Use and linkage of source and output repositories and the expectations of the chemistry research community about their use

Panayiota Polydoratou

Imperial College London
South Kensington Campus
London SW7 2AZ, UK
p.polydoratou@imperial.ac.uk

Abstract. This paper presents findings from a questionnaire survey that aimed to identify the issues around the use and linkage of source and output repositories and the expectations of the chemistry research community about their use. In the context of the StORe project (http://jiscstore.jot.com/WikiHome), which sought to develop new ways of linking academic publications with repositories of research data, thirty eight (38) members of academic and research staff from institutions across the UK provided valuable feedback regarding the nature of the research that they conduct, the type of data that they produce, the sharing and availability of research data and the use and expectations of source and output repositories.

1 Introduction

The StORe: Source-to-Output Repositories project (http://jiscstore.jot.com/WikiHome), funded by the Joint Information Systems Committee (JISC, http://www.jisc.ac.uk/) is a collaboration between seven universities across the UK and the Johns Hopkins University in the USA, who are focusing on seven disciplines, including chemistry. The project sought to develop new ways of linking academic publications to repositories of research data. One of the project deliverables are published surveys of researchers that identify workflows and norms in the use of source and output repositories, including common attributes across disciplines, the functional enhancements to repositories that are considered to be desirable and perceived problems in the use of repositories.[1]

This paper presents findings from an online questionnaire survey that aimed to identify the issues around the use and linkage of source and output repositories and the expectations of the chemistry research community about their use and is relevant to the above mentioned deliverable of the StORe project. The respondents to the questionnaire survey provided feedback about how useful or not they considered the linking of research data to publications, the types of data they produced and the

[1] StORe project description. Available at: http://jiscstore.jot.com/WikiHome (Last accessed 04/09/2006)

formats they are saved, the level of metadata that is considered important and the assignment of metadata. Furthermore they indicated perceived barriers and advantages in the sharing and access of their data and their preferred routes of searching at output repositories.

2 Definitions

Several terms were used in the questionnaire survey and throughout this paper. They are defined as:

- ***Repository***. A repository is a store where electronic data, databases or digital files have been deposited, usually with the intention of enabling their access or distribution over a network.
- ***Source repository.*** A database that contains primary research data on which a publication will eventually be based.
- ***Output repository***. A database that contains research publications, the published outcome of the research. Output repositories can function at an institutional, regional, or global level. They maybe organized accordingly to publication type (theses, working papers, post prints, etc.). They may include the commercial repositories maintained by publishers, since it can be argued that online journal services such as ScienceDirect qualify as output repositories.[2]

3 Literature Review

"*The most comprehensive and reliable source of chemical and physical property data is the chemistry literature. In many cases a literature search may be the best option for finding this type of data"* (NIST Data Gateway, http://webbook.nist.gov/chemistry/faq.html). Several studies had identified the extensive use that chemists make of the literature and as a research community they are considered to be those with the highest reading rates among scientists (Tenopir & King, 2002). Chemistry is a science with a long and established history. Some of the characteristics of chemistry research include its interdisciplinary nature, the production of vast amounts of data that lead to a comprehensive literature in which the relevance of the articles does not decline over the years, and the use of information technology to conduct research.

Despite this, there is a generalization that has followed chemists over the years, citing that in general they are reluctant in using information systems. More than 10 years ago, Philip and Cunningham in a British Library Research and Development study, surveyed chemists across the UK to find out about the availability and the use of automated chemical information systems. The study found that more than half of the respondents who did not make use of chemical information systems claimed this

[2] StORe Q. Screenshots of final draft. Available at: http://jiscstore.jot.com/WikiHome/SurveyPhase/StORe%2BQ.ppt?revision=1#256,1,StORe Q (Last accessed 04/09/2006)

was because they did not have a need for them. Those that were thought to make the most extensive use of automated chemical information were theoretical chemists, for whom it was noted that *although some physical chemists would have need for information based on chemical structures, the majority would not, as their information would be suited to alpha-numeric format"*.

Developments that the e-science programme (http://www.rcuk.ac.uk/escience/) initiated since its inception in 2001 have been well documented in the literature, in particularly for chemistry by the eBank project (http://www.ukoln.ac.uk/projects/ebank-uk/) which addressed the role of repositories in linking research data to peer reviewed papers and how such a service has an impact in the scholarly communication and publication (Lyon, 2003; Lyon et al, 2004). Coles and colleagues (2006) described how the UK National Crystallography Service has developed the eCrystals repository in which electronic files that are produced in the process of the crystal structure determination are captured and validated and also are assigned relevant metadata that is automatically generated and aim to support publication and dissemination of the information. Other relevant projects that were initiated by the e-science programme were the Comb-e-Chem (http://www.combechem.org/) and the ECSES (http://www.it-innovation.soton.ac.uk/research/grid/comb_e_chem.shtml). They both run by the iT Innovation research centre (http://www.it-innovation.soton.ac.uk/) at the University of Southampton and aimed to "*develop an e-Science testbed that integrates existing structure and property data sources within a grid-based information-and knowledge-sharing environment*".[3]

4 Methods

The StORe project (http://jiscstore.jot.com/WikiHome) employed two methods to gather information about the use and the linkage of source and output repositories, with regard to researchers working in seven scientific domains. These methods were: a) an online questionnaire survey and b) interviews with members of academic staff from institutions across the UK. This paper presents the results from the questionnaire survey among chemistry researchers.

The questionnaire survey was launched on the 13th March and closed on 21st April 2006. It was publicized among 728 members of the chemistry research community at the following universities: Imperial College London, Bristol University, Cambridge University, Southampton University, University of Durham, University of Oxford, and 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.

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. Thirty eight people

[3] CombeChem. About CombeChem. Information available at: http://www.combechem.org/about.php (Last accessed 04/09/2006)

responded to the questionnaire survey representing 10% of the overall response that the questionnaire received and 5.2% within chemistry itself. The 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.[4]

5 Results

The online questionnaire comprised four sections that are discussed below. These were preceded by an introductory section that aimed to gather information relevant to demographic characteristics of the researchers, such as the scientific domain they represented, their employing organisation, their occupation and contact details, if they wished to provide them. Almost half of the response (47%) came from postgraduate research students. 40% of the responses came from academic staff and the remaining 13% represented responses by postdoctoral researchers, research assistants and contracted researchers. Undergraduate students were not targeted as a group and therefore there was no response received from them. Also, there was no response from any independent researchers. Analytically the response is presented in the following table.

Table 1. Response to the questionnaire survey by role of the respondents

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

5.1 Section A. The need for linking repositories

The first part of the questionnaire comprised questions that aimed to identify the need for linking source and output repositories. The respondents were invited to indicate how advantageous it would be for their research if they had the ability to link from primary research data to their published outputs and vice versa. Some examples of potential future use included the ability to count actual papers' downloads and

[4] 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)

therefore argue that the impact of a research paper had been increased. Also, the ability to track the timeline in the process and outcome of a given set of research data. Or even link a data set to researchers that had downloaded and used it for their own research.

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 of the point of the survey as they had only recently commenced their doctoral studies and they were unable to judge the significance such a facility would have for their research. The reverse of this facility, 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 1).

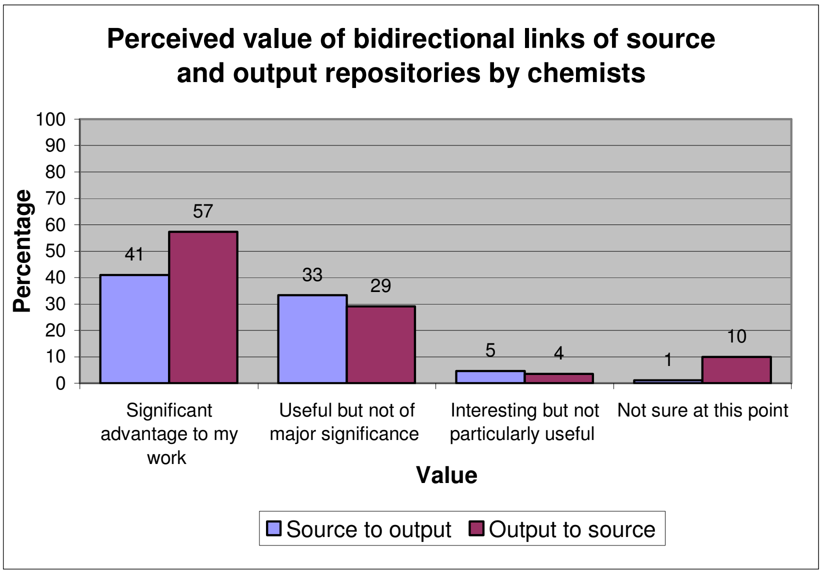

Fig. 1. Perceived value of bidirectional links of source and output repositories – Chemistry research

More than half of the chemists that were surveyed (65%) had not used a repository before and they were not familiar with the idea of open access repositories in general. They noted thought 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 thought again 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 they tended to specify mainly for application and use by their students rather than themselves.

5.2 Section B. Research data and source repositories

The second part of the questionnaire aimed to gain some understanding about the type of data that chemists produce and the formats in which it is stored. Such information could prove useful to people working in the set up of repositories in general as it provides useful insight of data types produced and could indicate software requirements for the deposition and retrieval such information. The questionnaire respondents were also asked to denote what metadata is assigned to their data and at what stage. The respondents to the questionnaire were invited to select from a range of different types of source data that they generate 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 that these were "mainly binary and text files from calculations, with figures and graphs derived from these".

The format in which this data is saved and held 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%). 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.

The respondents were invited to select from a list of metadata fields from which they were asked to indicate those that they considered 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%).

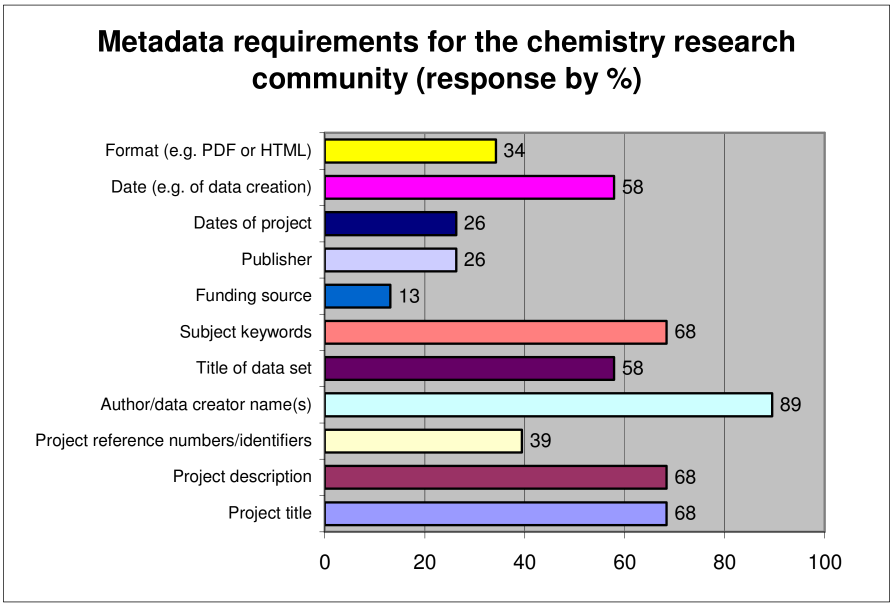

Fig. 2. Metadata requirements for chemistry researchers

The respondents were invited to indicate at what stage metadata is assigned to a resource as part of their own processes and practices, by making selections from key stages identified in the questionnaire. The responses to this question were fairly evenly spread across the options offered, which may indicate that the respondents were not familiar with the concept and the practice 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 another the smallest group of respondents (5%) indicated that they were not sure when metadata is assigned.

More than half of the chemistry respondents (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 they were unaware of who assigns the metadata to their resources, which again complements 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 was 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.

5.3 Section C. The accessibility and sharing of primary research data

The aim of the third part of the questionnaire was to gather some understanding about the perceived advantages and barriers in making research data available, and where researchers do so, to find out if they apply any restrictions on how it may be accessed. The respondents were invited to indicate what measures they normally use to control access to their data by other researchers. All respondents indicated a variety of measures. The majority of the responses from the academic staff indicated storage of their data on a private network/intranet (21%) as the main measure to control access. The same measure was also employed by a large proportion of the postgraduate research students (32%) as well. All of the contracted researchers noted that they use authentication of ID and passwords for controlling access to their data. The research assistants indicate that they tend to select storage of their data on standalone computers (16%) as the main measure for controlling who has access.

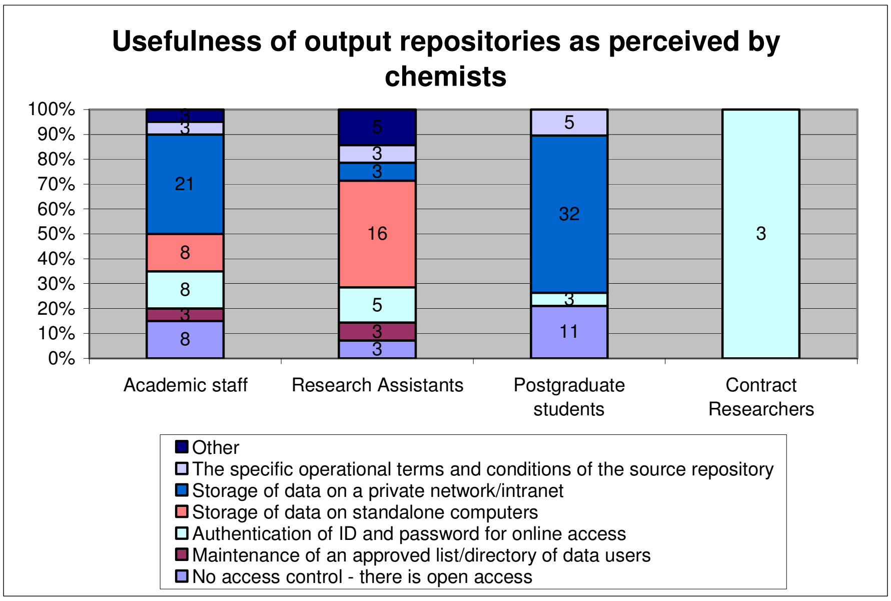

Fig. 3. Usefulness of output repositories for chemistry researchers

5.4 Section D. Output repositories

The fourth section of the questionnaire aimed to gather some information regarding the repositories that the respondents have used for their research and for teaching purposes. In addition, they were asked to indicate how they usually search repositories for information, the amount of support they have been given (and from whom) and how much they would have liked to receive. The majority of output repositories that chemists tend to use are those in the commercial sector, set up and managed by publishers. Academic staffs are the group who indicated that they used the widest range of repositories, including institutional, discipline based and publisher repositories. 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 is divided between institutional and discipline repositories. This pattern is similar to the repository usage indicated by contracted 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.

Although, the majority of the chemistry respondents to the questionnaire replied that they preferred to use the simple search option when they visited both source and output repositories, the response is quite spread again according to the different types of repositories. The majority of those who tend 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.

The respondents to the questionnaire survey were invited to indicate their preferred ways 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 route was via an Open URL resolver (11%). A small number indicated that they do not have a preferred route for accessing repositories. Few of the respondents (5%) indicated other than those prescribed routes and they specified "Web of Knowledge, SciFinder, or that they had only recently started their research, so they do not have any preferred routes of accessing a repository yet".

6 Conclusions

This paper presents results from an online questionnaire survey, undertaken as part of the StORe project, aiming to identify the issues around the use and linkage of source and output repositories and the expectations of the chemistry research community about their use. From the questionnaire survey response the following conclusions can be made:

- There is some indication from the questionnaire survey that the concepts of source and output repositories, as well as the model of open access, is not yet widely known and accepted in the chemistry research community as more than half of the chemists that were surveyed (65%) had not used a repository before and they were not familiar with the idea of open access repositories in general.

- In spite of this, bidirectional links between repositories and in particular, a bidirectional link between a source and output repository has been perceived as something that would be either a significant advantage or useful for the research conducted in the chemistry domain.
- Academic staff indicated a preference of 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.
- There are many variations in the type of data produced, their recording and storage and also in the perceived value of repositories. The most common type of data produced among chemists is SPECTRA data that it is represented in drawings, spreadsheets and image files.
- Although the majority of the respondents denoted that they use a simple search when they visit a publishers' repository, the use of subject specific thesauri and Boolean logic is used when they navigate institutional or discipline repositories.
- In general it was felt that the availability of a prototype that would illustrate the aims of the StORe project to developing a facility that can link source and output repositories, would have made it easier for the respondents to understand and comment upon advantages and barriers to use.

7 Bibliography

1. Philip, G and *Cunningham*, F P (1995). Availability and use of automated chemical information systems by academic chemists in the United Kingdom. British Library. *Research* and Development Department. BLRD Report; 6184 1995, 63p
2. Lyon, Liz. (2003). "eBank UK: Building the links between research data, scholarly communication and learning", Ariadne (Issue 36), Available at: http://www.ariadne.ac.uk/issue36/lyon/intro.html (Last accessed 05/09/2006)
3. Lyon, L., Heery, R., Duke, M., Coles, S J., Frey, J G., Hursthouse, M., Carr, L. and Gutteridge, C. (2004). eBank UK: linking research data, scholarly communication and learning. In, *eScience All Hands Meeting*. Swindon, UK, Engineering and Physical Sciences Research Council. Available at: http://eprints.soton.ac.uk/8183/ (Last accessed 05/09/2006)
4. Tenopir, C., & King, D.W. (2002) Reading behavior and electronic journals.Learned Publishing, 15(4), 259–266.
5. Coles, S., Frey, J., Hursthouse, M., Milsted, A, Carr, A., Gutteridge, C., Lyon, L., Heery, R., Duke, M., Koch, T. and M. Day (2006). Enabling the reusability of scientific data: Experiences with designing an open access infrastructure for sharing datasets. Presented at the, Designing for Usability in e-Science. International Workshop, Edinburgh, Scotland, 26-27 January, 2006. Available at: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html (Last accessed 05/09/2006)