Source-to-Output Repositories

Source to Output Repositories Project: Chemistry

Panayiota Polydoratou

Imperial College London
Central Library

August 2006

Contents

List of figures

List of tables

Acknowledgements

There are several people that I would like to thank for their help and support in the process of this research study and the writing of the report.

I would like to thank Jenny Evans, liaison librarian for Chemistry and Physics at the time of this research study, for her help in identifying and contacting relevant academic and research staff at Imperial College London. Furthermore for her help in publicizing the questionnaire survey and for reviewing written parts of this report and articles for publication.

Also, I would like to thank the IT team of the Imperial College Central Library, in particular Siamand Salehian for his help in identifying, installing and supporting the use of software to record and analyse recorded interviews.

Last but not least, Adrian Clark for his support and helpful suggestions throughout the research study.

Section 1 - Introduction

1.1 StORe project

The **StORe project** (http://jiscstore.jot.com/WikiHome) is a collaboration of seven universities across the UK and the Johns Hopkins University in the USA and under funding from the Joint Information Systems Committee (JISC, http://www.jisc.ac.uk/). The project sought to develop new ways of linking academic publications with repositories of research data. Within the scope of the StORe project, an online questionnaire survey was launched on the 13th March and ran until 21st April 2006, aiming to gain some understanding about how repositories can be used to support research and scholarly communication and to invite researchers to provide feedback about their use of source and output repositories and their expectations for their future development. In addition to the questionnaire survey, academic staff and postgraduate research students were interviewed in order to gain a deeper insight into their research methods.

This report presents the findings from this online questionnaire survey and from interviews with academic staff and postgraduate research students in the domain of chemistry.

1.2 Definitions

The terms source repository and output repository are used throughout this paper. They are defined as:

- ***Source repository.*** Source repositories contain the **source** or **primary data** produced during a programme of research, and comprise the source from which research publications will be developed.
- ***Output repository***. An output repository usually contains published articles or other texts, although it may hold other data objects that have been published. The contents of an output repository will typically include publications at a pre- or post-refereeing stage, working papers, research reports and PhD theses.

1.3 Methods

The StORe project (http://jiscstore.jot.com/WikiHome) employed two methods to gather information about the use and the linking of source and output repositories based on the opinions and experience of researchers in seven scientific domains. Those two methods were: a) an online questionnaire survey and b) interviews with members of academic staff from academic institutions across the UK.

1.4 Questionnaire survey sample

The questionnaire survey was launched on the 13th March and closed on the 21st April 2006 and was publicised among 728 members of the chemistry research community at the following universities:

- Imperial College London
- University of Bristol
- University of Cambridge
- University of Southampton

- University of Durham
- University of Oxford
- University College London

The target group included academic and research staff engaged in chemistry research and wherever the information was available, postgraduate research students were also contacted. The university departments surveyed were selected on the basis of their outcome in the 2001 Research Assessment Exercise, selecting those who had been rated 5 *.

Members of academic and research staff were contacted via email during the week beginning 27th March 2006 and were invited to participate to the survey. Members of staff and postgraduate research students at Imperial College London were additionally sent background information concerning the StORe project via the chemistry newsletter produced by our liaison librarian, alerting them to the questionnaire. An email reminding them to participate in the survey was sent during the week beginning 3rd April and again the week beginning 17th April.

For the purpose of this study, the areas identified in the 2001 RAE assessment in the field of chemistry were used to identify members of staff and students conducting research in each field. The intention was to obtain, if possible, representative examples of research patterns from all chemistry research fields.

1.5 Response

Thirty eight (38) chemistry staff and students participated to the questionnaire survey representing 10% of the overall response. The relatively low response has been attributed to several factors, such as survey fatigue, the timing of the survey which coincided with the exam period and then the Easter holiday and the fact that academic community did not appear to be familiar with JISC, digital repositories or repositories in general.[1]The majority of respondents were from Imperial College London chemistry staff and research students, followed by those from the University of Cambridge. The response from each academic institution contacting in the survey is presented in the following table.

	Contacts	**Response**	**Response rate within each institution**
Bristol University	99	1	1.0
University of Cambridge	68	10	14.7
Durham University	46	0	0.0
Imperial College London	330	14	4.2
Southampton University	57	4	7.0
University of Oxford	78	2	2.6
UCL	50	2	4.0
Unknown	n/a	5	n/a
Total	**728**	**38**	

Table 1: Number of contacts for the questionnaire survey and response rate by institution

Almost half of the response (47%) came from postgraduate research students. Another 40% from academic staff and the remaining 13% represents the response by postdoctoral researchers, research assistants and contracted researchers. There was no response from any independent researchers. The undergraduate community was not

[1]Pryor, Graham. Linking research papers and research data: possibilities for a generic solution. Presentation at the DRP Workshop - StORe at WWW06. http://jiscstore.jot.com/WikiHome/DisseminationPages/WWW06-SSVY.ppt (Last accessed 04/09/2006)

targeted as a group and therefore does not feature in the survey responses. Analytically the response is presented in the following table.

Role:	Number of respondents	%
Academic staff	15	**39.5**
Research Assistant	2	**5.3**
Postgraduate	18	**47.3**
Undergraduate	0	**0**
Contract Researcher	1	**2.6**
Independent Researcher	0	**0**
Other (*please insert*)	2	**5.3**
Total	**38**	**100**

Table 2: Response to the questionnaire survey by role of the respondents

1.6 Interviews sample

Respondents to the questionnaire survey were invited to indicate if they would be willing to participate to a more in depth interview. Eleven people replied that they would be prepared to participate further in the StORe project survey, however, some of these respondents did not reply when contacted directly to arrange this. Other researchers/academic staff however expressed interest in participating in an interview after their colleagues had done so. With the level of interest in the project increasing in this way, eventually seventeen (17) interviews were conducted with members of academic staff and postgraduate research students. The interviews took place during the months of May and June 2006. The majority of the interviewees were from Imperial College London and their colleagues at the University of Cambridge. The breakdown of the interviewees by academic institution is shown in the following table.

University	Number of interviewees
University of Cambridge	5
Imperial College	9
UCL	2
University of Southampton	1
Total	**17**

Table 3: Number of Interviews conducted by institution

All the interviews were conducted face to face They lasted between 25 minutes and one hour each and were recorded. In some cases, written notes were also taken. The majority of the respondents represented two fields of chemistry research; experimental/synthetic chemistry and theoretical/computational chemistry. There was also one representative from the crystallography community. The number of interviewees and information about their role and fields of interest are presented in the following table.

Role	Position	University	Field of interest	Type of interview
PGR	PhD student	IC	Biological	Face to face
Academic staff	Lecturer	Cambridge	Computational/ Theoretical	Face to face
Academic staff	Professor	UCL	Computational/ Theoretical	Face to face
Academic staff	Reader	IC	Computational/ Theoretical	Face to face

Academic staff	Professor	Cambridge	Computational/ Theoretical	Face to face
Academic staff	Professor	IC	Computational/ Theoretical	Face to face
Contracted researcher	Research associate (postdoctoral)	Cambridge	Computational/ Theoretical	Face to face
Contracted researcher	Research assistant (postdoctoral)	Cambridge	Computational/ Theoretical	Face to face
PGR	PhD student	Cambridge	Computational/ Theoretical	Face to face
Contracted researcher	Postdoctoral researcher	UCL	Computational/ Theoretical	Face to Face
Contracted researcher	Postdoctoral researcher	IC	Computational/ Theoretical	Face to face
Academic staff	Professor	Southampton	Crystallography	Face to face
Contracted researcher	Postdoctoral researcher	IC	Physical	Face to face
PGR	PhD student	IC	Polymer	Face to face
Academic staff	Reader	IC	Synthetic	Face to face
Academic staff	Lecturer	IC	Synthetic	Face to face
PGR	PhD student	IC	Synthetic	Face to face

Table 4: Presentation of the interviewees by role and field of interest

1.7 Background: Repository development in chemistry

1.7.1 Source repository (EPSRC UK National Crystallography Service)

The source repository, eCrystals – University of Southampton is the archive for Crystal Structures generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service.

The repository description states (http://ecrystals.chem.soton.ac.uk/information.html) "*The information contained within an entry in this archive is all the underlying data generated during the course of a structure determination from a single crystal x-ray diffraction experiment. This information is freely available and allows the reader to a) assess the validity of the dataset or b) repeat the experiment or c) use the data for further studies.*

An individual entry consists of three parts:

- *Core bibliographic data, such as authors, affiliation and a number of chemical identifiers,*
- *Data collection parameters that allow the reader to assess at a glance certain aspects of the crystallographic dataset,*
- *Files available for download. These files are: visualisations of the raw data (.jpg), the raw data itself (.hkl), experimental conditions (.htm), outputs from stages of the structure determination (_xs.lst, _xl.lst & .res), the final structural result (.cif & .cml) and the validation report of the derived structure (_checkcif.htm)".*

The eCrystals repository is running on the GNU EPrints open archive software and users can either browse or search the repository. Information can be browsed by year, person and compound class while it can be searched by various options More information about

the eCrystals repository can be found at the following URL: http://ecrystals.chem.soton.ac.uk/

1.7.2 Output repository (Imperial Eprints)

The Imperial Eprints (http://www3.imperial.ac.uk/library/digitallibrary/imperialeprints) is an open access repository that has been established as a pilot project, focussing on the research output from the Faculty of Physical Sciences at Imperial College London. Material stored in the archive is freely available over the internet. Information from the project's description states that development work on Imperial Eprints has been funded through the SHERPA-LEAP Project. SHERPA-LEAP is a University of London consortium, which has created institutional repositories at 7 institutions. The project is also an associate partner of SHERPA (http://www.sherpa.ac.uk/) a network of over 20 institutional repositories.

Adopted from the Eprints website, "*Imperial Eprints contains post-prints of journal papers (the final refereed version, which have been accepted for publication), and similar material such as book chapters, conference papers and technical reports. By storing the full text of papers deposited by members of Imperial and making them available on the web, the Imperial Eprints repository will help the worldwide scholarly community to discover and retrieve Imperial College London research.*

The repository has a simple interface for retrieval by browsing or searching. Information about the content of the repository will also be harvested by international Open Archiving services such as OAIster (http://oaister.umdl.umich.edu/o/oaister/) and Google Scholar http://scholar.google.com/).

Imperial Eprints is not a publishing mechanism, nor is it a substitute for peer-reviewed journals. It has been developed to host material that has already been, or is about to be, published elsewhere".

The Imperial Eprints runs on GNU EPrints / revision: EPrints 2.3.0 (Hoi Sin Duck) [Born on 2004-01-12] and users can either browse or search the repository. Information can be browsed by year and subject and searched by various options.

Information about Imperial Eprints can be found at the following URL: http://www3.imperial.ac.uk/library/digitallibrary/imperialeprints. It should be noted that Imperial College London is currently in the process of launching an institutional repository, building on the experience gained from the Imperial Eprints pilot project. In the near future, more comprehensive coverage of the research outputs from the Faculty of Physical Sciences (which from 2006 forms part of the Faculty of Natural Sciences), will be available from the Institutional repository.

Section 2 – Summary of significant observations from both questionnaire and interviews

2.1 Identities

- Thirty eight (38) chemists replied to the questionnaire survey and seventeen (17) were interviewed.
- 47% of the questionnaire response was received from postgraduate research students, 39.5% from academic staff, 5.3% from research assistants and another 5.3% from postdoctoral researchers and 2.6% from contract researchers.
- 47% of the interviewees were academic staff, 29% contract researchers and 23% postgraduate research students.
- 59% of the interviewees were researchers in computational/theoretical chemistry and 18% from experimental/synthetic chemistry. Other fields of interest represented in the study were members of the crystallography community and one from each of the following areas; biological, physical and polymer chemistry.
- Almost half of the interviewees (47%) were academic staff at a very senior level within their disciplines. In particular, four professors, two readers and two lecturers were interviewed. The remaining interviewees were either postdoctoral contract researchers (29%), or postgraduate research students (24%).

2.2 Project aims

- More than half of the chemists that replied to the questionnaire survey (65%) noted that they had not used a repository before and they were not familiar with the idea of open access repositories in general. Those chemists who participated in the interviews however, once there had been an opportunity to explain the terminology used in the survey, indicated that they had been long term and consistent users of certain source repositories, such as the Cambridge Structural Database.
- Academic staff were more interested in linking from the primary research data to the published outcome of the research, while PhD students and postdoctoral researchers were more interested in navigating from the published outcome to the primary data sets.
- More than two thirds of the response from academic staff (67%) indicated that they considered the linking from primary research data to the published outcome as useful but not of major significance for their work. The majority of the respondents in the same group (73%), replied that linking in the opposite direction, from the published research outcome to the primary research data would be useful for their work.
- The postgraduate research students (67%), research assistants (100%) and the contract researchers (100%) noted that it would be a significant advantage to be able to link from the published outcome of research to the primary set of data.
- The majority of the chemistry respondents noted that the ability to link from the published outcome of the research to the primary research data would be either a significant advantage to their work (57%) or a useful feature (29%). The opposite feature, to be able to link from a source repository to the published outcome of the research was greeted by almost half of the respondents (41%) as a significant advantage. Another one third (33%) indicated that this option would be useful for them but not of major significance.
- The respondents suggested the following as missing functionality from source repositories:
    - complex hyperdocuments
    - more cross linking to output repositories

    - a search engine like SciFinder or Google
    - access to raw data such as SPECTRA
- Missing functionality from output repositories included suggestions such as:
    - incomplete electronic coverage of hard copy sources, e.g. past journal issues,
    - peer reviewing,
    - a method to filter the journals not accessible from the researcher's institution,
    - in text linking of SPECTRA, procedures,
    - data to compounds,
    - a more direct link to URL directing to the primary file (minimal number of mouse clicks) and hyperlink from article on publisher's site to relevant source data.
    - Supporting information is woefully inadequate - eg coordinates of calculations are in pdf format - it takes a considerable amount of time to convert them into a form suitable for visualisation in molecular modelling software...
- Some of the concerns that the interviewees expressed about the use of source repositories were related to how such a service can compete with and/or complement existing services provided by commercial/publishers' repositories. Aspects such as the quality of data and mechanisms for quality assurance, the comprehensiveness of data and the ways it can be accessed, were raised as very important considerations if the use of source repositories of primary research data is to become more widespread. Other factors that it was felt would contribute to the success and/or the increased use of source and output repositories were the publicising of the potential benefits that such a service could have for the research community and the involvement of scientists in its development. Also important were issues regarding the sustainability, management and maintenance of both source and output repositories.

-

2.3 Source data

- There are many variations in the type of data produced, their recording and storage. The perceived value of repositories also varies widely throughout the chemistry research community. The most common type of data produced among chemists is SPECTRA data, which is represented in drawings, spreadsheets and image files. This finding was also confirmed by the survey interviews. Chemistry researchers produce a variety of different types of data, often in large volumes. The majority of the computational/theoretical chemists that were interviewed indicated that they do not produce raw data in the same sense that other scientists do. They are primarily involved in the development of the methods that test how molecules behave in certain conditions; their research is about the development and testing of methods that measure the energy, the geometry and the particular arrangement of molecules under certain conditions. Therefore although they produce data in the form of calculations and measurements' testing, they tend to apply their methods to other researchers' published outcomes.
- Other types of data referred to "mainly binary and text files from calculations, with figures and graphs derived from these".
- The most popular formats in which the data is saved and held included spreadsheets (76%), word-processed files (74%) and image files (68%).
- Other suggested formats included a variety of standards and software associated with the production and description of data in the chemistry research community such as:
    - .cif (crystallographic data),
    - binary data files,
    - chemdraw

  - cdx. xwin nmr files,
  - Chemdraw Word,
  - Chemical Markup Language,
  - Corel Draw,
  - Fourier induction decay files (generated from Bruker and Varian NMR instruments),
  - Spectra are in spectrometer specific code.
- Academic staff and research assistants noted the main reason why they might want to make use of a source repository was “to access data that are useful or necessary for my research”. The postgraduate students and the postdoctoral researchers noted the reason “to understand the broader context and orientation of my research”. The identification of useful contacts for their research is more important to the postgraduate students, in comparison to academic staff and the contract researchers. Also, testing the validity and uniqueness of the research objectives is more important to research assistants than any other group.
- The majority of the respondents indicated that they access other researchers’ data through access to source repositories. It is believed that the respondents might have confused the term source repository with repositories in general as this finding is in conflict with the initial questionnaire response indicating that more than half of the respondents (65%) had not used a repository in the past and in general they were not familiar with the idea of open access.

2.4 Source repositories

- More than half of the chemists that replied to the questionnaire survey (65%) had not used a repository before and they were not familiar with the idea of open access repositories in general. Those who had used one made specific mention of the National Crystallography Service, the Cambridge Crystallographic Data Centre, specific publishers and institutional repositories. As already noted in section 2.1 however, those chemists who agreed to an interview, indicated that they had been long term and consistent users of certain source repositories such as the Cambridge Structural Database, once the terminology that was used in the survey had been explained.
- Primary research data is obtained from various sources. Examples included the commercial companies that participate in research projects, existing established work models within a research group, use of repositories/databases/Internet resources and use of data produced mainly by crystallographers.
- Quality control of the data in the repositories, comprehensiveness and maintenance were considered issues of primary importance for the use of source repositories.
- Accessing and using both source and output repositories were envisioned as part of a wider context of supporting research conduct and publication.

2.5 Metadata

- The author/creator was the most important metadata element for 89% of the chemists. Other important metadata elements were the project’s description (68%), the project’s title (68%) and the assignment of subject keywords (68%). The date and the title of the data set (each at 58%) were equally important. The least important metadata was considered to be the funding source of the project (13%).
- More than one third of the chemistry respondents (37%) noted that metadata is assigned to resources during file saving which indicates the involvement of software for automatic assignment of metadata. The second most popular choice was that metadata is assigned prior to data creation (26%) while one quarter of the respondents noted that metadata is either assigned as part of the indexing process

for source files (24%) or no metadata is assigned (24%). Few of the respondents (8%) noted that metadata is assigned at a later stage, usually after the submission of the data to the repository and another small group of respondents (8%) indicated that they were not sure when metadata is assigned.

- More than half of the chemistry respondents to the questionnaire survey (53%) noted that they themselves decide both on the terms to use and the assignment of metadata. More than one third (29%) of the respondents replied that it was not known to them who assigns the metadata to their resources.
- The assignment and use of metadata as a minimum set of requirements that facilitate the deposition and retrieval of research data, did not appear to be popular or in some cases, even understood by the majority of chemistry researchers that were interviewed. Those chemists that were familiar with metadata though, possessed an in-depth knowledge of metadata's use, application and functions. The assignment of metadata automatically or by implementing a system that relieves the depositor of having to do it, was also commented upon positively.

2.6 Data access and sharing

- The response was spread and there is not one single factor that appears to be significantly important, that would encourage the respondents to share access to their data. For academic staff the potential benefits to the research community appear to be the most appealing factor (34%) followed closely by the demonstrable benefit for their research profile (32%). Similarly this is the most important factor for postgraduate research students (34%). It is the research assistants (21%) and the contract researchers (5%) that noted the requirement of a funding body/condition of funding, as the primer factor that would encourage them to share their data.
- In general, the threat of the loss of ownership is a strong factor for postgraduate research students (37%) along with the risk of premature broadcast of research findings (37%) preventing them from sharing access to their data. Academic staff noted the time/effort required to enable sharing (32%) and the risk of premature broadcast of their research findings (29%) as the most important reasons. The most important factor for research assistants were the risks to an established research niche (24%). These findings were also confirmed by the interviews. The majority of the interviewees denoted that they were reluctant to share access to their data whilst they were still in the process of conducting their research. Exceptions existed in the form of personal communication and contacts. Positively though, all interviewees were happy to publish and share access to their data once the project had finished and the research outcomes had been published.
- Academic staff (24%) and postgraduate research students (21%) replied that they do not apply any formal restrictions to their data. The research assistants though, noted that there are restrictions imposed on the research team and members (18%). A restriction noted by respondents from all groups was that they respond to individual enquiries/requests for access and they judge them based on their merits.
- Academic staff were the only group that noted other restrictions. In particular they specified: creative science commons, ownership retained - request acknowledgement on re-use. Restrictions may also vary with the maturity of project or be applied to respect collaborators requirements or those of the research programme sponsor.
- All respondents indicated a variety of measures that they use to control access to their data. The majority of responses from academic staff indicated storage of their data on a private network/intranet (21%) as the main measure to control access. The same measure was indicated as the main one by the majority of the postgraduate research students (32%) as well. All of the contract researchers noted that they use authentication of ID and password for online access, for controlling access to their data. The research assistants replied that they tend to select storage of their data on standalone computers (16%) as the main measure for controlling who has access.

2.7 Output repositories

- The output repositories that the majority of chemists tend to use are those in the commercial sector set up and managed by publishers. Academic staff indicated they used the widest range of repositories, including institutional, discipline, publisher and other. The academic staff (18%) and the research assistants (3%) were the two groups that denoted they do not deposit their data in any type of repository.
- Although, the majority of chemistry respondents to the questionnaire survey replied that they preferred to use the simple search option when they visited both source and output repositories, the response is quite spread in relation to the different types of repositories they consult. Those who tend mostly to use the publishers repositories, prefer to search employing simple methods. The use of subject specific thesauri and the use of Boolean logic are only mentioned in the searching of institutional and discipline repositories.
- Results from the face to face interviews with the chemistry staff and students revealed that Internet based services such as Google Scholar are amongst the most popular means of searching for information. Theoretical/computational chemists tend to use many Internet based repositories and services, depending on the type of data that they were looking for. The speed of information retrieval, the quality of the information and the comprehensiveness of the repository are important factors in their choice of output repositories

2.8 Support

- Although the majority of the responses denoted that they use a simple search when visiting a publishers' repository, the use of subject specific thesauri and Boolean logic is fairly common when they navigate institutional or discipline repositories.
- In general it was felt that the availability of a prototype that would illustrate what the StORe project proposes would have made it easier for the respondents to understand and comment upon advantages and barriers to use.
- The postdoctoral researchers and the research assistants that completed the survey were not aware of the level of support provided in their institutions or by the output repositories. The majority of academic staff (13% of all survey respondents) and all the contract researchers, indicated that there was repository-enabled support in their use of the repositories. The majority of the postgraduate research students (18% of all survey respondents) replied that they did not know the level of guidance/support that was provided.

Section 3 – Questionnaire analysis

3.1 Questionnaire results

The majority of the responses in the chemistry domain came from postgraduate students (47%) and academic staff (39%). Undergraduate students were not targeted as a group, as the use and submission of data to source and output repositories requires the participation in a research based study and the production of research data, and therefore there was no response received from them. Similarly, no response was received by any independent researchers. Those who selected the option "other" noted their current role in postdoctoral research. The response by the role of the respondents is presented in the following table.

Role:	**Number of respondents**	**%**
Academic staff	15	39.47
Research Assistant	2	5.26
Postgraduate	18	47.37
Undergraduate	0	0
Contract Researcher	1	2.63
Independent Researcher	0	0
Other (*please insert*)	2	5.26
Total	**38**	**100**

Table 5: Response to the questionnaire survey by role of the respondents

3.2 Perceived value of source-output repository links relative to different research roles

The replies of the respondents in the chemistry domain regarding the perceived value of bidirectional repository links were representative of the generally limited deployment and use of source and output repositories by this research community. More than two thirds of the response from academic staff (67%) indicated that they considered the linking from primary research data to the published outcome as useful, but not of major significance for their work (Figure 1). The majority of the respondents in the same group (73%) however replied that providing links in the opposite direction, from the published research outcome to the primary research data, would be useful for their work (Figure 2). All the researchers (including research assistants, contract staff and postdoctoral researchers) replied that they considered it either a significant advantage for their work or useful, to be able to link from a source to an output repository (Figure 1).

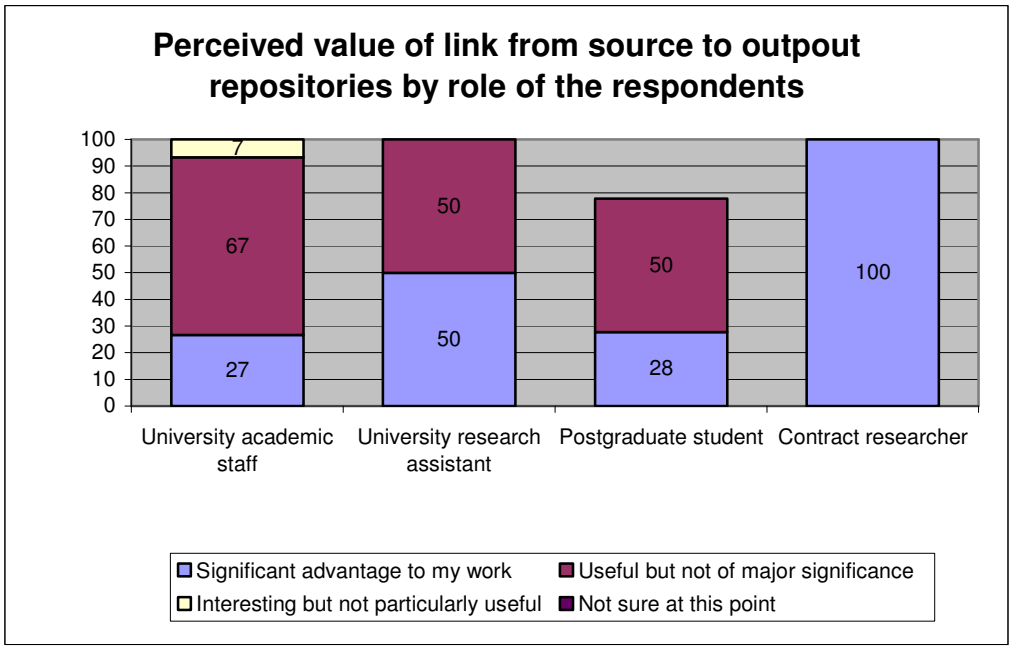


Figure 1: Perceived value of source to output repositories link by role of the respondents

Being able to link from an output to a source repository was considered by almost three quarters (73%) of the academic staff in the chemistry domain as useful but not of major significance. In view of the general cautiousness in the response regarding the use of source and output repositories in the chemistry domain, this finding should be considered encouraging. Other groups of chemistry researchers have been much more positive in the value they place on repositories. A large majority of postgraduate research students (67%), and all research assistants and contract researchers, noted that it would be a significant advantage to be able to link from the published outcome of research to the primary set of data.

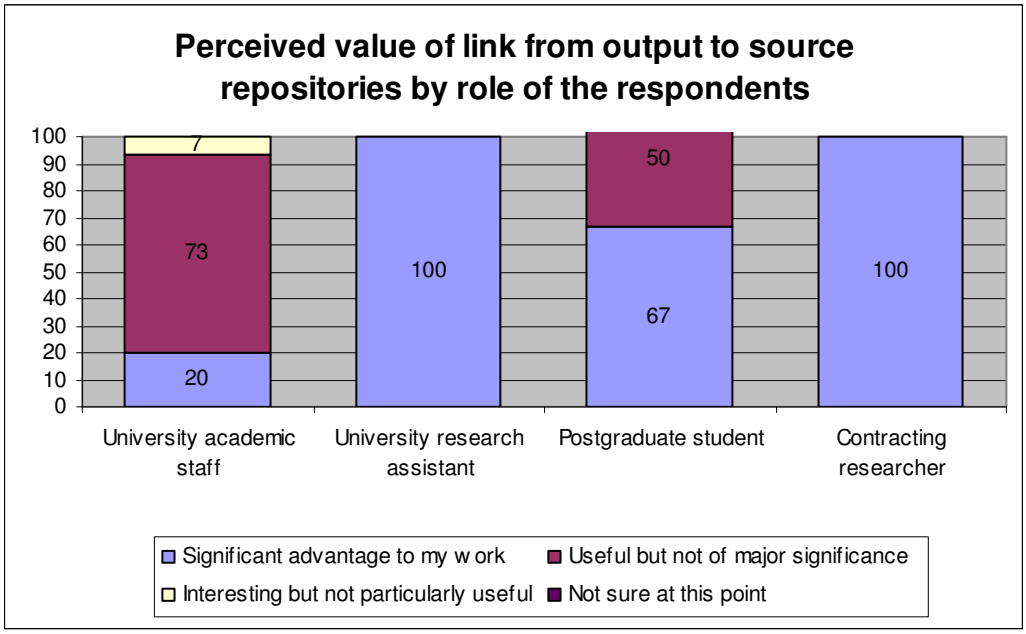
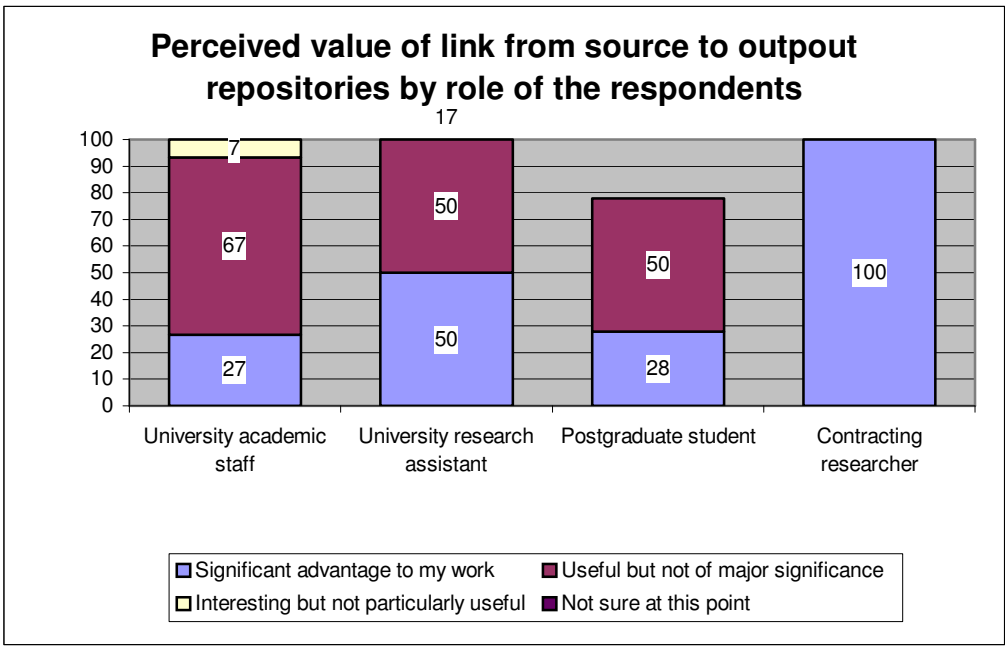


Figure 2: Perceived value of output to source repositories by role of the respondents

3.3 Perceived value of source-output repository links relative to different repository communities

The majority of the chemistry respondents noted that the ability to link from the published outcome of the research to the primary research data would be either a significant advantage to their work (57%) or a useful feature (29%). Only one of the respondents replied that they were not sure, as at the time of the survey, they had only recently commenced their doctoral studies and so could not estimate the significance such a facility might have for their research. The reverse feature, to be able to link from a source repository to the published outcome of the research was greeted by almost half of the respondents (41%) as a significant advantage. Another third (33%) indicated that this



option would be useful for them, but not of major significance.

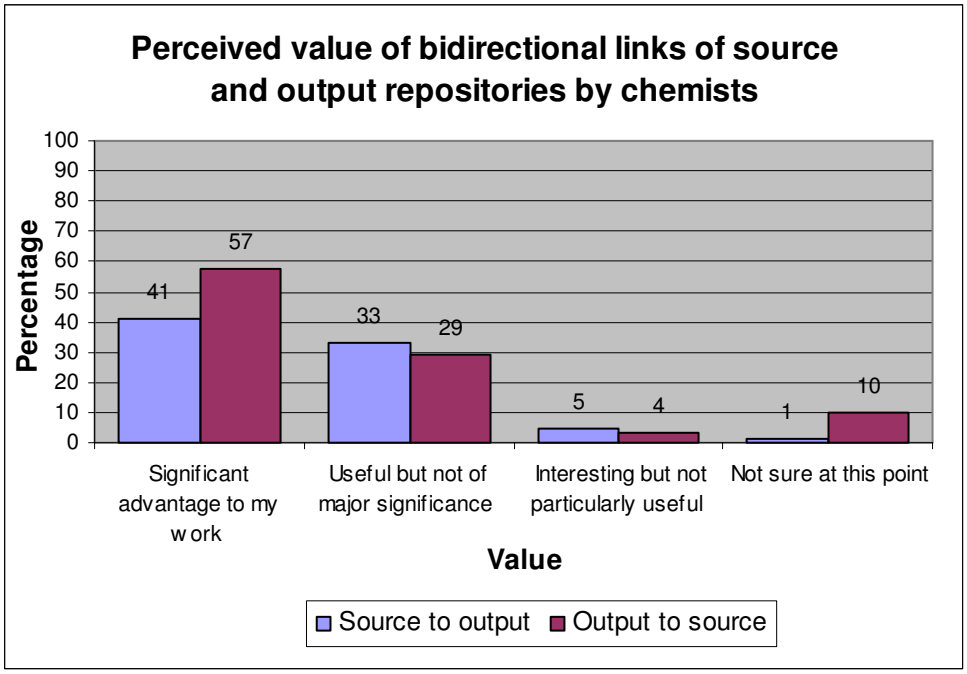
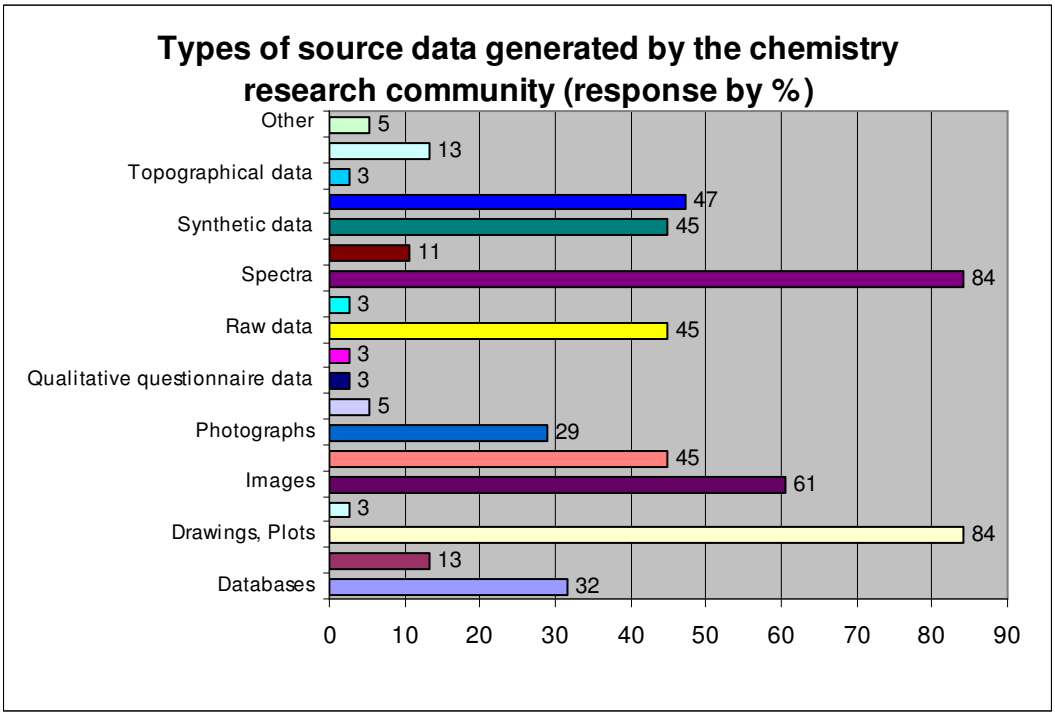


Figure 3: Perceived value of bidirectional links of source and output repositories - The case of chemists

More than half of the chemists surveyed (65%) had not used a repository before and they were not familiar with the idea of open access repositories in general. They noted however that they thought the ability to be able to link from the primary research dataset to the published outcome of the research could be either a significant advantage for their work or useful but not of major significance. Those who had used a source or output repository on a frequent basis, or on several occasions, also thought that it would be a significant advantage for their work. In general, academic staff although they considered the use of bidirectional links between repositories as either significant or useful for their research, tended to specify that this would be mainly for the application and use by their students, rather than themselves.

3.4 The different types of source data generated/held according to different repository communities

The respondents to the questionnaire survey were invited to select from a range of different types of source data that they generated in their research field. The dominant types of data in the chemistry domain were SPECTRA (84%) and drawings and plots (84%). Other types of data that were noted by almost half of the respondents were images (61%), text based data (47%), instrument data (45%), raw data (45%) and synthetic data (44%). Those who indicated other types of data, specified them as “mainly binary and text files from calculations with figures and graphs derived from these”. The types of source data that are generated by the chemistry respondents to the questionnaire survey are shown in the following figure.



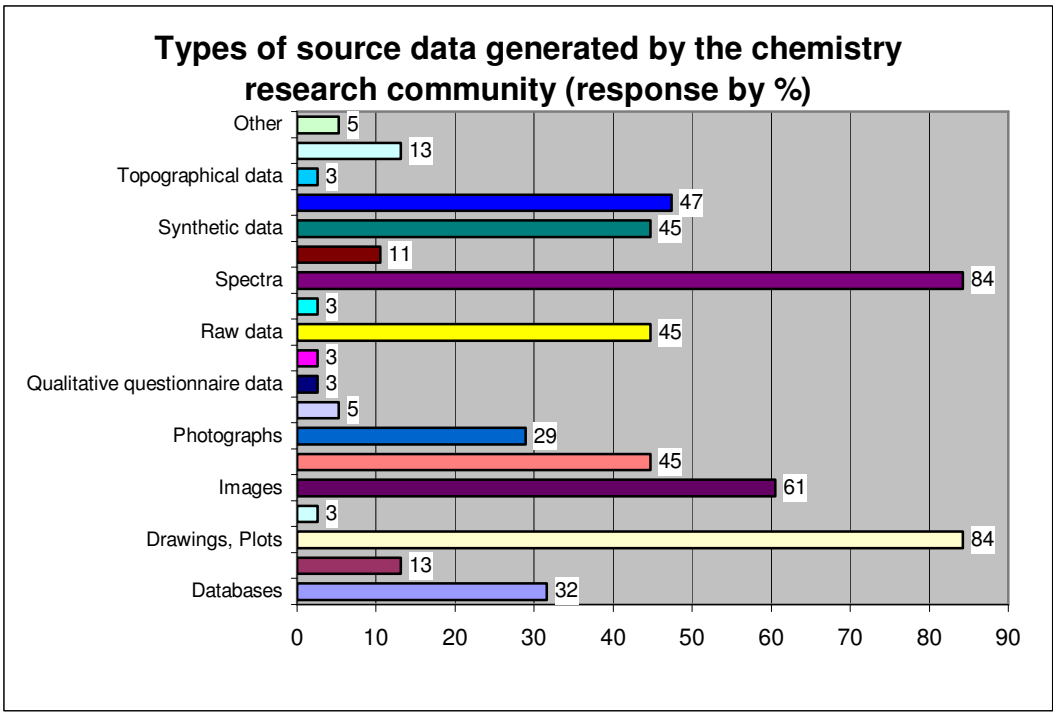
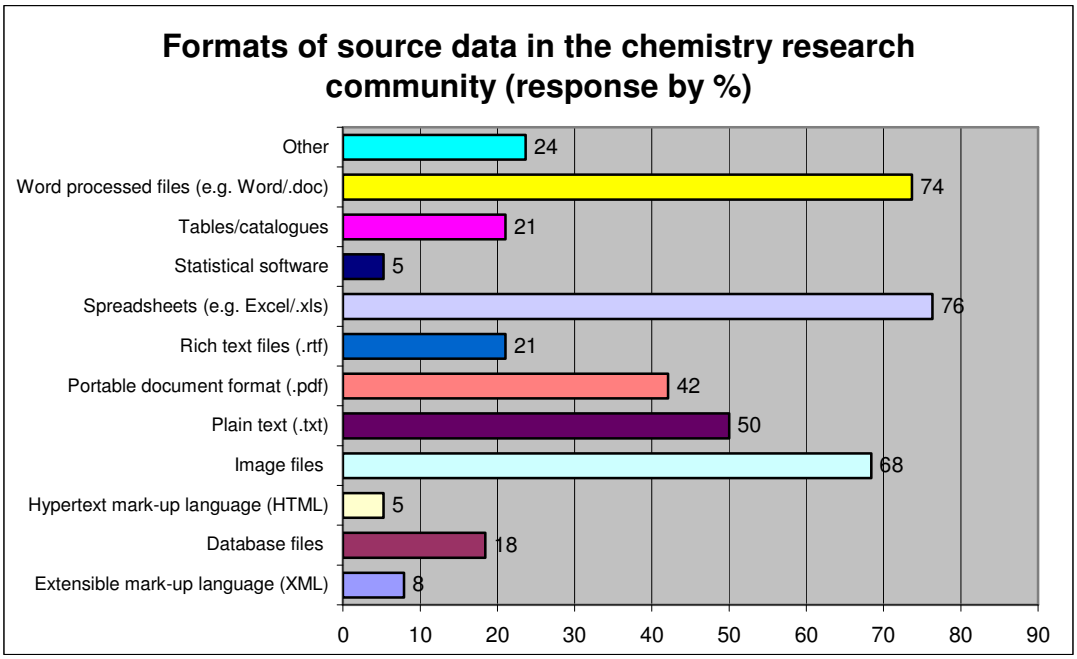


Figure 4: Types of data generated/held by chemists

The most commonly used formats in which this data is saved includes spreadsheets (76%), word processed files (74%) and image files (68%). Other popular formats among the chemistry respondents were plain text files (50%) and portable document format (42%, Figure 5). Other suggested formats included a variety of standards and software associated with the production and description of data in the chemistry research community such as:

- .cif (crystallographic data),
- binary data files,
- chemdraw
- cdx. xwin nmr files,
- Chemdraw Word,
- Chemical Markup Language,
- Corel Draw,
- Fourier induction decay files (generated from Bruker and Varian NMR instruments),
- Spectra are in spectrometer specific code.



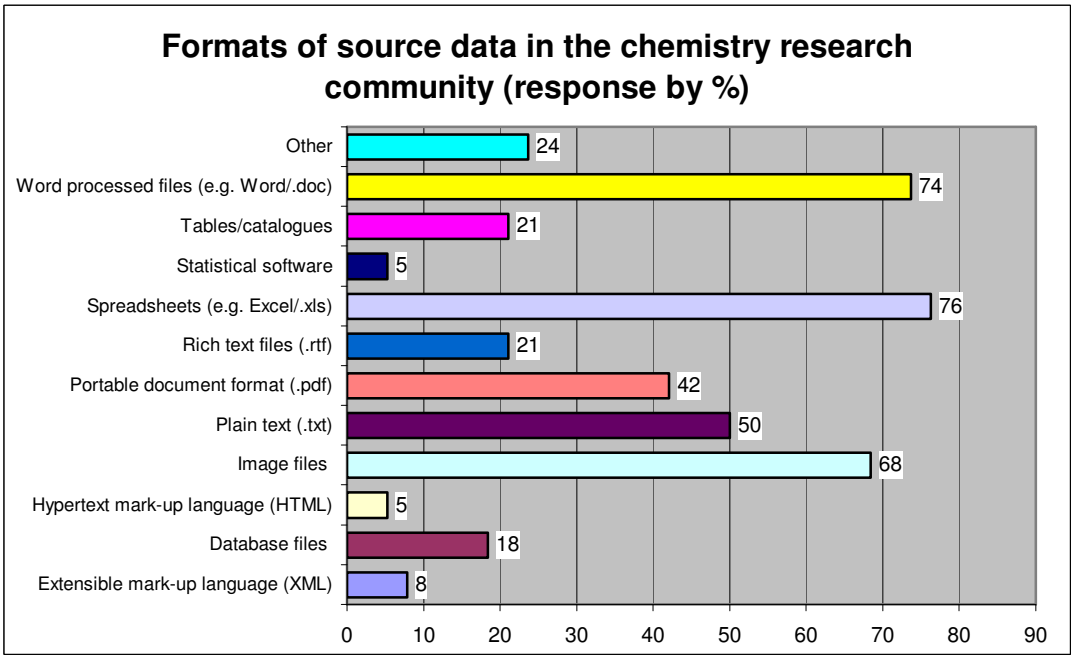


Figure 5: Types of data formats in the chemistry research community

3.5 Metadata requirements according to different repository communities

The respondents were invited to select from a list with metadata options, choosing those they considered were the most important to assign to their data. The majority of the chemistry respondents (89%) noted that the author and/or creator's name was the most significant metadata element for their data. Other important metadata elements were the project's description (68%), the project's title (68%) and the assignment of subject keywords (68%). The date and the title of the data set (each at 58%) were equally important. The least important metadata was considered to be the funding source of the project (13%).

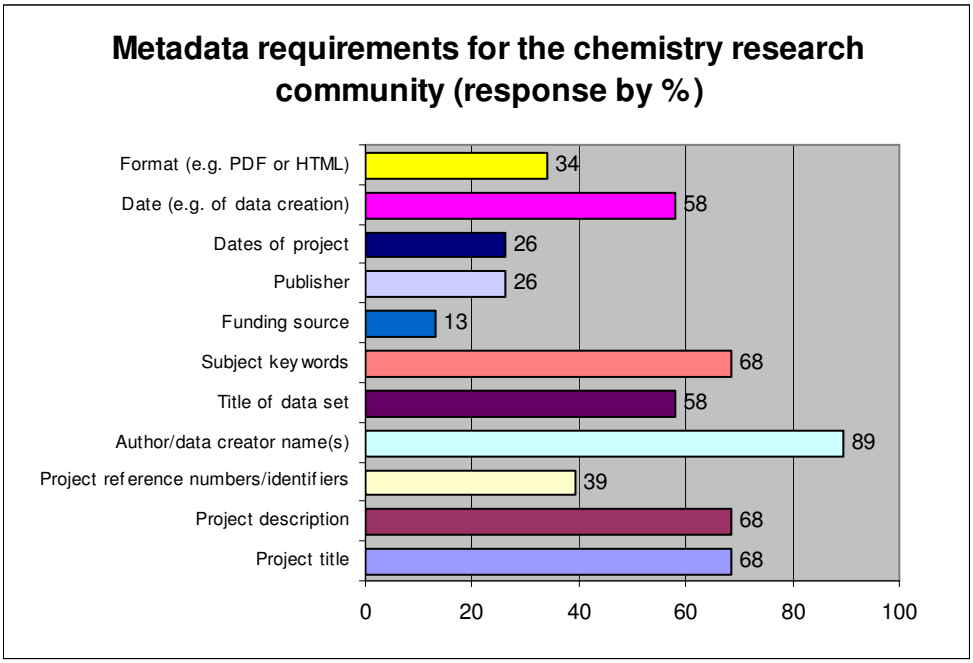


Figure 6: Metadata requirements for chemists

3.6 Metadata assignment practices relative to the level of support provided in the use of repositories

The respondents to the questionnaire survey were invited to select from a range of stages that metadata is assigned to a resource and to indicate those that were applicable to their own processes and practices. The responses to this question were spread amongst the options offered, which indicates that the respondents were not familiar with the process or the concept of assigning metadata to their resources. More than one third of the chemistry respondents (37%) noted that metadata is assigned to resources during file saving, which indicates the involvement of software for automatic assignment of metadata. The second most popular choice was that metadata is assigned prior to data creation (26%) while one quarter of the respondents noted that metadata is either assigned as part of the indexing process for source files (24%) or no metadata is assigned (24%). Few of the respondents (8%) noted that metadata is assigned at a later stage, usually after the submission of the data to the repository and small group of respondents (8%) indicated that they were not sure when metadata is assigned. Results are presented in the following figure.

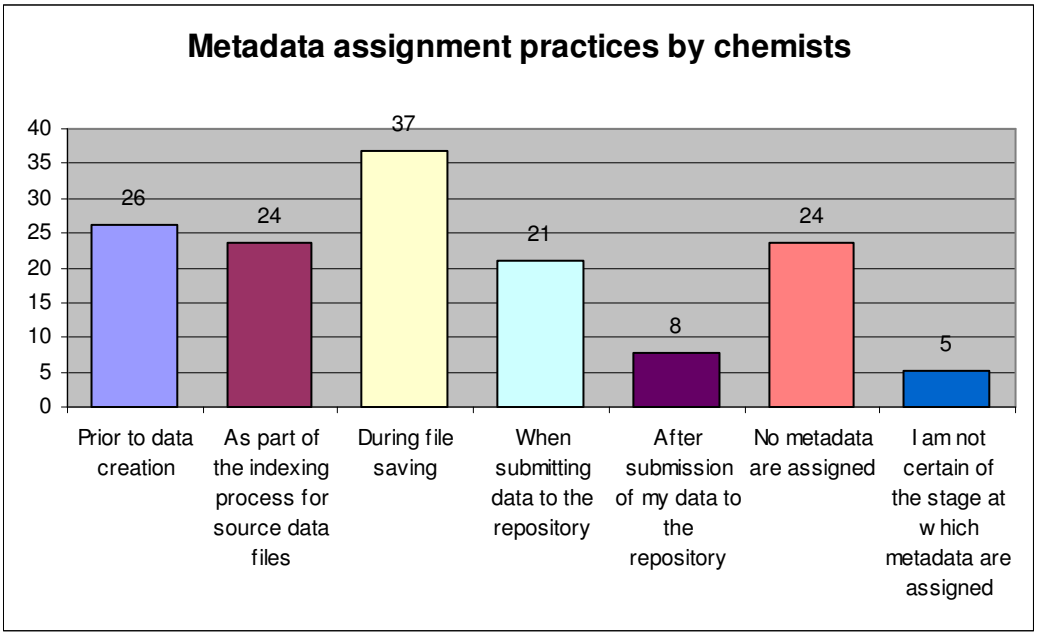


Figure 7: Metadata assignment practices for chemists

More than half of the chemistry respondents to the questionnaire survey (53%) noted that they themselves decide both on the terms to use and the assignment of metadata. Almost a third (29%) of the respondents replied that it was not known to them who assigns the metadata to their resources, which again supports the finding in the previous section, that showed a spread in the way chemistry respondents assigned metadata to their resources. The remainder of the responses were divided between those who replied that metadata is automatically generated (16%), metadata is assigned by research colleagues (11%), by research support staff (8%) and repository administrators (8%). One of the respondents noted that no one decides nor assigns metadata to their resources.

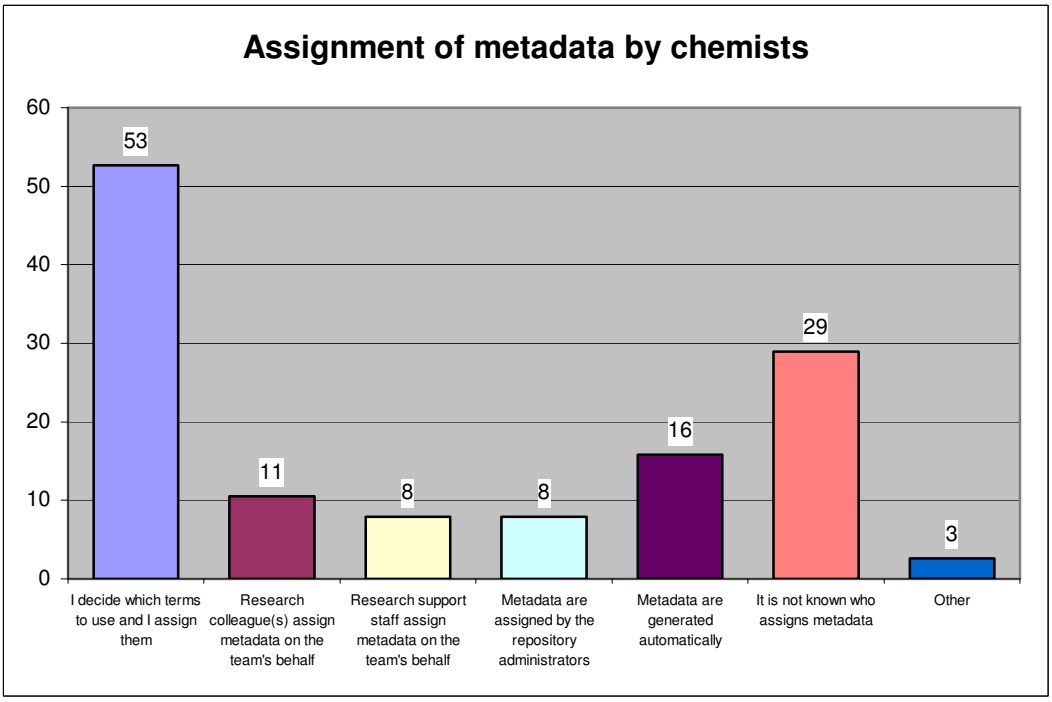


Figure 8: Level of metadata assignment by chemists

3.7 Usefulness of output repositories compared by users of named source repositories

The questionnaire respondents were invited to indicate what measures they normally use to control access to their data by other researchers. All respondents indicated that they employ a variety of means. The majority of the responses from academic staff indicated storage of their data on a private network/intranet (21%) as the main measure to control access. The same measure was indicated as the main one employed by the majority of postgraduate research students (32%) as well. All of the contract researchers noted that they use authentication of ID and password for online access for controlling access to their data. The research assistants replied that they tend to select storage of their data on standalone computers (16%) as the main measure for controlling who is able to access it.

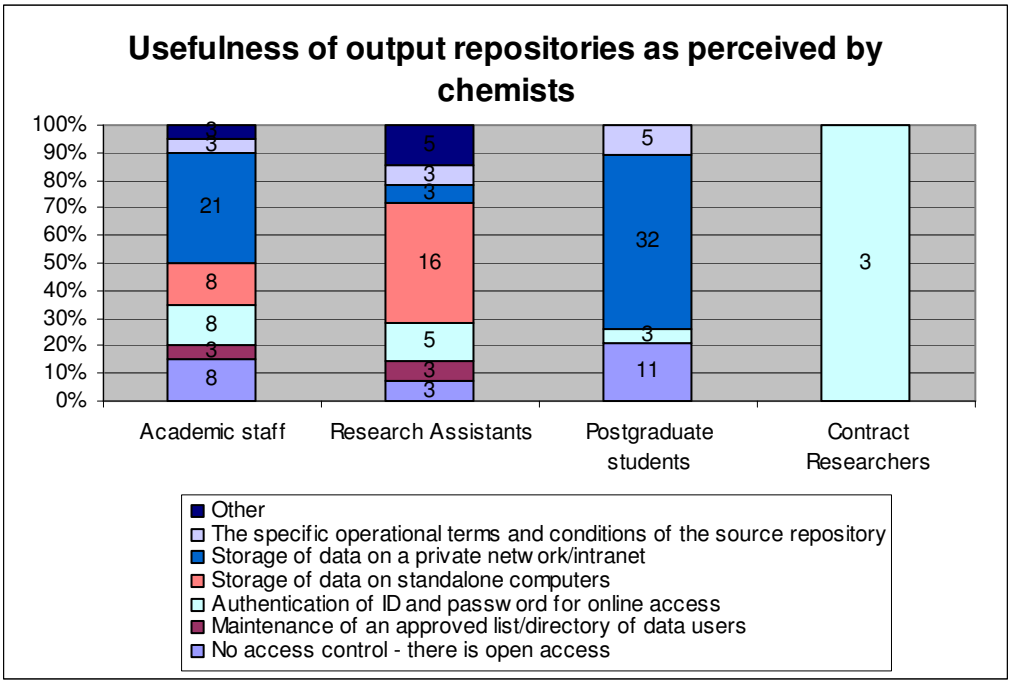


Figure 9: Usefulness of output repositories for chemists

The output repositories that the majority of chemists tend to use are those in the commercial sector set up and managed by publishers. Academic staff are the group who indicated they used the widest range of repositories, including institutional, discipline, publisher and others. Few of the academic staff replied that they do not use any repositories at all. Half of the postgraduate research students replied that they use publisher repositories for their research and the other half of the response was divided between institutional and discipline repositories. This is similar to the usage patterns indicated by the contract researchers as well. The research assistants also replied that they tend to use many different repositories such as institutional and publisher repositories and a few of them also noted that they do not use any repositories in particular. Results are presented in the following figure.

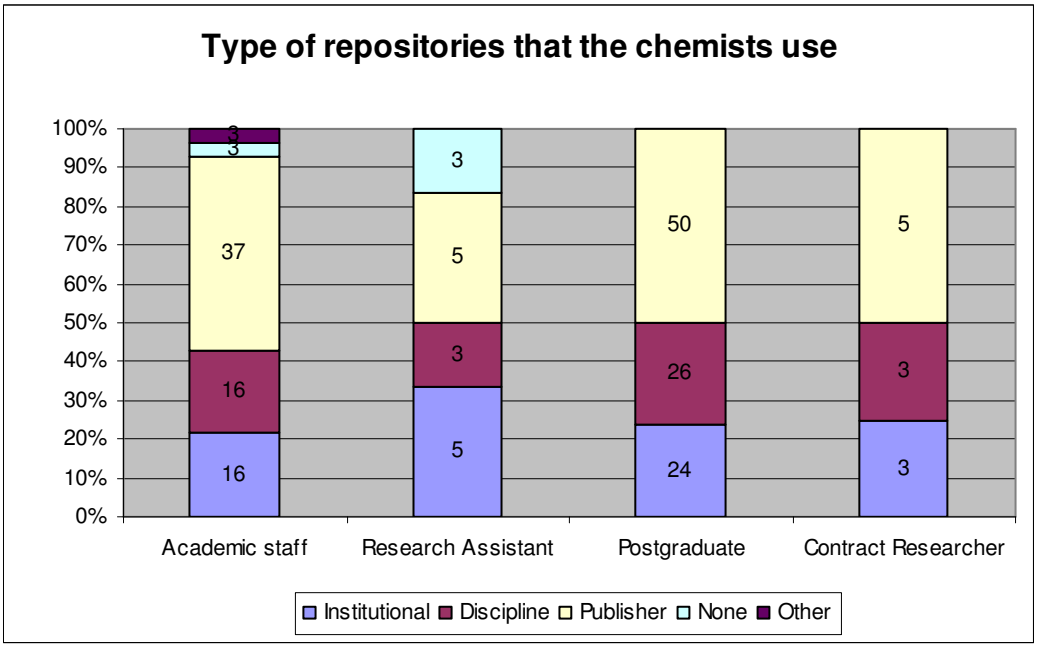


Figure 10: Type of output repositories that are used by chemists

3.8 The level of searching that is sufficient to researchers across different types of output repository and what enhancements might be considered

Although, the majority of the chemistry respondents to the questionnaire survey replied that they preferred to use the simple search option when they visited both source and output repositories, the response is quite spread again in relation to the different types of repositories. The majority of those who tend to use publishers repositories generally prefer to search them by employing simple methods. The use of subject specific thesauri and the use of Boolean logic are only mentioned in the searching of institutional and discipline repositories.

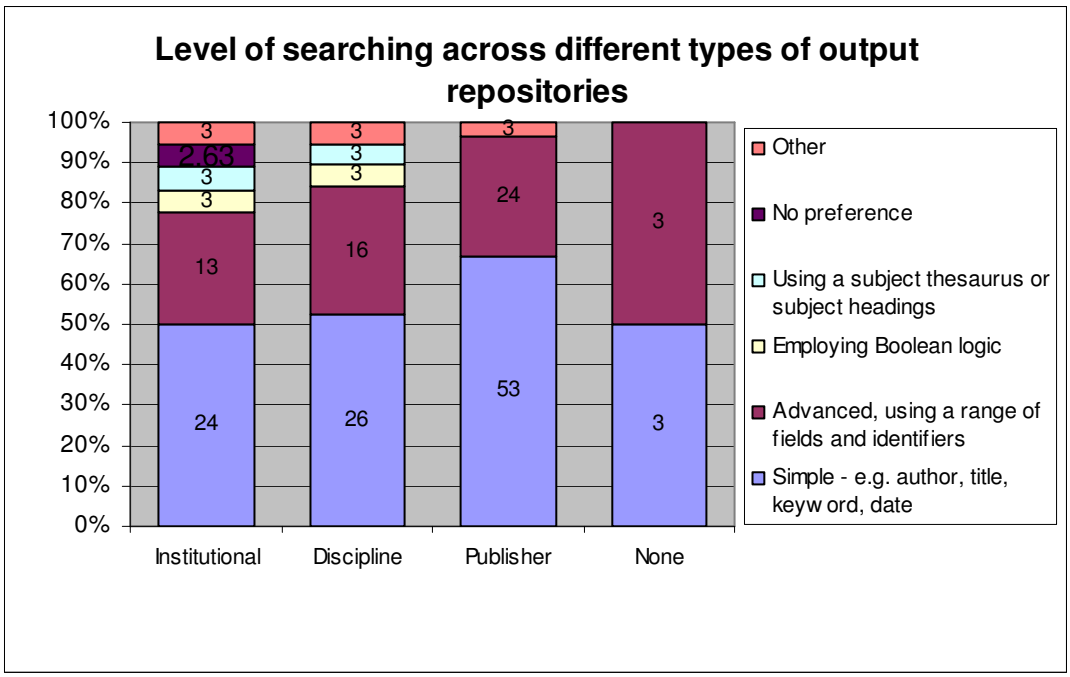


Figure 11: Level of searching that is sufficient for chemists across different types of output repositories

3.9 Preferred routes to output repositories compared by users of named source repositories

The respondents to the questionnaire survey were invited to indicate their preferred methods of accessing repositories. They were provided with a list of options that included access:

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

Half of the respondents replied that they preferred to search from an Internet search engine and from a publisher's online service. Other popular routes were via a library catalogue that links directly to an article in a repository (45%), directly through a specific journal's own web page (42%) and via a known repository's URL (39%). The least preferred routes were through an author's personal web page (18%) or via an Open URL resolver. 11% or respondents indicated that they do not have a preferred route for accessing repositories. Few of the respondents (5%) indicated "Other" means than the prescribed routes and they specified bibliographic services such as "Web of Knowledge or SciFinder, or stated that they had only recently started their research and as yet do not have any preferred routes to access repositories.

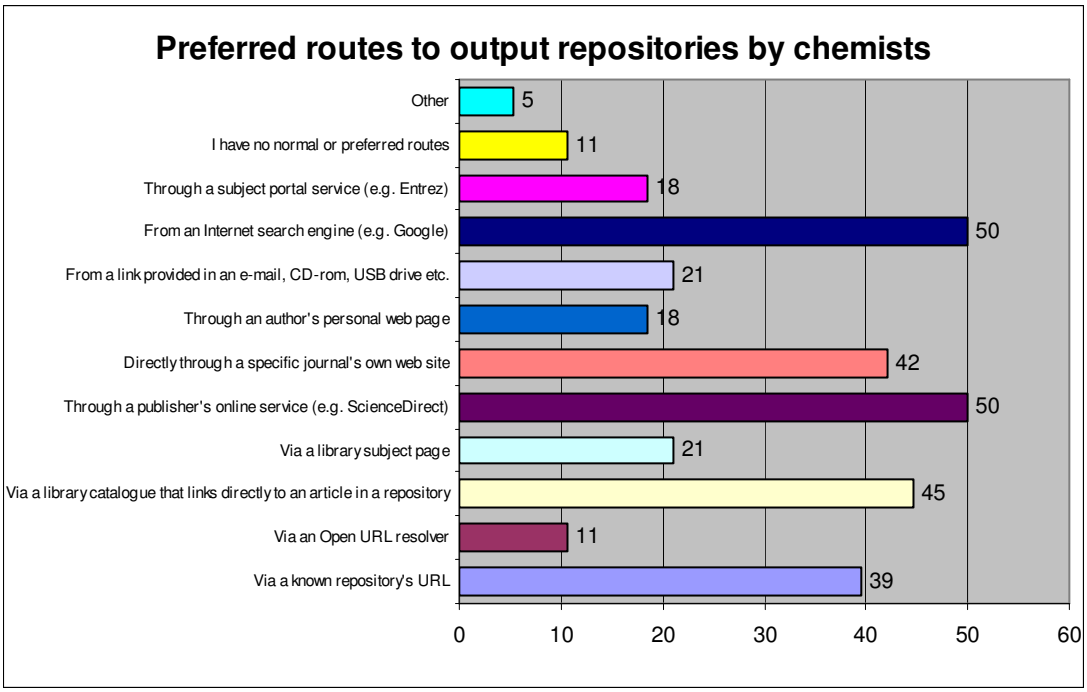


Figure 12: Preferred routes to output repositories by chemists

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

The respondents were asked if they receive support and/or guidance in their use of output repositories. They were also asked to indicate from a list of options that included documentary support, repository-enabled support, personal support provided by an intermediary, no support provided, unknown and other, those that were applicable to their case. The postdoctoral researchers and the research assistants that completed the survey did not know the level of support that was provided in their institutions or by the output repositories. The majority of the academic staff (13% of all respondents) and all the contract researchers indicated that there was repository-enabled support in their use of the repositories. The majority of the postgraduate research students (18% of all respondents) replied that they did not know the level of guidance/support that was provided. Analytically the response is presented in the following figure.

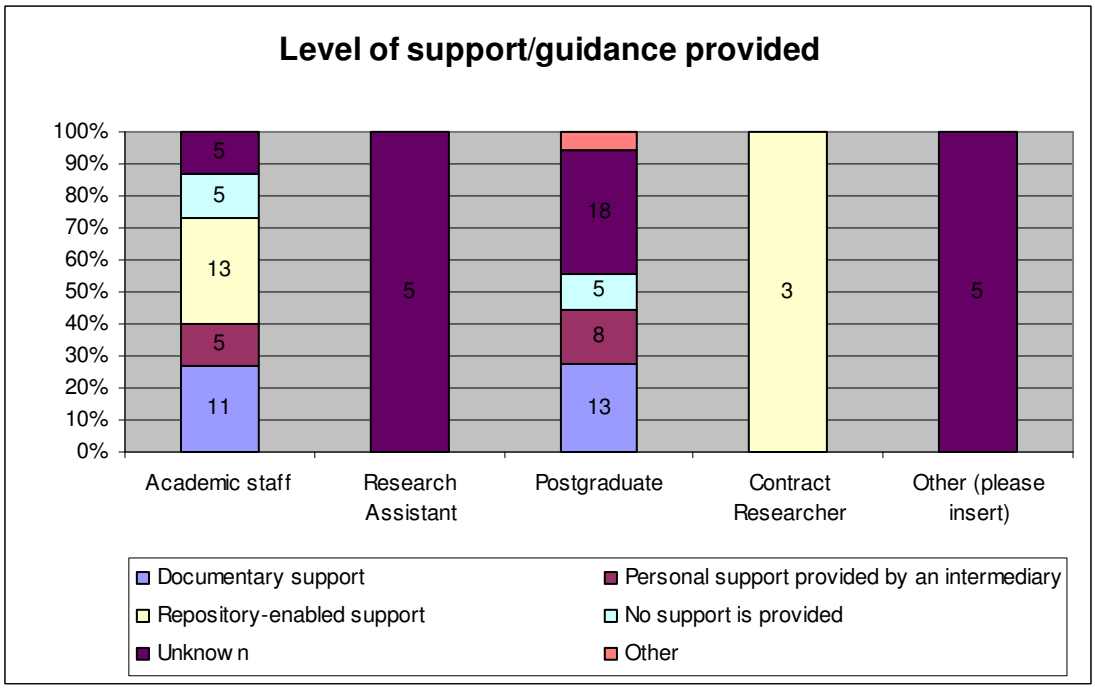


Figure 13: Level of support/guidance provided

Section 4 – Interviews analysis

4.1 Identities

Seventeen members of academic staff and postgraduate research students were interviewed. Almost half of the interviewees (47%) were academic staff at a very senior level within their disciplines. In particular four professors, two readers and two lecturers were interviewed. The remaining interviewees were either postdoctoral contract researchers (29%), or postgraduate research students (24%).

Role	**Number of interviewees**	**%**
Academic staff	8	**47.06**
Research Assistant	0	**0**
Postgraduate	4	**23.53**
Contract Researcher	5	**29.41**
Independent Researcher	0	**0**
Other (*please insert*)	0	**0**
Total	17	**100**

Table 6: Number of interviewees by role

The majority of the respondents represented two fields in chemistry; experimental/synthetic chemistry (18%) and theoretical/computational chemistry (59%). There was also one representative from each of the following areas: crystallography, biological, physical and polymer chemistry. All interviews were conducted face to face and lasted from 25 minutes to one hour, providing lengthy and insightful information. The number of the interviewees by field of interest is presented in the following table.

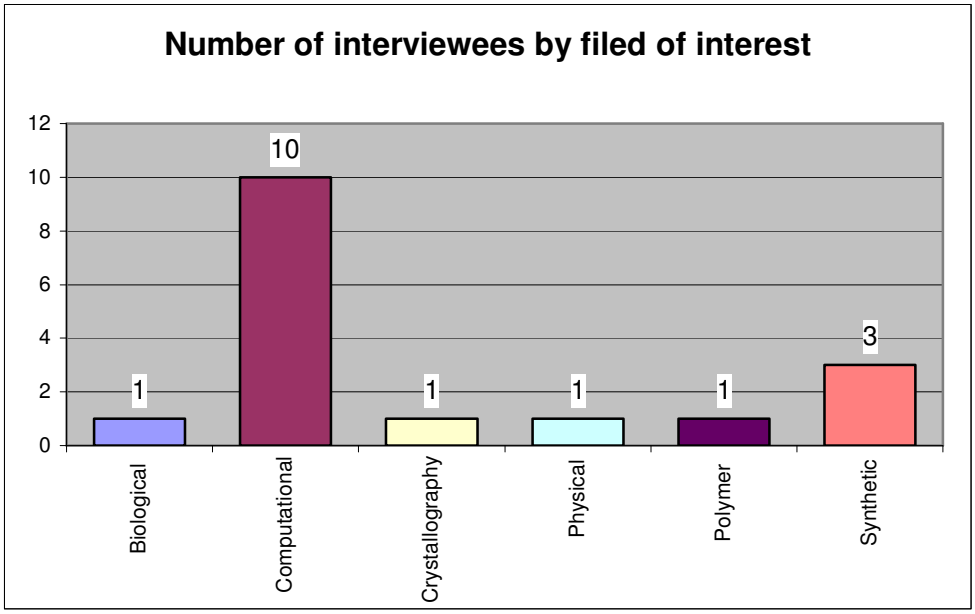


Figure 14: Number of interviewees by field of interest

4.2 Project aims

The bi-directional linking between source and output repositories has been acknowledged by the chemistry researchers as an interesting concept that could be useful for their research. One of the interviewees, representing the crystallography

community stressed the volume of research data that is getting lost in the process of publishing research results, in the following statement:

> *"…our standard cry [concerns] the long winded protocols that … are still required not only by the journals but by many scientists. It takes so long to write a paper and prepare all the data for it that we can not catch up with the updated data and only about 1/5 of the structures published in academic laboratories ever get into the public domain".*

In general, chemists were positioned favourably towards the idea of linking between research data and the published outcome of the research. Some of the interviewees were enthusiastic about the potential benefits that such a model could have on the way research is conducted. In particular they noted:

> *"That is my model. My model is that you are reading a paper on your computer screen. I want to link to it, that' s got to happen. There is no point in going digital and not being able to use the full power of the resource that we have. And if we have a growth of institutional repositories or national repositories and they are properly managed, I think it can revolutionise the way we are doing science. It will stop duplication, it will feed new data into scientific thinking, it will progress intellectual development, everything will go so much better..".*

There were concerns though, how such a service could complement or compete against existing publisher repositories that they provide access to this type of data already. Also, how it could change the way research data is stored at the moment and how it could be used to change the way research is conducted,

> *"… you create a database with all the metadata in and when you write your publications whether they are in open access [repositories] or within a journal – it doesn't matter, depends on your preference – you write a paper where you can discuss the use of the data, you can draw diagrams and graphs of what happens if you compare 50 entries, but you don't have to supplement the paper with all the data which is the traditional thing at the moment".*

Aspects such as the quality of data and mechanisms for quality assurance, the comprehensiveness of data and the ways it can be accessed, were raised as very important considerations if the use of source repositories of primary research data is to become more widespread. Other factors that it was felt would contribute to the success and/or the increased use of source and output repositories were the publicising of the potential benefits that such a service could have for the research community and the involvement of scientists in its development. Also important were issues regarding the sustainability, management and maintenance of both source and output repositories. Several of the comments that the chemists made are presented below:

> *"A lot of the data that is not getting published is because it's not of a publishable quality. So that's one thing. If I was to class a compound my biggest desire would be to have fully comprehensive worldwide sources of data. I wouldn't want to have to search three or four different databases if I can do it all in one. So that's why another group came up with the crystallographic database, because you can do all in one. There are some unpublished ones in there although the actual structures themselves would go through a review process, saying yes, they are of suitable quality to go [in a source repository]. So I think you have to have a similar thing there, but then in which case, why not just use them and say we send it to you. All it takes people to say if we are not publishing it,*

then we send it off there. To have another database that duplicates a lot of what is out there…there are so many databases to search anyway".

"Yes, I think [the quality of data] needs to be assessed. Yes, it has to be peer reviewed before it's available. I mean I know it's very different for people who are into astrophysics and the satellite downloads of their data and then that's raw data and then they get 6 months to do it and then they go on, and that's it, their target. And then if they haven't published in that time someone else will use it. It's not like that in chemistry because you've kept it and it's personal to you and the way you did it and the way you carried it out. So it's not like someone can pick and choose and whoever collects it, it's going to be the same".

"You have to convince a lot of people. Scientists are traditionally still in the mud. Scientists change the world but they don't like their world to be changed. We are so old-fashioned. The problem is that "I don't want my data to be exposed until I have a good look at it.
-"How long do you need for that?"
-"Oh, 3-4 years at least!!!"
They are imprisoned by their circumstances. And of course it will stop because if people sit on their data not writing papers, they are not going to make their names, they are not going to get the citations, the grants. It will be so determining, that's what I think anyway!"

"The difficulty is sustainability. For example, we have put up all our DSpace 250,000 molecules with their properties on open access. We can do it. The university has the DSpace project and they were happy to put this in. But it's not a sustainable model".

4.3 Source data

One of the prominent findings from the interviews was that chemistry researchers produce a variety of different types of data, often coming in large volumes. The majority of the interviewees were computational/theoretical chemists and they indicated that they do not produce raw data in the same sense that other scientists do. They are primarily involved in the development of methods that test how molecules behave in certain conditions; their research is about the development and testing of methods that measure the energy, the geometry and the particular arrangement of molecules under certain conditions. Therefore, although they produce data in the form of calculations and measurements' testing, they tend to apply their methods to other researchers' published outcomes.

The data is usually stored in their own computers or on shared network drives accessible by the project partners only. The file structure of the stored data for computational/theoretical chemists appears to adopt a tree structure with numerous files stored in individual subfolders. Access to this data outside of the context of the project and the individual file was thought to be not usable. For example, one of the interviewees noted:

"*First of all it would have to be everything associated with that compound. There is no point having an NMR without a picture of what it is. Then it's useful to have a synthesis scenario and say, oh that could fit with that, but I want proof and then that really is a paper. You know you can waste a lot of time trying to follow what people have done before that isn't properly published and never have any worth. It's not always, but is it worth the risk of wasting too much of your time?".*

The production of multiple types of data and data files that can be unusable outside of the context of the research project, or that their production enhances the quality and understanding of the research process, was also noted by an organometallic chemist who is primarily conducting experimental research. When they were asked to indicate some of the source data they produce they noted:

> *"…physical properties, experimentally derived. Physical properties like melting points of materials that we are looking at, lots of different SPECTRA and structural details, some x-ray like crystallography […] sometimes some computational work that goes with it".*

And another chemist graphically indicated that the outcome of their research is not always data in the computerised representation as it is now, but it could be in a different form:

> *"…but the actual outcome is actually like this (points at small containers with crystals) – STUFF! Real things, we make STUFF!"*

Other descriptions of the source data that is produced include:

> *"It usually comes to bits of paper; you usually submit it to a machine. You get the data back either in a sheet of paper from the guy who sends it to you or a file that you download from a central server. And then you can print it out, analyse it yourself. So you can have a hard copy and a print copy most of the time"*

> *"The main type of spectroscopy that you use to test everything to begin with is protein NMR so what you get back looks like a graph like this [starts drawing to a piece of paper], numbered 1-10 and it's a flat based line and you see signals. This isn't anything in particular but it looks something like this [points to the drawing] and from the pattern and the position of these signals you can ascertain if you have the correct molecule. So, it is quite pictorial because you can look at it as trying to elucidate the structure that you have actually made".*

When the interviewee was asked whether in order to get this graph they would have to enter some type of data into a computer, they replied:

> *"No, not really. Essentially, the whole basis of it is that the protons act like magnets and you are looking at magnetic resonance so basically you put in the spectrometer and [points at paper] this range of 1-10 represents the magnetic frequencies and then essentially you get these patterns back again. So what tends to happen is you get these pictures back and then what you tend to do when you publish your paper you publish the number 1-10, the integral, so that corresponds to how many protons and the pattern as well, so it shows if that's triples or duplets and that is how people tend to quote their data rather than publish in these graphs, because if you've got an organic paper that they made 20 different things, they cant publish 20 different graphs in a journal so they tend to shorten in out in numbers".*

4.4 Source repositories

As mentioned in the previous section, one of the most prominent findings from the interviews with the chemistry researchers is the breadth of the data that is produced and the diversity in the ways research is conducted within chemistry sub disciplines. The chemistry researchers were asked whether they had used any source repositories and to give examples of repositories that they had used to obtain useful information for their research. The computational/theoretical chemists that were interviewed indicated that that the data that they tend to work with could be obtained from various sources. They provided examples from their current research, which included the following:

- Obtaining data from the project partners such as commercial companies that funded the project. In this particular case a pharmaceutical company that had the data in their archives and which was then given to the researchers.
- Using established work models such as obtaining crystallographic data that is extracted from crystallographic journals in an automatic manner, on a monthly basis. In particular it was noted *"The best defined project that we have, is extracting crystallography from crystallographic journals so we are working with the International Union of Crystallography and every month or so they publish in Acta Crystallographica E a record of one paper per structure. You look at the paper and it will tell you the structure and also associate the data. We read that automatically so every month we have 300 research papers and we can analyse the data".*
- Developing the data themselves or searching for it in repositories, databases or the Internet. In one particular case, the example referred to the drawing of a molecule using software that the researcher had purchased, although software may also be developed for this research purpose. Alternatively researchers may use search engines such as Google Scholar to find and download a molecule to their desktop, or,
- Use data produced by chemical crystallographers, which is likely to be fundamental data that can be used in many of the sub disciplines within chemistry. For example, a chemical crystallographer described what a research process entailed *"Basically, we can take any pure compound that was made by the synthetic chemists or extracted or minerals or whatever. We use a single crystal technique so we have to have a single crystal, but then by using X-RAY diffraction we can create an image of the crystal structure of the atomic level. And crystal structure is easily represented by almost always alpha numeric file relating to parameters of the crystallographic unit. The atom types, the position of the atoms and the representation of the symmetry. And in fact crystal structures are probably the first scientific topic to be databased. The Cambridge Crystallographic database that started in the 1950s really has been building up ever since. So the data is still in a very rigorous way and can be used in a variety of other applications. You can download the results of the crystal structure, you can display a molecule, you can calculate interactions between atoms by bonding and non bonding and electrical interactions and you can feed the data into a chemical calculation and property prediction routines and so on".*

In most cases and due to the nature of the scholarly publication model in chemistry which requires the deposition of the primary data along with the submission of the paper to the journals, primary research data is available and accessible to researchers via the Cambridge Crystallographic Data Centre. Even if an institution does not subscribe to a particular commercial journal, supplementary data to a paper can be obtained, free of charge by the CCDC. Therefore, in theory, all chemists have access to a comprehensive source of primary research data that has been peer reviewed and quality assessed. The chemical crystallographer provided information regarding the source repositories that their research group have been using.

"*We use source repositories that are held at Daresbury like DTherm for example, a database of thermal data. We use NMR databases, again it is primary data, again, anything that you have as a compilation of data in a general standard format. So that it's easy to set up a search query. Then it's very useful. In many respects, these databases, if they exist, are isolated. And the linking, yes, this is something that we are trying to address in some respects in our project".*

Another aspect emphasised by the interviewees was the access to quality assured, primary research data that is supported by both a technological and social framework of research. In particular it was noted:

"[Primary research data] can come from the primary literature. The advantage of that is that it is peer reviewed and there is some measure of quality. Not always as good as it should be and so forth. The difficulty is that either many journals don't make it easy to archive data or they hide it away somewhere or put copyright restrictions on it and so on and so forth. There is a big problem in chemistry in getting the primary data out from the primary publication process, but that is one of the things that we are actively pursuing. So developing the technology but also the social consciousness of that, is a useful thing to have".

4.5 Metadata

The assignment and use of metadata as a minimum set of requirements that facilitate the deposition and retrieval of research data did not appear to be popular, or in some cases, even understood by the majority of chemistry researchers that were interviewed. It was perhaps surprising then, to see that those chemists that were familiar with metadata, did possess an in-depth knowledge of its use, applications and functions. The assignment of metadata automatically or by implementing a system that relieves the depositor of having to do it, was also commented upon positively. For example a computational / theoretical chemist described:

"Well, there's lots of different types of metadata. There is metadata for discovery, there is metadata for semantics, there is metadata for intellectual property and so on and so forth. They are all important. If I find some piece of information and it's not in open access then I can't use it. If I find some piece of metadata and it's in a language that my machine does not understand and there is no metadata, then it is uninterpretable, I can not use it. If I am particularly concerned about the quality of data I need provenance metadata. So there are different needs for different people..."

"No, we store a lot of metadata. Reports from the equipment, the settings of the equipment, the temperature, the detector, the conditions that show you can do the experiment. That's done automatically now and it's generated as part of our data workflow process, but when it comes to creating the archive and there will be a toolbox that the student will create on the job and then he will go through a set of guidelines and go through the process and then, will create an entry for the archive. It will be checked by a senior member of the group and then it will go to the archive. In that process, a certain selection of metadata is put onto the front page and these are also searchable items. Everything is done automatically".

The level of metadata assignment and the preference for particular metadata standards was discussed by some of the interviewees. Although the Dublin Core metadata standard was indicated as the de facto standard that supports the minimum set of requirements for both the management and retrieval of information, other information requirements were also noted. These included:

> *"We would always put certain features, certain details of the crystal structure data space group, authors, institution, quality indicators, colour of the crystal, keywords, what kind of compound that is, organic or inorganic, organometallic compound…All the kind of things that we think people may want to search on".*

> "*I think there is far too much attempt to try and classify things by rigid schemes. I believe in free text classification. And you need metadata for things like dates, and owners and things like that. To say that something is organic chemistry, it's a 19th century way of doing things"*

One of the aspects that the interviewees commented upon was that there should be a wider organisational/institutional requirement that supports and manages the repositories, should they be source, output or institutional.

> "…*sustainability depends on a business model. And it's a major problem that confronts everybody at the moment in aggregating data, whether it would be raw data, processed data, metadata, primary publications, abstracts, things like that".*

4.6 Data access and sharing

The majority of the chemistry researchers noted that they were reluctant to share access to their data whilst they were still in the process of conducting their research. Exceptions existed in the form of personal communications and contacts. For example, if someone knew another researcher personally and they were contacted on a direct basis they would be inclined to provide access to research data. The majority of the interviewees replied that it is not common to share access to their data with anyone outside the research project, while they are still working on their project. Reasons such as infringement of their work, early publication of their research and misinterpretation of research outcomes were cited as those that would make researchers reluctant to share access to their research. This situation changes once the research is complete, and all interviewees were happy to publish and share access to their data once the project had finished and the research outcomes have been published.

Regarding copyright, the majority of the chemists were not happy to sign it away although one person noted that they are happy to sign away the copyright, as they felt that they gained in prestige from being able to publish their research in high standard journals and being read by their fellow researchers.

The other chemistry researchers expressed their reluctance to sign away the copyright of their research. The majority of the computational/theoretical chemists noted that they never sign away the copyright for their research and one in particular said that should they have the opportunity, they would rather licence the copyright to their work. Although the reluctance to sign away the copyright was more common among the researchers that were in a senior position within their disciplines, the Ph.D. students noted that they would be happy to trust their supervisor's judgement on this issue. One of the more senior academic staff noted:

"I never sign away the copyright for my data. I think that is something that we keep. This stems from my realisation, and other people's realisation, that many of the commercial journals make millions just by handling data".

Regarding data sharing, there was a common reason among the chemists, for not being willing to share their data while they were still working on their project. They were happy to make their data available after they had published their research results but not before. As was shown in the questionnaire survey, personal communication is an important factor in the chemistry research community, for the exchange of data and discussions concerning the progress of projects. This is shown in the interviews as well. In particular it was noted:

"Well, yes [it would be a barrier]. While you are still working on it. Yes, that is a serious problem. Because a lot of chemical research, unlike some other areas, it is possible to make interesting, useful chemical research very quickly. You know, within a week, sometimes that's possible. It's not like big physics or big bio projects where you have 20 people working on something that takes ages to do. It would be possible for one of my students, supposing we picked up this information about making this compound, for the student to go into the lab today, make this compound and then do what we wanted to do with it tomorrow and have the result by Friday. I mean that is not common, but it is possible. And what this chap emailed me back, very recently, I am not going to give you the details, it wasn't in a nasty way, but because I have been scooped before. So, he said, oh you will have to wait for the publication.

Yes, if I meet the people. Within your research group there is not too much competition. You go and talk to people".

Reasons that would make chemistry researchers hesitate in sharing access to their research data were noted as concerns about the quality and validity of the data and its comprehensiveness outside the context of the research project.

"If I publish work on that data and I am not going to publish any more on it, then fine. I am happy to make it publicly available. But no, not on general open access while I am still working on it, for two reasons. One is because it hasn't been verified and secondly a lot of it doesn't make any sense without writing about it. You can have a SPECTRA there, but unless you say what you did and what you are trying to do [it is of no use]. If I use a mix of two or three different things, then do the SPECTRA and see how they reacted, is that reaction of what I did, associated with that SPECTRA and stuff, and the actual techniques you used, and then if I am going to write it all up, then I would have published it".

Another member of academic staff differentiated between access to data and copyright of the research work conducted.

"I believe that all facts are not copyrightable and I believe that all publishers should make factual information publicly available without copyright restrictions. Now, copyright and access are two different things but they are often conflated. Many publishers do not provide access to the factual data, some provide access and it's their copyright. I would say that both of these are undesirable…and then…about copyright of full text, this is one of the worst things that the scientific community ever did in the early 1970s, to sign away their copyright. Nobody else does it. J K Rowling doesn't sign away the copyright to the publisher, she retains it. So, it's the creative work of the author. There is absolutely no reason why

it should be owned by the publisher. What I genuinely do when I get the copyright form, I write not applicable or I write not exclusive or something like that over it and they don't have enough time to argue".

4.7 Output repositories

Results from the face to face interviews with the chemistry staff and students revealed that Internet based services such as the Google Scholar are amongst the most popular means to look for information. Theoretical/computational chemists tend to use many Internet based repositories and services depending on the type of data that they are looking for. The speed of information retrieval, the quality of the information and the comprehensiveness of the repository are important factors in their choice of output repositories. For example, interviewees noted:

"…If I want to find out if someone has done something already then I tend to use Web tools first of all. They could be Google Scholar, they could be Google, Wikipedia, it could be anything. And normally they would come up with a hit pretty quickly. So, myself, I never use Web of Science or any of these things. If I want to find a chemical molecule then it's much more difficult because most of the chemical molecule information at the moment is managed by two major chemical information suppliers, Chemical Abstracts and Beilstein, and they are commercial..."

"Chemical Abstracts for the chemist is THE answer, the comprehensive answer. And it's pretty good and it's pretty easy to search. There are some other databases. Now, the organic chemists use something called Beilstein, which contains a lot of the physical details, the sort of thing you are talking about. So, that's used quite a bit. The reason why we don't use it very much is because it's very organic and that's less interesting to us and it's also not comprehensive. So, sometimes students say "Oh, I've looked on Beilstein and it's not there". To me that's not good enough because Beistein is not comprehensive. It doesn't try to be comprehensive. So you might as well start with CA as far as I am concerned. The other main database that we do use is one called the Cambridge Crystallographic Database. Have you come across this? So, that's run as an online database. I mean CA does exist physically, in the library, in thousands of volumes on the shelves. The Cambridge Crystallographic Database never existed physically, I think. And that allows you to search, that's for crystallographic structure data. So it is quite specific, it's big and it allows you to extract 3 dimentional pictorial data and that is based on the sort of thing that you are interested in. That's made by taking the fundamental crystallographic data as a computer file originally, you submit it as a computer file to Cambridge and they check it and they are fussy! They don't just take whatever you give them. They say, you know, they will have a look at it and they will say, this makes sense, the data you give, the drawings, yes".

Most of the output repositories in chemistry have a structural search feature, which is very popular amongst chemists. In the structural search, the researchers draw what is called a "picture" of the molecules that they are interested in and the software looks for identical structures or molecules with similar structures. This search option is preferred in source repositories such as the Cambridge Structural Database. The simple search is preferred in the output repositories. Interviewees commented upon the search options:

"…*you can search in lots of different ways. We often do it for a particular sub structure. Very useful, being able to draw little bit of a molecule. I can draw in that fragment there and I can say, are there any other compounds*

with that chunk in? They might be huge and contain that little bit, or it could be that little bit just with a couple of extra atoms on. You actually draw that on screen. You draw the little chemical structure on the screen and you say are there any other compounds like that? I don't know what compound I am looking for but it should have that little bit in it. And it will tell you if there are 5000 of them. So, that's really powerful as a chemist because it's a pictorial representation of what you want"

"Sometimes you can search by author, sometimes by keywords, for example, are there any papers about hydrazine imidazole derivatives? That's a powerful database. The Cambridge database you usually do by structural search. You can draw a little structure and say are there any compounds like this in the database? You can do it by author. Actually the Cambridge database…is very compound based. They all have a little, six letter code. So, sometimes you see them in publications. And you say, that's the quick way to get them. Oh, I want this code compound, they are unique. So, yeah, it's quite easy to look".

"There are a couple of good databases that do that. There is one called Beilstein and there is one called SciFinder. So you can look up their structures and they will come up with papers that have been published before. And they seem to be reasonably good. If you search in both of them, they pretty much cover everything that has been published".

4.8 Support

The majority of the interviewees did not have any former training for using support services in either source or output repositories and they tend to rely on personal communication and interactions in their research groups when they need to ask for information and/or help with a search. In libraries where the presence of librarians is strong in the academic departments, chemists noted that they would contact the liaison librarian, should they require any assistance. Observation or consultations with fellow research students, in the case of the Ph.D students, were also noted. The different approaches that chemists tend to adopt in the use of support services are presented below:

"Yeah, [person's name omitted] in the library would be somebody to ask. I might ask a colleague. It's hard to say because I am quite experienced with most of the things I use, so it doesn't happen very much but yes, you might visit a FAQ section or an online help guide. However, because we have a research group here, you may, if [person's name omitted] is not available, I may jump in the lab and say has anybody tried this".

"[I don't use any] Not that much. I think first when I was learning how to use them I think I may have done that, I can't really remember. There are more advanced, sort of like data mining things, that they do offer but I haven't used it so much".

"Not really. I tend to ask other people, someone in the group that has been around for a couple of years and they have experience in using them and you will sit down and show them to you. Or you can sit down and see how other people do it. No formal library experience or training when you come into the group".

Appendix A - Section 5

5.1 Reasons why chemists might want to access the research data generated by other research programmes

The questionnaire respondents were provided with several options to choose from regarding the reasons why they might be interested in accessing other research programmes' data. The primary reason noted by academic staff and research assistants was "to access data that are useful or necessary for my research". The postgraduate students and the postdoctoral researchers noted the reason "to understand the broader context and orientation of my research". The identification of useful contacts for their research was more important to the postgraduate students compared to academic staff and contract researchers. Also, testing the validity and uniqueness of the research objectives was more important to research assistants than any other group.

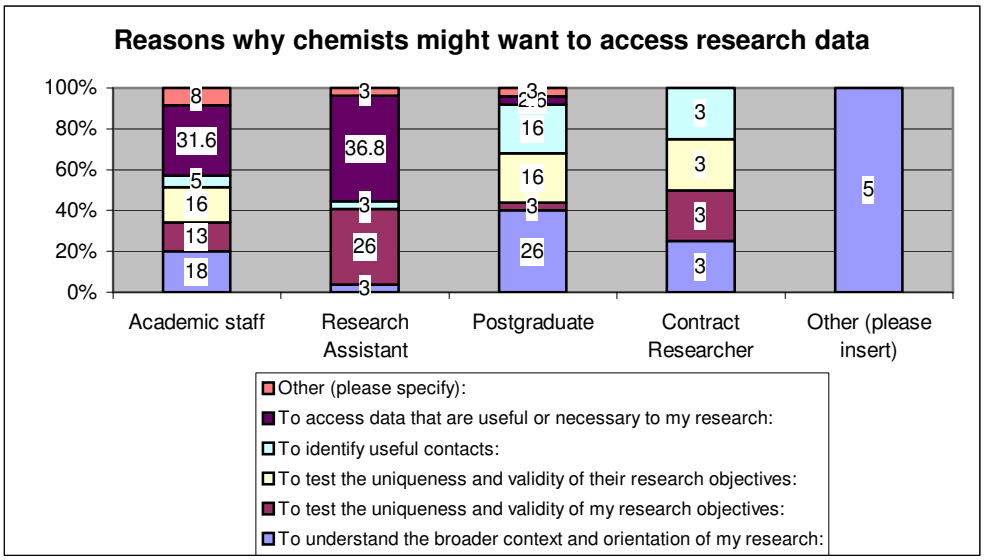


Figure 15: Reasons why chemists might want to access research data

5.2 How would you normally access the research data of others

The questionnaire respondents were invited to select from a range of options to indicate how they would normally access the data of other researchers. The options were: through online access to source repositories; by access to networked file servers at other institutions; by access to networked fileservers at my own institution; through the exchange of data held on portable media, or, I do not normally access others' research data. Surprisingly the majority of the respondents in all groups noted that they access other researchers' data through access to source repositories. It is believed that the respondents might have confused the term source repository with repositories in general as this finding is in conflict with the initial observation that more than half of the respondents (65%) had not used a repository in the past and in general they were not familiar with the idea of open access.

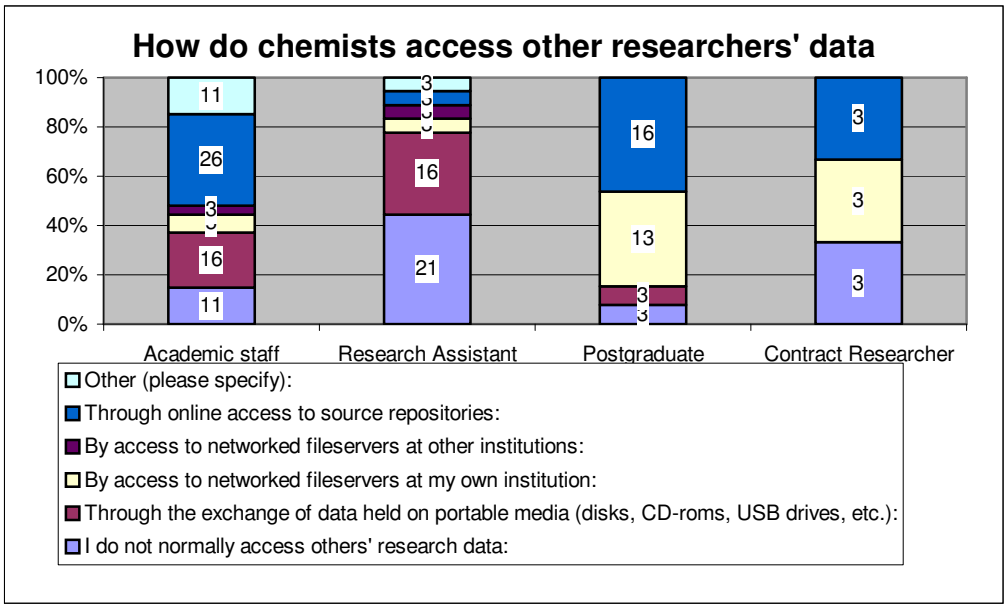


Figure 16: How chemists access other researchers' data

5.3 What factors would encourage you to share your data

The chemistry researchers were invited to note the factors that would encourage them to share their research data. The response was spread and there was not one single factor that appeared to be significantly important. For academic staff, the potential benefits to the research community appear to be the most appealing factor (34%) followed closely by the demonstrable benefit for their research profile (32%). Similarly, this was the most important factor for postgraduate research students (34%). It was the research assistants (21%) and the contract researchers (5%) that noted the requirement of a funding body/condition of funding as the primer factor that would encourage them to share their data.

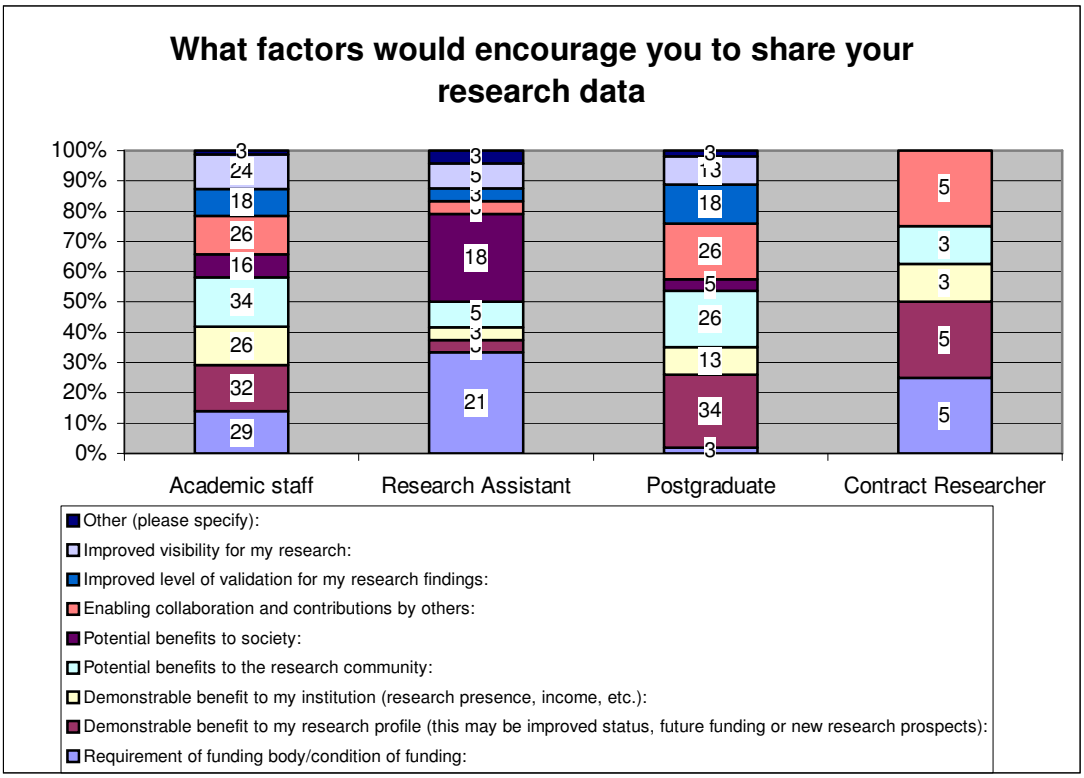


Figure 17: Factors that would encourage chemists to share their data

5.4 Factors that would discourage the sharing of data

The chemistry researchers were invited to indicate the factors that would discourage them from sharing their research data. As with the previous questions, there is not one single answer there was not a single predominant view expressed for each group. In general, the threat of the loss of ownership was a strong factor for postgraduate research students (37%) along with the risk of premature broadcast of research findings (37%). Academic staff noted the time/effort required to enable sharing (32%) and the risk of premature broadcast of their research findings (29%) as the most important reasons. The most important factor for research assistants was the risks to an established research niche (24%).

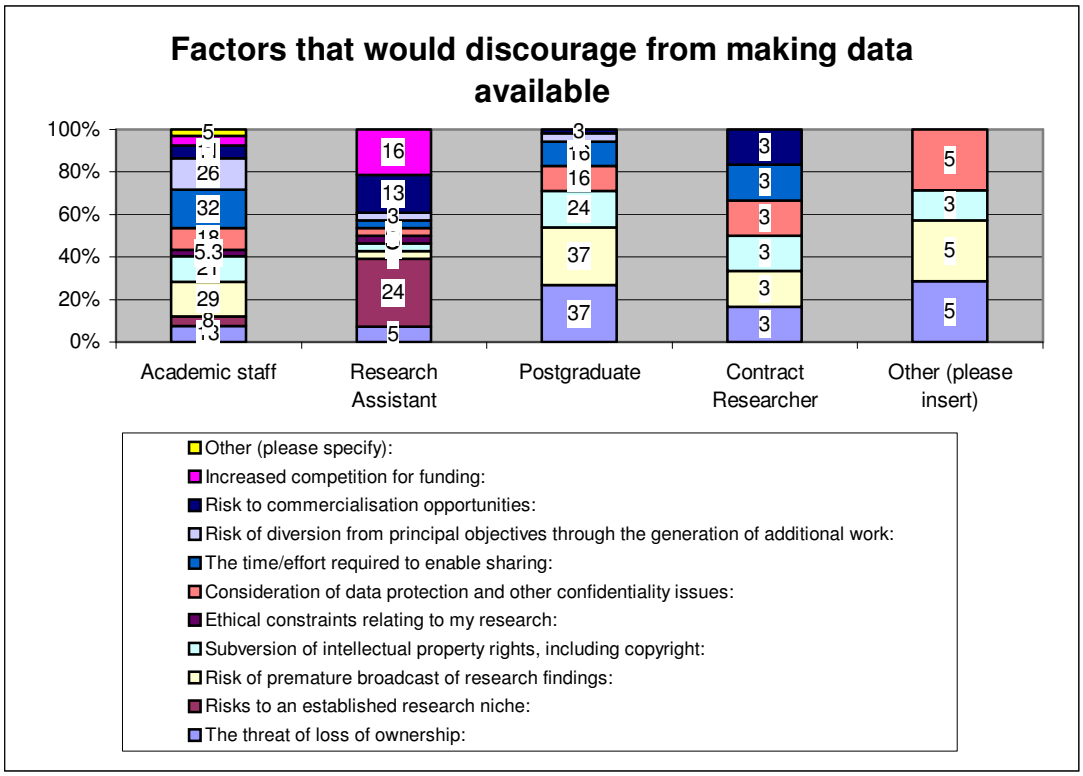


Figure 18: Factors that would discourage chemists to share their data

5.5 Kind of formal restrictions that chemists apply

Respondents to the questionnaire survey were invited to select from a list of restrictions that they apply to their research data. Academic staff (24%) and postgraduate research students (21%) replied that they do not apply any formal restrictions to their data. The research assistants however, noted that there are restrictions imposed for the research team and members (18%). A restriction noted by respondents to all groups was that they respond to individual enquiries/requests for access and they judge them based on their merits. The academic staff were the only group that noted other restrictions. In particular they specified: creative science commons, ownership retained; request acknowledgement on re-use; varies with maturity of project; to respect collaborators requirements. It depends on the programme and sponsor which factor may apply in any particular case.

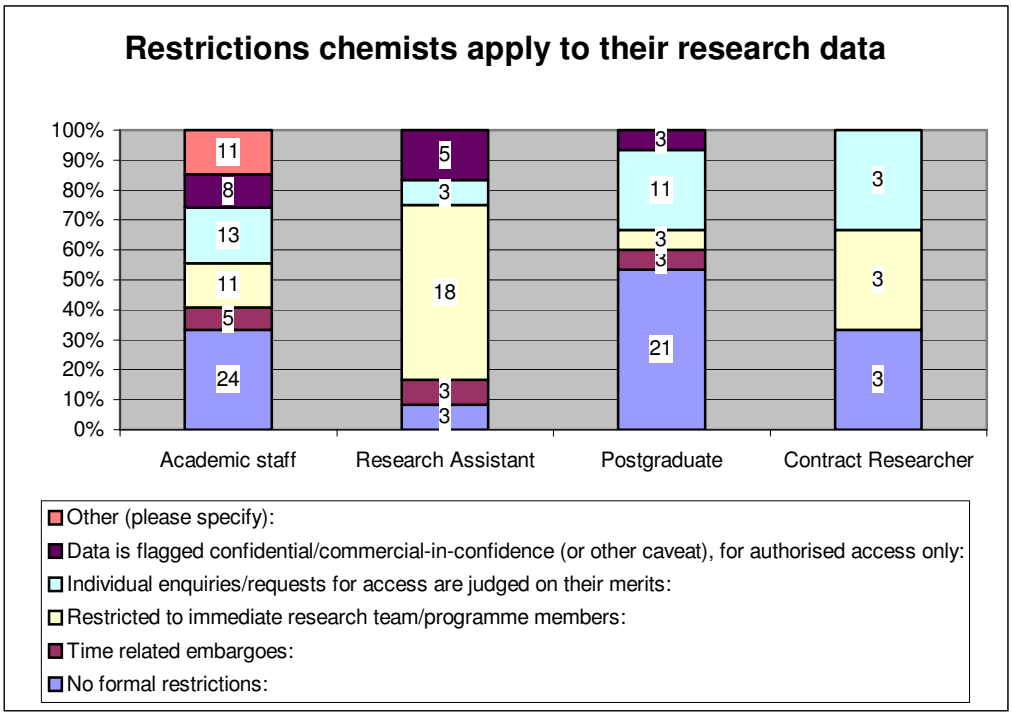


Figure 19: Types of formal restrictions that chemists apply to the access of their data

5.6 Output repositories where chemists submit their research data

The majority of the respondents indicated publishers' repositories as the main output repositories where they submit their data. Academic staff (18%) and the research assistants (3%) were the two groups that indicated they do not submit their data in any type of repository, an opinion probably based on their understanding of a repository as distinct from electronic access to journal publications.

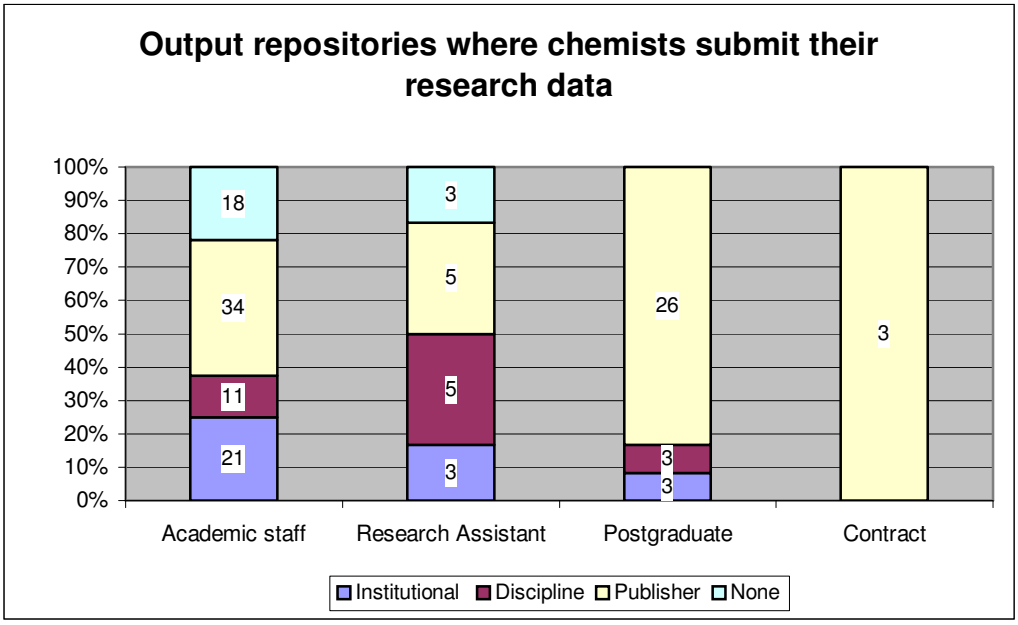


Figure 20: Output repositories where chemists submit their research data

Appendix B - Section 6

Free text from questionnaire edited to reflect :

6.1 Perceived value of source-output repository links relative to different research roles

NONE

6.2 Perceived value of source-output repository links relative to different repository communities

NONE

6.3 The different types of source data generated/held according to different repository communities

Those who indicated other types of data they specified as “mainly binary and text files from calculations with figures and graphs derived from these”.

Other suggested formats included a variety of standards and software associated with the production and description of data in the chemistry research community such as:

- .cif (crystallographic data),
- binary data files,
- chemdraw
- cdx. xwin nmr files,
- Chemdraw Word,
- Chemical Markup Language,
- corel draw,
- Fourier induction decay files (generated from Bruker and Varian NMR instruments),
- Spectra are in spectrometer specific code.

6.4 Metadata requirements according to different repository communities

Chemical identifier (InChi)
None

6.5 Metadata assignment practices relative to the level of support provided in the use of repositories

Query1

10 "At what stage are metadata assigned to your data?"	25 "What assistance in your use of repositories is provided by
As part of the indexing process for source data files	Only recently started my research. Unknown
Calculations typically contain creator name and short description. Summaries are generated by whoever ran the calculations to analyse their own results.	Assistance is available, but with the combination of web of knowledge / SFX / Digital Object Identifier working almost transparently, I've not used it - unless the combination doesn't work for a technical journal-specific reason, when I've e-mailed for support.

6.6 Usefulness of output repositories compared by users of named source repositories

Query1	
18 "What measures do you "'normally use"' to control access	**8e National Crystallography Service**
For that which is not deposited in central stores access is currently by request	On several occasions
Maintenance of an approved list/directory of data users	Never, but I am intending to do so soon
Since it is not available I would expect others to email a request	Never
storage on paper (e.g. lab-books),The specific operational terms and conditions of the source repository. Storage of data on standalone computers	Frequently

6.7 The level of searching that is sufficient to researchers across different types of output repository and what enhancements might be considered

Query1	
19 "Which kind of output repositories do you use to find and r	**23 "What level of searching do you normally find sufficient wh**
I usually go through our library web-site to the online version of journals - does this count? Institutional.	Simple - e.g. author, title, keyword, date
None	Simple - e.g. author, title, keyword, date
Only recently started my research. None.	No preference

Query1	
19 "Which kind of output repositories do you use to find and r	**23a "What further options, features or functionality would en**
Institutional.	Chemical substructure search
Only recently started my research. None.	
I usually go through our library web-site to the online version of journals - does this count? Institutional.	

6.8 Preferred routes to output repositories compared by users of named source repositories

Web of Knowledge (wok.mimas.ac.uk)
SciFinder

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

NONE

Appendix C – Section 7

7.1 Interviews structure

The following section includes the range of the questions that the interviewees were asked. Depending on the flow of the conversation some questions may have been omitted in some interviews and/or all questions were asked in other interviews.

7.1.1 Identities

- Role (e.g., academic staff, Ph.D. student, post doctoral researcher, etc.)
- Field of interest (based on the RAE 2001 list of options for chemistry)
- Research experience (years of research and post held, e.g. lecturer, reader, professor, etc.)

7.1.2 Research process

- Could you please describe what would be **a typical process for a research project** for you?
- In particular, how would **you access and share information** throughout and after that process?
- As well as sharing the results of your research through publication, do you currently **share access to your data**?

7.1.3 Source data

- What **types of electronic data** do you produce?
- In what **formats** is it held?
- Is the data you generate sometimes a **combination or group of different data formats**?
- Where is this data **kept/stored**?
- **Intentional use of research data** – why and how they would use data obtained from source repositories or elsewhere.
- How would you **normally access the research data** of other researchers?

7.1.4 Project aims

- Introduction to the StORe project and explanation of what source and output repositories are.
- If those repositories included a standard feature to **link to the published outcome** of those data would that be advantageous for you?
- Similarly, if repositories included a standard feature to **link from primary research data to the published outcome** of the research would that be advantageous for you?
- Do you see any **benefits** from such a feature? What would be a **desirable feature** in a system that facilitates bi directional links?

- Could you please name three features that you consider essential in any system?
- Could you please name three features that are considered overstated?

7.1.5 Source repositories

- Have you used any source repositories? (e.g., Cambridge crystallography service. National Crystallography service, etc.)
- What is the main type of information that you look for in source repositories?
- What output repositories have you used?
- How important is it to you that your research data is accessible to others?

7.1.6 Output repositories

- Explain what an output repository is.
- Have you **used** any output repositories?
- What is the **main type of information that you look for** in output repositories?
- **Which output repositories** have you used to find and retrieve information?
- How important is it to you that your **research data is accessible** to others?
- How do you choose where **to publish the different outcomes** of a research project?
- How much do you publish in **online journals**?
- In which **output repositories** do you **deposit** your research publications?
- What are your **normal or preferred routes** to the **contents** of output repositories?
- What **level of searching** do you normally find sufficient when using an output repository?
- What further **options, features or functionality** would enhance your level of searching?
- To what **extent** do you **publish** in publications which are available **online**? Is this a prerequisite for you?

7.1.7 Metadata

- Explain term metadata and for what purposes it is used.
- **How** do you describe/characterise/name your data?
- **Who assigns metadata** to your research data?
- **At what stage** is metadata assigned to your data?
- Do you use **different types of metadata** for describing and different for retrieving information?
- How would you like to be able to **search** in a repository?

7.1.8 Data access and sharing

- What **measures** do you use to make your research data available?
- What **factors would '"encourage"'** you to share your research data?
- What **factors would '"discourage"'** you from sharing your research data?
- Normally, what kind of **formal restrictions '"do you apply"'** to your research data?
- What measures do you '"normally use"' to **control access to your data** by others?

- How do you feel about the **copyright** issues sometimes inherent in such a process?
- Do you feel that allowing other researchers to **link** from your online publications to the data which lies behind such research would **enhance** the general **quality of research** in your field in the future?
- How much would you be willing to make **available on open access**, and at what stage?
- To what extent do you already **allow** others to **access your data**?

7.1.9 Support

- How would you **usually search** for information in a repository – what are the options that you would search by? Keywords, compounds, name of author, etc.
- Which of the **support features** of output repositories do you tend to **use** more often? (e.g., FAQs, HELP, personalised settings, alerts)
- Have you come across any **features that you would wish to see** more often in all repositories?
- Do you find there is sufficient information **support available** to help in **data deposition**?