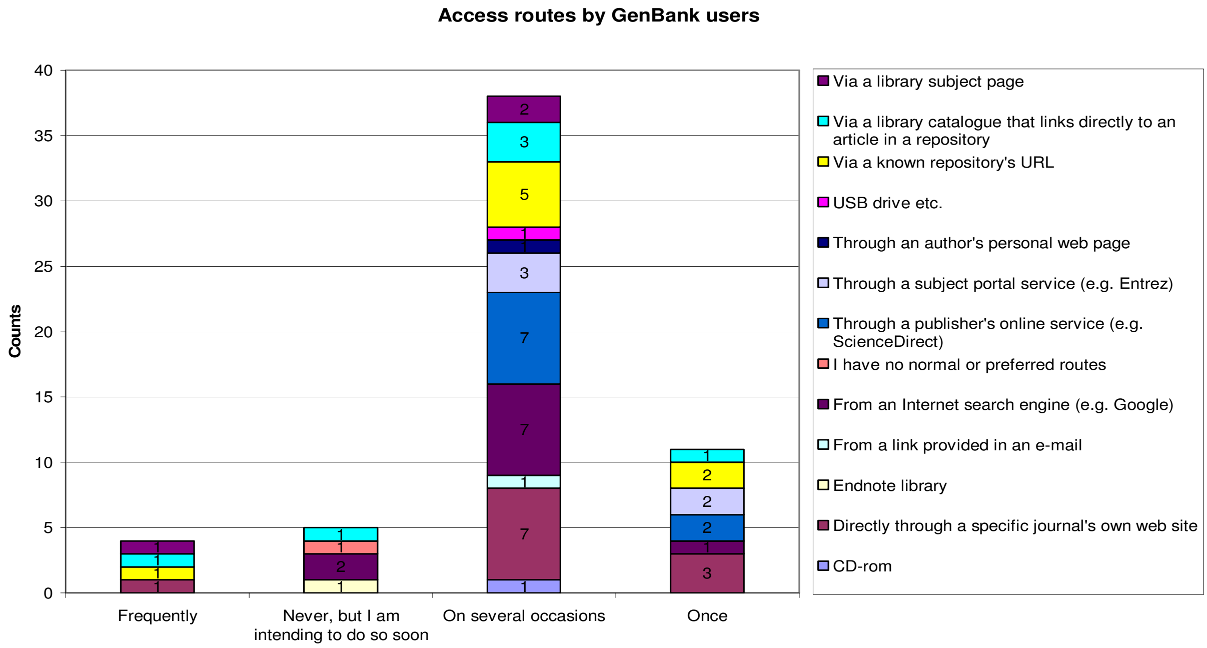
4.8 Preferred routes to output repositories by users of named source repositories

▪ Answers to question 22 were depicted against answers to question 8, showing a preference for access through journal web sites, publisher's online services and internet search engines among GenBank users.

Table 46: Access routes by GenBank users

GenBank user	Access route	Count of access route
Frequently		**4**
	Directly through a specific journal's own web site	1
	Via a known repository's URL	1
	Via a library catalogue that links directly to an article in a repository	1
	Via a library subject page	1
On several occasions		**38**
	Directly through a specific journal's own web site	7
	Through a publisher's online service (e.g. ScienceDirect)	7
	From an Internet search engine (e.g. Google)	7
	Via a known repository's URL	5
	Through a subject portal service (e.g. Entrez)	3
	Via a library catalogue that links directly to an article in a repository	3
	Via a library subject page	2
	Through an author's personal web page	1
	From a link provided in an e-mail	1
	CD-rom	1
	USB drive etc.	1
Once		**11**
	Directly through a specific journal's own web site	3
	Through a publisher's online service (e.g. ScienceDirect)	2
	From an Internet search engine (e.g. Google)	1
	Via a known repository's URL	2
	Through a subject portal service (e.g. Entrez)	2
	Via a library catalogue that links directly to an article in a repository	1
Never, but I am intending to do so soon		**5**
	From an Internet search engine (e.g. Google)	2
	Endnote library	1
	I have no normal or preferred routes	1
	Via a library catalogue that links directly to an article in a repository	1
Grand Total		**58**

Figure 27: Access routes by GenBank users



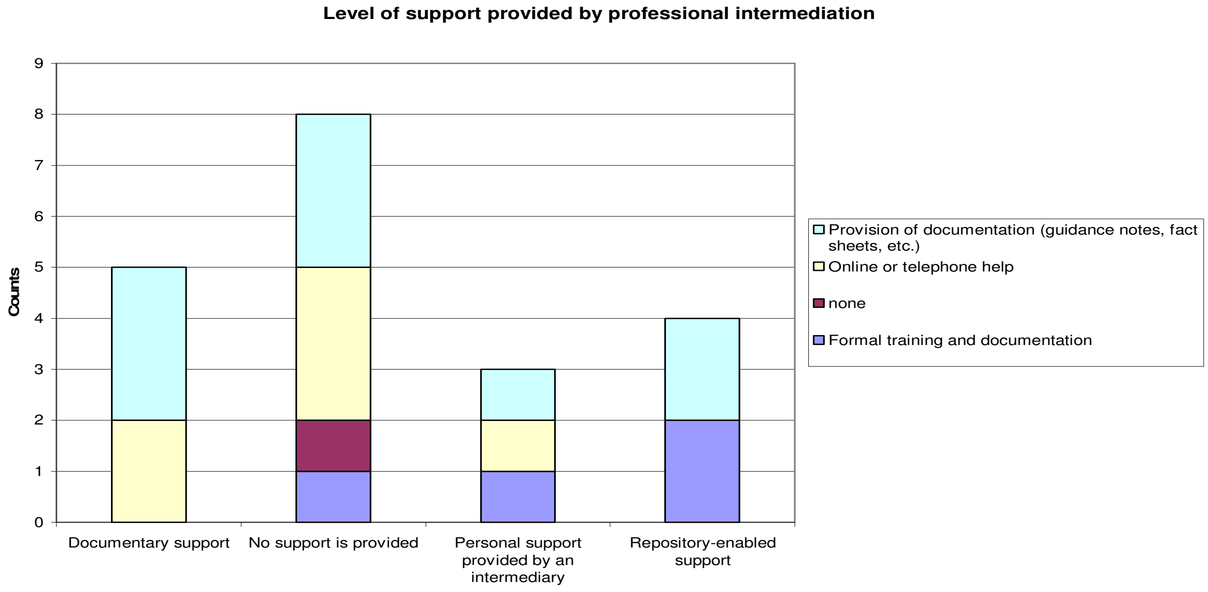
4.9 The level of support/guidance provided matched against professional intermediation

▪ Answers to question 24 were depicted against answers to question 25, showing a preference for self-sufficiency among GenBank users, which is also the case for biosciences researchers in general.

Table 47: Level of support provided matched against professional intermediation

Level of support	Professional intermediation	Count of Professional intermediation
Documentary support		**5**
	Online or telephone help	2
	Provision of documentation (guidance notes, fact sheets, etc.)	3
Personal support provided by an intermediary		**3**
	Online or telephone help	1
	Provision of documentation (guidance notes, fact sheets, etc.)	1
	Formal training and documentation	1
Repository-enabled support		**4**
	Provision of documentation (guidance notes, fact sheets, etc.)	2
	Formal training and documentation	2
No support is provided		**8**
	Online or telephone help	3
	Provision of documentation (guidance notes, fact sheets, etc.)	3
	Formal training and documentation	1
	none	1
Grand Total		**20**

Figure 28: Level of support provided by professional intermediation



5. Additional data from the interviews

Most observations from the interviews are listed separately for 'Responders' (interviewees who also responded to the StORe questionnaire) and 'Non-Responders' (interviewees who did not respond to the StORe questionnaire), to see if there is any difference between both groups. In addition, the sum of both is given.

5.1 Source data

▪ Question 13: How would you normally access the research data of other researchers?

▪ Unlike respondents to the StORe questionnaire, where the prevalent access route to research data seems to be through online access to source repositories, two thirds of the interviewees seem to favour the exchange of portable media, followed by networked fileservers at the own institution, and to a lesser extent on-line access. As for the respondents to the StORe questionnaire, access to other institutions' networked fileservers and e-mail attachments do not seem to be of great importance when accessing the research data of other researchers.

Table 48: Research data access routes by type of interviewee

Category	Portable media	Own networked fileservers	Online access	Other networked fileservers	E-mail attachments
Responders	4	2	2		
Non-Responders	4	4	2	1*	1
Total	8	6	4	1*	1

* In this case a shared virtual drive on the internet

The access routes were:
Through online access to source repositories
I do not normally access others' research data
Through the exchange of data held on portable media (disks, CD-ROMs, USB drives, etc.), including DTL
By access to networked fileservers at my own institution
By access to networked fileservers at other institutions
Other: Please specify (all respondents wrote 'e-mail attachments')

5.2 Source repositories

▪ Question 7: To which source repositories do you submit your data?

▪ Of all the interviewees who answered this question (7 out of 12), most researchers submit data to an internet-based repository, with GenBank being possibly of less importance here than it is for the questionnaire constituency. However, the numbers are too small to make more than an assumption.

Table 49: Source repositories by type of interviewee

Name	Internet	GenBank	EBI
Responders	2	1	
Non-Responders	2		2
Total	4	1	2

▪ **Question 8: How often have you submitted data to any of these source repositories?**

▪ Responders and Non-Responders are added together here because of the overall low numbers. Again, source repositories on the internet seem to be favoured by all interviewees, followed by EBI repositories.

Table 50: Submission frequency by type of interviewee

Submission frequency	Internet	EBI	GenBank
Frequently	3	2	1
Occasionally	1		

5.3 Metadata

▪ **Question 9: By selecting from the following options, please would you indicate what types of metadata you consider it important to assign to your data.**

▪ Following the trend for both questionnaire responders and interviewees, a thorough description of experimental conditions is crucial for Non-Responders especially. This is followed by author name, project title and date, with subject keywords featuring low in the rank of importance.

A possible explanation for this fact may be that Non-Responders work in molecular biology and bioinformatics, where submission may include a higher number of e.g. gene/protein sequences. These would require different metadata from scientific publications and would not necessarily require subject keywords. Instead different metadata, such as database accession numbers would be required and they seem to be especially important for the Non-Responders.

Table 51: Metadata type by type of interviewee

Metadata type	Responders	Non-Responders	*Total*
Description of experimental conditions	4	6	10
Author or data creator name	4	5	9
Project title	3	5	8
Date (e.g. of data creation)	3	3	6
Database accession numbers		5	5
Links to websites, networked files		5	5
Subject keywords	1	3	4
Project description	1	3	4
Project reference number or identifier	1	2	3
MIAME	1	2	3
Title of data set	1		1
Funding source	1		1

▪ **Question 10: At what stage are metadata assigned to your data?**

▪ Metadata seems to be assigned during file saving or when submitting, for both Responders and Non-Responders. This is consistent with the answers to the questionnaire.

Table 52: Metadata stage by type of interviewee

Category	During file saving	When submitting	After submission
Responders	3	2	1
Non-Responders	1	2	1
Total	4	4	1

▪ **Question 11: Who assigns metadata to your research data?**

▪ It seems that more than half of the interviewees assign metadata to their research data themselves, which is similar to the results from the questionnaire, with all Non-Responders stating that they decide and assign metadata themselves.

However, half of all interviewees leave the task of assigning metadata to repository administrators, which might be explained by the fact that EBI and internet-based repositories use automated metadata assigning systems to ensure a consistent standard.

Table 53: Metadata assignment by type of interviewee

Category	I decide and assign	Repository administrators
Responders	1	3
Non-Responders	6	3
Total	7	6

5.4 Data access and sharing

▪ **Question 14: What measures do you use to make your research data available?**

▪ It is interesting that the use of a publicised URL seems to be of much higher importance for interviewees that for respondents to the questionnaire. In contrast, data availability through a publisher is of high importance for questionnaire respondents, but does not seem to be of any importance at all for interviewees (it was not mentioned). One explanation could be that interviewees were more possibly more focused on sharing their data through public domain repositories rather than through scientific publications.

Table 54: Data availability by type of interviewee

Data exchange through …	Responders	Non-Responders	*Total*
Publicised URL	2	3	5
E-mail	3	1	4
Portable media		3	3
Source repositories		3	3
Printouts and post	1		1
Verbal data exchange only	1		1
I don't make my data available		2	2

Table 55: Data availability - Additional statements from the interviews

Public Access Repositories:
I give people access to our standards lab protocols. Experimental protocols are freely available on the web and we tend to use them. We usually find them with a Google search.
We send our data to the EBI public repositories and once it is stored in the public domain we do not keep a copy of it (the file size is too big)
We run several public repositories. Some of these are quite popular, for example the Chicken EST one receives about 7 hits a day from various labs.
We publish in BMC journals and they require existing links from our publication to our source data
Other:
I have colleagues all over the world and they have different preferred contact routes. One example is a colleague in a faraway region, with whom I keep in contact through letters only. Not everybody in the whole world seems to be all that familiar with computers, and that is me included.

▪ Question 15: What factors would encourage you to share your research data?

▪ Many researchers seemed to be in favour of sharing teaching materials. Individual researchers were also stating collaborative work processes, protection of ownership and supporting free access in general as factors to encourage them to share their data.

Table 56: Other factors encouraging the sharing research of research data

Teaching:
Anything that is used as teaching material can be freely shared and I have already accessed some teaching materials from the web.
The data we generate is visually striking – it makes a good picture to give to students as teaching aid.
Material from institutional talks can be stored as teaching material.
Institutional repositories could contain the microarray handbooks from commercial companies.

Collaboration:
My students put their results in the public domain, because they are interested in the feed-back, for example to find the errors in a program.
Open Access Initiative:
I support free access to information in general. We publish in BMC.
Ownership protection:
Publishing everything gives you some sort of protection of ownership- hiding does not. If you say ‘It’s mine’ in front of 25 000 people, they are much less inclined to lift your stuff and they will also spot people who do.

▪ Question 16: What factors would discourage you from sharing your research data?

▪ Two interviewees stated that they did not share any of their data. For one interviewee, this was because he considered his findings to belong to his employer only (the university), so he would only share data within the university. Another researcher stated that he feared increased competition for funding and ‘stealing of results’ by other researchers. This was also given as reason for a certain reluctance in the case of a third researcher who would share teaching materials only and gives access to his research source data based on individual merits.

▪ Question 17: Normally, what kind of formal restrictions do you apply to your research data?

▪ Most researchers in the biosciences give access to their research data on an individual basis. This trend was consistent throughout questionnaire and interviews. In addition, some Non-Responders from the interviews stated more specific restriction types such as the GNU license terms for free software.

Table 57: Formal data restrictions by type of interviewee

Restriction Type	Responders	Non-Responders	*Total*
Judged individually, on merits*	4	3	7
Restricted to research partners	1		1
Not prior publication		2	2
GNU license		2	2
Data unlikely to be understood		1	1
No formal restrictions	1	2	3

The restriction types were:
Individual enquiries/requests for access are judged on their merits
* including: not to commercial companies, public sector only
Restricted to immediate research team/programme members
Downloads covered by GNU license terms
Own data is extremely complex. Not many people would understand it, which serves in itself as an access control.

▪ **Question 18: What measures do you normally use to control access to your data by others?**

▪ Most biosciences researchers seem to keep their data non-network accessible. Some use storage of data on a private network or intranet as measure of access control.

Table 58: Access control measures by type of interviewee

Access control	Responders	Non-Responders	*Total*
Storage on standalone device	5	6	11
Private network/ intranet	3	4	7
Online authentication	1		1
Source repository level	1	1	1
Data encryption	1		1
No access control		1	13

The access control measures were:
Storage of data on standalone computers (or experimental devices)
Storage of data on a private network/intranet
Authentication of ID and password for online access
The specific operational terms and conditions of the source repository
I encrypt the data and people have to contact me to get the key.
No access control – there is open access

5.5 Output repositories

▪ **Question 22: What are your normal or preferred routes to the contents of output repositories?**

▪ Interviewees stated that their main route to an output repository is via a known repository URL. An internet search engine also featured highly in the access routes of Responders. In contrast, non-responders mainly access output repositories through a subject portal with journal websites also featuring highly.

Table 59: Routes to output repositories by type of interviewee

Access route	Responders	Non-responders	*Total*
Repository URL	4	5	9
Subject portal	3	6	9
Journal web site	2	5	7
Internet search engine	4	2	6
Library catalogue	2	2	4
Publisher's service		2	2
E-mail link		1	1

The access routes were:
Directly through a specific journal's own web site
Via a known repository's URL
From an Internet search engine (e.g. Google)
Through a subject portal service (e.g. Entrez)
Through a publisher's online service (e.g. Science Direct)
Via a library catalogue that links directly to an article in a repository
From a link provided in an e-mail, CD-ROM, USB drive etc.

▪ Additional question: Do you read research publications on-line or do you print them out?

This came up as an additional issue during the interviews. 3 researchers stated that they prefer to read research publications in paper form, with one saying that he tends to read those on the bus, with the existing electronic link to the appendices of no use to him in this case. 2 researchers stated explicitly that they read everything on-line only.

▪ Additional question: Do you 'scan' or browse key journals or web sites on a regular basis, in order to get an overview about recent developments in biosciences research?

▪ 7 out of 12 interviewees gave information browsing as part of their general information search strategy during their research work.

One researcher stated that within his group they collaboratively perform about 50 keyword searches per week, and also scan the table of contents of important journals. The research group have split up the task of searching, with all of them knowing what the others might be interested in.

Table 60: Browsing habits by type of interviewee

Browsing	**Journals**	**Web-based**	**Discipline repository**
Responders	3	1	1
Non-Responders	2		
Total	*5*	*1*	*1*

Examples stated:
Journals: discipline-specific journals, e.g. Nature, Science
Web-based: Google, Internet (non-specific)
Discipline repository: NCBI web site

▪ Question 23: What level of searching do you normally find sufficient when using an output repository?

▪ Throughout the whole survey, researchers stated that their normally employed level of searching was a simple search. This is similar for both respondents to the questionnaire and interviewees. Other search modes do not seem to be very widespread.

Table 61: Levels of searching by type of interviewee

Category	**Simple**	**Advanced**	**Boolean**	**Thesaurus**	**No preference**
Responders	4	1	1	0	0
Non-Responders	5	1	0	0	0
Total	9	2	1	0	0

The levels were:
Simple – e.g. author, title, keyword, date
Advanced, using a range of fields and identifiers
Employing Boolean logic
Using a subject thesaurus or subject headings
No preference

5.6 Support

▪ Question 24: Do you receive support and/or guidance in your use of output repositories?

▪ Most interviewees stated that they do not receive any support in their access of output repositories, which is consistent with the trend for all biosciences researchers.

Table 62: Output repositories support by type of interviewee

Support	Documentary	Personal	No support
Responders	1	1	4
Non-responders	1		5
Total	*2*	*1*	*9*

The support options were:
Documentary support
Personal support provided by an intermediary
Repository-enabled support
No support is provided

5.7 Additional topic: Cross-disciplinary information search

▪ Additional question: Do you regularly access source data or scientific publications from other disciplines?

▪ Two thirds of the interviewees seem to access information from different disciplines on a regular basis.

Table 63: Frequency of cross-disciplinary access, by type of interviewee

Category	Yes	No
Responders	4	1
Non-Responders	4	3
Total	*8*	*4*

▪ Additional question 30 b): What are the disciplines or the publications from other disciplines that you would be interested in?

▪ It should be noted that medicine and pharmacy were not regarded by the researchers as different disciplines. Most interviewees seem to access information from disciplines related to the life sciences. However, mathematics (as source of algorithms useful for bioinformatics) and geology (as determining soil structure which in turn determines plant ecology) were also stated.

Table 64: Target disciplines for cross-disciplinary search

- Veterinary sciences
- Physics
- Toxicology
- Structural biology
- Trade magazines
- Molecular chemistry
- Geology
- Governmental news alerts
- Zoology
- Ornithology
- Mathematics
- Philogeography
- Computer sciences

▪ Additional question: What would be your reasons for you not accessing information from other disciplines electronically?

▪ Some researchers seem to be in favour of personal communication rather than digital access, with for example the clarification of different terminology being stated as a reason. It is interesting that on researcher would not want to access source data (spectroscopic data) in other disciplines because of the large file size.

Table 65: Reasons for not accessing other disciplines electronically

I tend to contact other researchers by e-mail rather than accessing their data on-line.
I contact the researcher because then we can have an interesting conversation over the phone and work out together what the different (discipline-specific) terms mean.
I am interested in spectroscopy data of other disciplines but would not access or download that kind of data, because the file size is far too big.
I don’t think I would need to access the source data from other disciplines.

▪ **Additional question: What is your experience with accessing source data from other disciplines? Do you find it easy to get through to the data you want?**

▪ The researchers' experience with cross-disciplinary searches seems to be quite varied, with the different terminology being stated as a potential problem for understanding other disciplines.

Table 66: Experiences with cross-disciplinary access

Positive:
Yes, fairly easy (3 entries)
It is only difficult if the web site is difficult to navigate.
I have long experience with the terminology, so it is fairly easy.
Negative:
No, not easy (2 entries)
The information can be quite difficult to understand, for example maths is barely comprehensible to me because of the different terminology they use.
Beilstein is a nightmare – did not figure out how to use it beyond a very basic level.

▪ **Additional question: What would be a reason for you not finding the information you want?**

▪ Variable reasons for unsuccessful information searches were given here, with IT problems, insufficient metadata capture and inherently unreliable information all being stated.

Table 67: Problems with information retrieval, cross-disciplinary and in general

Unreliable information:
Sometimes the information in some journals is wrong and that can be detrimental.
Information referring to some undergraduate course is usually rubbish.
Teaching notes often contain errors.
Metadata capture:
The experimental methods are not always very well standardised. This is why an extensive description of experimental conditions and properties would be needed, but it is not always there.
Getting all the experimental conditions captured in enough detail to be meaningful can be a problem. The information then has gaps and some cannot be used. However, I am currently setting up systems for the Faculty of Life Sciences which will help to standardise the data.

Other:
We have problems accessing the latest publications because of the delay caused by the whole publishing process. We need the very latest research information, especially at the beginning of a project.
Sometimes the publication says ‘It’s all on our web site’ but then it’s not there.
I would not know what is in a commercial database. If you are a bioinformatician and you have the choice between a commercial and a free site, you will always go for the free option.
I sometimes need older geological maps, e.g. from the 1880s. Most of them are not digitised and then I will have to order a paper copy.

▪ Additional question: What would you think about a common standard for access to repositories in different disciplines?

▪ The researchers’ views on a common cross-disciplinary access standard seemed to be varied, with many researchers being doubtful about practicalities.

Table 68: Perceived benefit of common access standard

Positive:
Yes, that should absolutely be standardised
Good idea
Neutral:
I would change the formats for articles, away from the traditional layout, and much shorter. That could then be the standard for different disciplines.
I think they are all fairly similar already
No opinion – too little experience
Negative:
Google works fine for me
Probably not feasible because of the different terminology
That would be rather difficult. The data would be consistent with a particular standard of publication, which might be a different one.
The data itself might be of a very different nature, so would e.g. astronomy data be very different from my data, I guess.

5.8 Additional topic: Help functions

▪ Additional question: Would you find help facilities useful when you are accessing a repository? (e.g. on-line help, two-directional conversation, user ratings, FAQ, terminology help)

▪ This question aimed to explore the researchers' opinions on advanced help functions. Again, the views on the usefulness of help facilities were somewhat diverse, but many researchers could actually see a potential benefit of technically improved help functions.

Table 69: Perceived benefit of help facilities

Positive:
Used all these in the past and found them useful.
I would certainly give it a try. I can imagine that a live on-line help could be useful.
Yes, helpful. Sometimes there is a specific style, for example for bioinformatics tutorials, and an explanation of this style can be very helpful.
User ratings and FAQ lists can be quite helpful.
Neutral:
The information I access is usually fairly self-explanatory. One example is algorithms from public domains.
For a program the full documentation should be with it. However, it is usually the last thing to use.
Negative:
Librarians might have problems to understand the exact details of what is important for me. They simply don't have the technical knowledge. So I don't use their help.
No, not helpful. Self-explanatory web sites are best.
I have no time to read the on-line help.
Wouldn't know because I don't like using help functions.

▪ **Additional question: What is your experience with using help functions?**

▪ The researchers' experience with help functions seemed to be rather varied.

Table 70: Personal experience with help functions

General opinion:
The help often doesn't address my specific problem.
Not always easy to understand.
Don't really use them.
FAQs:
I read the FAQs first or the 'How to' section.
I read user ratings and FAQs on web sites and I find them helpful
FAQ lists contain too many items. It takes too long to scroll through everything. A search function within FAQs would be helpful.
I don't use FAQs, because what I need is never in there.
Keywords and functionality:
I'm moderately comfortable with accessing web sites, but I do use the on-line help, and I read the help menu and the specific advice.
The keywords are hard to guess.
You have to express your question in exactly the way they want, but you don't always know how to phrase the question or what the preferred terminology is.
You have to think of the exact terms and the input has to be exactly what they want.
I'm quite experienced with MeSH headings, tags, data lists, but I don't use them because you have to be experienced in order to use them correctly.
Terminology servers are interesting but not all that useful. Experience with how to research for information is the most useful.
Other:
Web-based programs usually contain some basic instructions and these are helpful. They are also quite user-friendly.
Entrez is good because it is self-explanatory and there is no need for on-line help.
I tend to look for a quick and to-the-point answer but finding an answer takes too much time.
It can be frustrating when there is only an e-mail address given and then you have to send them an e- mail.

▪ Additional question: How do you find information if you are not using a help function?

▪ Again, personal communication seems to feature highly in the researchers' preference. Some researchers would use search engines and specific web sites for help and some would read the manual or tutorial. It is interesting that one researcher stated that he would prefer to 'try things out', since he would consider that the most informative procedure.

Table 71: Alternative routes of information retrieval support

Personal communication:
If I'm really desperate for a piece of information I might go and ask other people for help.
I contact the researcher.
It is much faster to ask a colleague a specific question and then I like telephone conversations with experts anyway.
If I am completely 'thrown' by an article, I use shortcuts to experimenters.
Web sites:
I use Google to understand what's in an article.
Things I don't understand I check up in KEGG.
Tutorials:
I read the manual which is usually put together by experts in the field.
User tutorials and step-by-step guides, but they tend to be quite lengthy.
Other:
I tend to try things out. I find it more informative to try out how everything works.

6. Scenarios

These scenarios provide typical user situations from biological sciences research work. Since they are meant to be typical examples only, they do not provide extensive covering of all eventualities. However, many extensions occurring in a real life situation were incorporated, e.g. the use of different access routes.

Scenario 1 had to be written as a generic version because it is cross-referenced by other scenarios. It is somewhat ‘padded out’ in a real life situation in Scenario 2. Scenario 1 describes the central information research process within sciences. This process is crucial for research and is performed throughout all stages of research work.

6.1 Scenario 1

Researching scientific information – generic version

Narrative:

A biosciences researcher wants to undertake a comprehensive overview about a certain topic in a scientific field. This knowledge will then form the basis for subsequent research steps.

Action steps:

1. The researcher performs an information search.
2. The researcher finds the desired information.
3. With his/her background knowledge and scientific training, the researcher evaluates the information.
4. Based on the results, the researcher plans subsequent research steps.

Alternative pathways:

1.1 within his/her own resources
1.2 through a discussion with his/her contacts
1.3 in the university library: OPAC, specific journals, subscribed databases
1.4 through a scientific search engine, such as Scirus
1.5 through an unspecific search engine, such as Google

Extensions:

1.a at any point: the researcher does not find the desired information.
1.a 1. The researcher abandons this particular information search pathway.
1.a 2. The researcher asks an expert for help.

6.2 Scenario 2

Researching scientific information – example

Narrative:

An infectious disease microbiologist receives a phone call from the local authority. There has been an outbreak of tuberculosis in the local Somali community, with the causative organism belonging to the Somali substrain. The local authority now seeks expert advice on how to ensure the health and safety of their social workers working in this community. The microbiologist checks a range of relevant information in order to produce comprehensive advice.

Action steps:

1. S/he remembers a talk about human immune response towards different tuberculosis substrains that was given at a recent conference and skims through the material from the conference, of which some is in paper form and some on his/her laptop.
2. S/he knows that there is some relevant information on the WHO and the EuroTB web sites. S/he accesses the sites and reads the information, and bookmarks the interesting pages.
3. S/he searches for relevant information in her university library databases. This involves calling up the university library web site, logging on to Athens and then accessing Ovid with selecting a range of publication databases, such as Medline and Embase. S/he performs a search across these databases, using 'MYCOBACTERIUM/ and tuberculosis and Somalia' as search term.
4. From the list of results, s/he picks the ones with the most promising titles, then abstracts, then full text, and s/he prints out the most interesting ones in order to read them on the way home.
5. The next day, s/he performs a similar search in Google. S/he checks the content descriptors and the URLs from the results list and decides to follow a link through to an article in the Somaliland Times. This contains useful information about simple protection measures when chewing Qad, and s/he bookmarks the article.
6. S/he follows two of the three links in the Google Scholar results list, reading the content descriptors first and then the full text article. In order to find related articles she follows some of the links at the bottom of the page, to other articles citing this one. S/he prints the most interesting publications out and reads them carefully, highlighting key facts and scribbling thoughts on the edge of the pages.
7. S/he gathers all the stored information together.
8. S/he evaluates all information and produces a comprehensive facts summary.
9. Based on this information s/he produces a concise piece of written advice, targeted towards the specific situation of the local authority, and sends it by post to his/her contact at the authority.

Extensions:

1.a Unfortunately, the talk dealt with the different substrains on a too broad basis. She e-mails the researcher who gave the talk to enquire about his exact findings for the Somali substrain. His/her answer is compiled together with all other information.

2.a S/he is not exactly sure where to find the information on the web site and performs a search with using the site search engine and the keyword 'Somali'.

3.a She has second thoughts about her choice of databases and e-mails the health sciences subject librarian, who tells him/her to include BNI and CINAHL in the search.

4.a Some of the sources contradict each other. S/he discusses the details over the phone with a respiratory disease physician s/he knows from a previous collaboration and then decides on one viewpoint only.

5.a The link does not work. S/he abandons it and goes back to the list of Google results.

6.a The Google search did not produce any relevant hits. S/he performs the same search in Mamma (16)[26].This brings up an internal SSI publication about a review of occupational TB infections with the Somali substrain in social workers in Copenhagen, and s/he bookmarks the page.
9.a S/he is not familiar with the specific situation of the local authority and checks their web site for further information on practices and procedures.

6.3 Scenario 3

Investigation of an unknown nucleotide sequence – homology search

Narrative:

A molecular biologist is part of a group investigating the expression of many known and unknown genes by using microarrays. In one of the experiments s/he has noticed an interesting human gene that s/he decides to examine further. A common method is to perform a homology search to relate this gene sequence to other known sequences.

Action steps:

1. S/he copies the gene sequence from the output of the sequencing experiment, transforms it from plaint text to FastA format and pastes it into the web-based utility BLAST (17)[27], adjusting the BLAST settings in an appropriate way for his/her query.
2. BLAST compares the sequence with all GenBank CDS entries, translates the gene sequence into a protein sequence and finds similar known proteins from a range of databases, such as PDB and SwissProt.
3. The output is a list of alignments of the query sequence and one subject sequence each, together with relevant information and links.
4. The researcher follows the links which lead to other divisions of the NCBI database system, in order to find out more about the subject protein.
5. With his/her extensive background knowledge, the researcher decides on one appropriate subject protein sequence and copies the alignment plus further information into his/her experiment database. This information will provide the basis for further experiments.

Extensions:

1.a Web failure of any sort during setup. User tries again, preferably not during US working hours.
1.b S/he is not sure which substitution matrix to pick and asks his/her colleague who runs the BLAST introductory course for postgraduate students. After a discussion with the colleague, s/he decides to use BLOSUM62 because of its wide application range.
2.a There are no similar proteins at all, or a true homology is very unlikely. Continue at 2.b
2.b The molecular biologist enlists the help of a bioinformatician to assign a function to his/her gene. S/he sends the query and the sequence by e-mail to an expert - see Scenario 4.
3.a Alternatively or additionally: The researcher performs a search across a range of output repositories - see Scenario 1.
4.a The researcher has not enough information to decide on one protein sequence.
4.a 1. S/he performs a search for further information - see Scenario 1.
4.a 2. S/he discusses the subject in a group meeting. His/her colleagues contribute useful ideas and also information in which repositories to look further - see Scenario 1.

[26] see bibliography list, page 86

[27] see bibliography list, page 86

6.4 Scenario 4

This scenario has been shortened to provide a clearer layout and usefulness for a wider audience. For the original version, see also under appendices: A 3 Scenario 4 – detailed version.

Predicting the function of a novel gene – search for motifs

Narrative:

A bioinformatician receives a request from a molecular biology group, by e-mail. They have found a new genetic sequence of potential interest.
However, when they used it for a BLAST search, the only result was a list of proteins of uncertain homology (see Scenario 3). A complementary method for predicting the function of an unknown gene sequence is to compare the translated sequence to patterns of conserved features or motifs in known protein families. This is a standard task in bioinformatics.

Action steps:

1. S/he runs a search of the query sequence against InterPro, the integrated protein family database. However, knowing that InterPro is not fully synchronised with its source databases, s/he runs separate searches against these too, using the search tools available on each of the database home pages. For each search, s/he compares the results to ensure consistent answers, and cross-references them back to the BLAST result.
2. The bioinformatician uploads the translated sequence into CINEMA, an in-house manual sequence alignment editor (18)[28]. S/he creates a multiple alignment of the query sequence with the InterPro/BLAST-consistent sequences, highlighting any common motifs.
3. S/he also carefully checks the Swiss-Prot Feature Table (19)[29], which contains sequence annotations (metadata and biological information), often derived from sequence analysis tools - this might highlight, for example, any hydrophobic domains.
4. S/he compares the highlighted motifs with the annotated regions in the Feature Table and, with his/her extensive background knowledge, s/he decides which of these features might be structurally or functionally significant - s/he is likely to check the literature cited in the Swiss-Prot entry to see if any of the regions annotated in the Feature Table are supported by an experimental finding - see Scenario 1.
5. The bioinformatician produces a motif recommendation, together with relevant biological information and sequence annotations, and sends this back to the molecular biology lab.
6. The lab verifies the results. Both 6 a and 6 b apply.

Extensions:

1.a The query does not give any significant results, i.e. there are no discernible common motifs within query and BLAST sequences. This could be due to errors in the sequence. S/he aborts the search and asks the molecular biology group to re-run their sequencing experiments.

2.a S/he needs more information about the motifs and performs an information search -see Scenario 1.

3.a There is only minimum annotation. S/he decides that s/he needs more information and performs an information search - see Scenario 1.

4.a There is minimal or no information on the motifs. S/he decides to create a discriminator (e.g. a fingerprint) to perform more specific searches of Swiss-Prot/TrEMBL. S/he opts to use in-house software to create the fingerprint.
S/he excises the motifs, runs separate motif searches of Swiss-Prot/TrEMBL, and feeds the hits which identify all the sequences in the correct order back into the database search. The result is

[28] see bibliography list, page 86

[29] see bibliography list, page 86

then annotated with whatever minimal information could be gleaned from Swiss-Prot and passed back on to the molecular biology group.

4.b There is minimal or no information on the motifs, but one of the matched sequences has had its 3D structure determined. S/he decides to use the Swiss-Model server to produce a homology model of the query sequence using the known structure of the matched homologue as a template.

6.a The molecular biologists save the motif into a FastA file and use it to run a new BLAST search. As this search has now been narrowed down considerably, the resulting homologues are far more significant.

6.b The molecular biologists run a new series of gene sequencing experiments to confirm the predicted structure.

6.5 Scenario 5

Depositing a genetic sequence in a public repository

Narrative:

A molecular biology research group has identified a human genetic sequence which corresponds to a new subfamily of cytochromes. They decide to deposit the sequence into a repository in the public domain, prior to publishing an article in a molecular biology journal. One member of the research group takes the responsibility for performing the submission to GenBank.

Action steps:

1. The researcher calls up the web-based submission tool Sequin (20)[30].
2. In the submission form, s/he chooses GenBank as repository of choice.
3. S/he fills in the authors, contact, and affiliation forms, and provides further information, such as the name of the organism, coding sequence location, and sequencing date. Some of this annotation is mandatory and some optional.
4. S/he uploads the single sequence in FastA format. S/he then uses the automated annotation features of Sequin, for example to identify and mark the open reading frames.
5. S/he then lets the program generate a view of the submission as a GenBank entry and makes appropriate corrections and changes where necessary.
6. S/he submits the whole piece of information, which is then loaded into the temporary hold division of GenBank. This repository is updated nightly, after which the sequence is uploaded into the subject-specific division of their database system.
7. This information is now ready to be linked to a scientific publication.

Extensions:

1. 2. 3. 4. 5. 6 a) Web failure of any sort. User tries again, preferably not during US working hours.

3.a S/he needs more information in order to perform the annotation. Since his/her laboratory has not installed an electronic lab book system yet, s/he organises a group meeting. In this meeting the group compiles more details about the experimental conditions and further information. Most of this data exists as handwritten experimental protocols in individual researchers' lab books only.

3.b Fully annotating a new gene requires a considerable amount of specialist knowledge. S/he decides to invite the molecular biology research community to give their expert comments and suggestions on his/her gene annotation and sets up a tool with a Wikipedia-type front end on his/her research group's web site. Following on from the resulting community annotation, s/he submits extensive information together with his/her genetic sequence.

7.a The link does not work. S/he contacts GenBank by e-mail. Apparently, his/her submission has not been registered. S/he re-submits the genetic sequence with all its annotation.

[30] see bibliography list, page 86

Bibliography

1. JISC Digital Repositories Programme http://www.jisc.ac.uk/index.cfm?name=programme_digital_repositories

2. JISC StORe web site http://www.jisc.ac.uk/index.cfm?name=project_store

3. StORe wiki http://jiscstore.jot.com/WikiHome

4. John Rylands University Library StORe web site http://www.library.manchester.ac.uk/projects/store/

5. Chapter 7, in: Basic Marketing Research, A. Burns and R. Bush, International Edition, Pearson Prentice Hall, 2005

6. Chapter 9, in: Human-Computer Interaction, A. Dix et al., 3rd edition, Pearson Prentice Hall, 2004

7. Personal experience in market research data collection through interviews and questionnaires, 2005

8. Chapter 7, in: Writing Effective Use Cases, The Crystal Collection for Software Professionals, Alistair Cockburn, Addison-Wesley, 2001

9. UniProt at http://www.ebi.uniprot.org

10. Leinonen, R. et al. (2004) UniProt archive. Bioinformatics, v 20, I 7, pp 3236 – 3237

11. Apweiler, R. et al. (2004) Protein Sequence Databases, Current Opinion in Chemical Biology, v8, I 1, pp 76-80

12. Definition in http://en.wikipedia.org/wiki/Genbank

13. GenBank at http://www.ncbi.nlm.nih.gov/Genbank

14. Chapter 10 (GenBank), in: Beginning Perl for Bioinformatics, J. Tisdall, O'Reilly, 2001

15. PubMedCentral at http://www.pubmedcentral.nih.gov

16. Mamma meta search engine at http://www.mamma.com/

17. BLAST at http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

18. CINEMA at http://aig.cs.man.ac.uk/research/utopia/cinema/cinema.php

19. Swiss-Prot access at http://www.ebi.ac.uk/swissprot/access.html

20. SEQUIN at http://www.ncbi.nlm.nih.gov/Sequin/

Appendices

A.1 Free text responses from the questionnaire

A.1.1 Project aims

▪ Question 2: 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?

Free text answers:

- In my field (structural biology) we already link our primary data to publications
- Only useful for a few key papers per year
- Crucial for bioinformatics and as a group we are often quite good at this

▪ Question 3: How advantageous to you would it be if it were possible to go directly from within an online publication (electronic article or other text) to the primary source data from which that publication was developed?

Free text answer:

- This is something that I spend a lot of time trying to do. I sort through the data to find the primary source, but it is not always easy, for example if there are several papers from a group at one particular time.

A.1.2 Source Data

▪ Question 4: What kinds of electronic source data do you produce?
Free text answer:

- Structure factors (crystallography)

A.1.3 Source repositories

▪ Question 7: To which source repositories do you submit your data?
Free text answer:

- Internet

▪ Question 8: How often have you submitted data to any of these source repositories?
Free text answer:

- On several occasions

A.1.4 Metadata

▪ Question 9: By selecting from the following options, please would you indicate what types of metadata you consider it important to assign to your data.
Free text answer:

- I don't understand the question, probably because I don't use metadata

▪ Question 11: Who assigns metadata to your research data?
Free text answer:

- No formal metadata is used

A.1.5 Data access and sharing

▪ Question 17: Normally, what kind of formal restrictions '''do you apply''' to your research data?

Free text answers:

- Occasionally we don't release data until an associated publication comes out
- I have not shared my research thus far

▪ Question 18: What measures do you '''normally use''' to control access to your data by others?
Free text answers:

- I have not shared my research thus far
- Only give access to individuals with which we collaborate
- The question has never arisen
- Only data of published works are made available
- Requests judged on merit

A.1.6 Output repositories

▪ Question 19: Which kind of output repositories do you use to find and retrieve information for use in your research?
Free text answer:

- Aspergillus web site

A 2: Interview questions

StORe interview No:

Interview list:

Date:

Name:

Position/Role:

Institution:

Discipline:

Your field of interest?

Research process

1. What would be the typical points for you to access information, to store, and to share information during a research project?

Source repositories

2. What sort of data do you generate? Where and how do you store it? In which format is this data stored? Do you find it easy to store your data? Would you want anything different?
3. Do you currently share access to your data? Why would you want to control access to your data? What measures do you use to make your research data available? What factors would discourage you from sharing your research data?

Metadata generation – *in source repositories or when you send an article to a publisher*

4. How do you assign metadata to your data, such as when it was created and by whom, what project it is related to, or maybe keywords so that it can be searched?
5. Who assigns metadata to your data?

Output repositories

6. How do you access output repositories? Which repositories do you access? Which routes do you use?
7. How do you search for relevant information (keyword search, advanced search, author search)? Do you receive support and/or guidance?
8. What are your experiences with accessing output repositories? Are there any that you don't like using and why? What are the problems?
9. Do you access the source data directly from publications? Do you contact the researcher? Would you find it convenient to be able to link through?
10. What are the data formats that you are interested in?
11. You may know that there are plans to create an institutional repository for the FLS. What do you think would be the advantages and disadvantages for you and your research?

Cross-disciplinary search

12. Do you regularly access source data or scientific publications from other disciplines? Which publications? What is your experience with accessing source data from other disciplines? Do you find it easy to get through to the data you want? What would be a reason for you not finding the information you want?
13. What would you think about a common standard for access to repositories in different disciplines?

Help functions

14. Would you find help facilities useful when you are accessing a repository?
    (e.g. on-line help, two-directional conversation, user ratings, FAQ, terminology)
15. What are your experiences with help functions? How do you find information if you are not using a help function?

A 3: Scenario 4 – detailed version

This is the original, detailed version of a standards task in bioinformatics. It has been developed together with a bioinformatician in a second interview, where a typical user situation in bioinformatics was evaluated.

Predicting the function of a novel gene – search for motifs

Narrative:

A bioinformatician receives a request from a molecular biology group, by e-mail. They have found a new genetic sequence of potential interest.
However, when they used it for a BLAST search, the only result was a list of proteins of uncertain homology (see Scenario 3). A complementary method for predicting the function of an unknown gene sequence is to compare the translated sequence to patterns of conserved features or motifs in known protein families, a standard task in bioinformatics.

Action steps:

1. S/he runs a search of the query sequence against InterPro, the integrated protein family database. However, knowing that InterPro is not fully synchronised with its source databases, s/he runs separate searches against these too (i.e., against PROSITE, PRINTS, Pfam, etc.), using the search tools available on each of the database home pages. For each search, s/he compares the results to ensure consistent answers, and cross-references them back to the BLAST result.
2. The bioinformatician uploads the translated sequence into CINEMA, an in-house manual sequence alignment editor (10). S/he creates a multiple alignment of the query sequence with the InterPro/BLAST-consistent sequences, highlighting any common motifs. If the alignment is hard to do by hand, s/he might use an automatic alignment tool (e.g., ClustalW or T-coffee), but then use CINEMA to manually refine the result.
3. S/he also carefully checks the Swiss-Prot Feature Table (11), which contains sequence annotations (metadata and biological information), often derived from sequence analysis tools - this might highlight, for example, any hydrophobic, potential transmembrane domains; glycosylation sites; phosphorylation sites; lipid-attachment sites; repeats; low-complexity regions; disulphide bridges, and so on. S/he compares the highlighted motifs with the annotated regions in the Feature Table and, with his/her extensive background knowledge, s/he decides which of these features might be structurally or functionally significant - s/he is likely to check the literature cited in the Swiss-Prot entry to see if any of the regions annotated in the Feature Table are supported by an experimental finding.
4. There is minimal or no information on the motifs.
5. The result identifies a common domain. The bioinformatician produces a table of motifs, together with relevant biological information and sequence annotations, and sends this back to the molecular biology lab.
6. The lab verifies the results.

Extensions:

1.a The query does not give any significant results, i.e. there are no discernible common motifs within the query and BLAST sequences. This could be due to errors in the sequence. S/he aborts the search and contacts the molecular biology group, so that they can re-run their sequencing experiments.

2.a S/he needs more information about the motifs and performs an information search -see Scenario 1.

3.a There is only minimum annotation. S/he decides that s/he needs more information and performs an information search -see Scenario 1.

4.a There is minimal or no information on the motifs. S/he decides to create a discriminator (e.g., a fingerprint, a regular expression, a profile, etc.) to perform more specific searches of Swiss-Prot/TrEMBL. In this scenario, s/he opts to use in-house software to create a fingerprint.

S/he excises the motifs, runs separate motif searches of Swiss-Prot/TrEMBL, then compares to the hitlists to discover which sequences have been identified that match all the motifs in the correct order. Sequence information from any newly identified family members is added into the motifs, and the database searches are repeated until the process converges. The result is then annotated with whatever minimal information could be gleaned from Swiss-Prot, and uploaded into the PRINTS database - once there, it provides a diagnostic signature to recognise new family members whenever Swiss-Prot and TrEMBL are updated. The full family membership identified by the fingerprint is passed to the molecular biologist- see Scenario 3.

4.b There is minimal or no information on the motifs, but one of the matched sequences has had its 3D structure determined. S/he decides to use the ExPASy Swiss-Model server (14) to produce a homology model of the query sequence using the known structure of the matched homologue as a template.

5.a The bioinformatician saves this region of the sequence into a separate FastA file, and then uses this to run a new BLAST search. As this search has now been narrowed down to a specific domain, the resulting homologues are far more significant.

6.a The lab runs a new series of molecular biology experiments.

1.a 1. Re-sequencing confirms the original sequence - hence there are still no homologues identified by BLAST and no common motifs identified in a multiple alignment. The bioinformatician has a few options remaining:

1.a.1.1 Again, s/he looks at the Swiss-Prot Feature Table to identify any possible functional sites – go to 3.

1.a.1.2 S/he runs standard bioinformatics software tools to identify features not already annotated in Swiss-Prot - e.g., on the Web, there are tools for identifying potential post-translational modifications (glycosylation sites, GPI anchors, signal peptides, etc.) and protein sorting signals (PEST regions, leucine zippers, etc.); for calculating protein physicochemical parameters (PI, aliphatic index, composition, etc.); for plotting hydropathy profiles; for identifying low-complexity regions, and so on;

1.a.1.3 S/he runs structure prediction algorithms to attempt to 'guess' the secondary or tertiary structure of the protein.