Beyond Publication – A Passage Through Project StORe

Graham Pryor

University of Edinburgh
Digital Library Division
George Square, Edinburgh, Scotland
e-mail: graham.pryor@ed.ac.uk

Abstract

The principal aim of Project StORe is to provide middleware that will enable bi-directional links between source repositories of research data and the output repositories containing research publications derived from these data. This two-way link is intended to improve opportunities for information discovery and the curation of valuable research output. In immediate terms, it is expected to improve citation rates as a consequence of increasing the accessibility of research output. A survey of researchers in seven scientific disciplines was used to identify workflows and norms in the use of source and output repositories, with particular attention being paid to the existence of common attributes across disciplines, the functional enhancements to repositories considered to be desirable and perceived problems in the use of repositories. Cultural issues were also investigated. From the results of the survey, a generic technical specification was designed and a pilot environment created based upon the *UK Data Archive* (source repository) and the London School of Economics' *Research Articles Online* (output repository). A further link to a prototype institutional repository at the University of Essex was used as a control mechanism. The StORe middleware was designed using a Web 2.0 approach similar to existing FOAF (Friend Of A Friend) services such as Flickr and MySpace, but incorporating a federation of institutional, source and output repositories rather than one central area where digital objects are deposited. Researchers can deposit digital material in various formats at their institutional repositories until the data and publications are made available at linked source and output repositories. An enabling central portal provides an OAI-based aggregator service, which harvests the contents of the federation's repositories and provides a simple search facility. Whilst all digital objects are title visible, a key feature of the middleware is the Flickr-like option for regulating access, which gives researchers control over who can see objects they have designated 'non-public'. Using the StORe middleware, it will be possible to traverse the research data environment and its outputs by stepping seamlessly from within an electronic publication directly to the data upon which its findings were based, or linking instantly to all the publications that have resulted from a particular research dataset. It has already been endorsed by participating researchers as having the potential for integrating multiple data sets from different publications. Following completion of the pilot demonstrator, an independent evaluation undertaken by the National Centre for e-Social Science found it effective and easy to use. It may also be said to have broadened the meaning of the terms *publish* and *publication*.

Keywords: interoperability; research publications; institutional repositories; middleware

1 Introduction

Project StORe is an initiative funded by the UK's Joint Information Systems Committee within its 2005-7 Digital Repositories Programme.[1] StORe's principal aim is to attach new value to published research through the provision of two-way links between the output repositories that contain research publications and the source repositories of original and processed data from which those publications originated. Hence the project name, which is an acronym of **S**ource **t**o **O**utput **Re**positories. This bi-directional linkage is predicted to increase opportunities both for information discovery and the curation of valuable research data. Specifically, it will provide members of the research community with the means to navigate directly from within an electronic article to the source or synthesised data from which the article was derived; conversely, direct access will also be provided from source data to the publications associated with those data. Researchers will benefit from this linkage through an enhanced capacity to track the use and influence of their published research, as well as to engage in the more comprehensive dissemination of research and scholarship, which it is anticipated will increase the citation rate for research papers linked to their sources. Scientific researchers involved in the development phases of the project have already identified other advantages, such as the ability to conduct a reanalysis of source data as new methods emerge, a feature that should lead to improvements in the integrity of

published results, whilst the potential for integrating multiple data sets from different publications has been perceived as promising time saved and more productive research.

On the subject of reanalysis, an incident reported in *Science* late last year[2] underwrites the potential value from being able to take a critical look at a published paper alongside its data. In a September 2006 paper in *Nature*, Swiss researchers cast serious doubts on a protein structure described in a 2001 *Science* paper by Geoffrey Chang's pioneering group at the Scripps Research Institute, San Diego. Upon investigation, Chang found that his homemade data-analysis program had inverted the electron-density map from which he had derived the final protein structure. Consequently, Chang and his colleagues had to retract three *Science* papers and report that two papers in other journals also contained erroneous structures. If his original paper and its data had been published together perhaps this mistake would have been discovered earlier.

Having referred here to the dual *publication* of a paper with its data opens up a more controversial realm than is first suggested by the design of a piece of functional middleware, since one may speculate that the provision of a mechanism for accessing not just electronic publications but also their underlying data raises fresh questions about the nature and meaning of the terms *scholarly* or *scientific publishing*. Reflecting on the open access publishing and repository movements, one detects a strong current of opinion that making data available does not constitute publication, which benign strategy contributes of course to the avoidance of unhelpful quarrels with publishers; but greater flexibility of interpretation and less defensiveness would be both appropriate and defensible, since the publication of scholarly papers and the dissemination of data are necessarily distinct acts, each being defined by their particular purpose. Any set of data selected specifically for inclusion in, or as the basis for a scientific paper is chosen with the principal purpose of helping to persuade the reader to accept a hypothesis or theory as proven, and its value is gauged by the degree to which it supports the effectiveness of the set piece of rhetoric that is the paper. The larger collection of data from a research programme, possibly archived in a source repository, does not serve that same purpose of persuasion. Indeed, it may be argued that by making this broader cache of data accessible via a link from a scientific paper to its source repository could even subvert the arguments in the paper, should there be weaknesses in the data or the research, although from a different perspective this does strengthen the case for the bi-directional link as a mechanism for ensuring the integrity of the source data. So whether making data from a source repository publicly available is an act of dissemination or publication, the answer is probably irrelevant. What is more to the point is the impact from enabling dual accessibility.

The impetus for Project StORe came from a belief held by members of the research library community that an achievable set of functional enhancements to both source and output repositories could be identified and built, on a generic basis, as a piece of middleware, and that this might be approached in a manner similar to the way in which digital library technologies have produced generic tools in other heterogeneous environments, such as ‘metasearch’ interfaces to publisher and local databases, metadata harvesters and link resolvers. These tools are based upon recent digital library protocols and standards such as OAI-PMH,[3] qualified Dublin Core[4] and OpenURL.[5] Project StORe was therefore conceived as a vehicle for undertaking the essential groundwork preparatory to building a production system solution that would meet the requirements for permitting useful interoperation between the two repository types, and it would be undertaken using the systems, standards and metadata protocols developed and used in other JISC projects, where appropriate, to ensure the widest possible interoperability. Its rationale would be that of a proof of concept, but from the start there was a firm aspiration to deliver an authentic pilot infrastructure capable of translation across multiple disciplines.

2 Methodology

In the first phase of the project a survey of researchers was conducted across seven scientific disciplines in the UK to understand their workflows and working philosophies, as well as to identify norms in the use of source and output repositories. The disciplines investigated were archaeology, astronomy, biochemistry, the biosciences, chemistry, physics and social sciences. The astronomy survey had a broader base, including members of the astronomy research community in the USA, in recognition of the internationally collaborative work undertaken by astronomy research teams at Edinburgh and Johns Hopkins universities and the discipline’s separate Mellon-funded analysis of repositories and applications. The survey, which was carried out over four months in 2006, first through an online questionnaire and subsequently by one-to-one interviews, addressed such issues as the existence of common attributes across disciplines (in terms of the data formats employed, the quality and method of metadata assignment, and the volume of data produced), the functional enhancements to repositories that were considered to be desirable, and the nature of problems experienced in the use of repositories. Cultural and organisational issues were also investigated, ranging from attitudes towards the concept of open access publishing to the measures employed for sharing and protecting data. Invitations to

participate in the online questionnaire were sent to 3,700 scientific researchers and produced a return in excess of 10%, whilst the in-depth interviews were held with between 10 and 15 respondents per discipline, selected to ensure an equitable representation from all stages of the academic/research career path. Each individual discipline survey produced a published study that described the source and output repositories used by members of that discipline, including a brief history and statistical information on their use, with a detailed analysis of responses to the questionnaire and the structured interviews. These reports, which have been archived in the Edinburgh Research Archive (ERA), also incorporate scenarios and use cases.[6]

Project partners at university libraries identified staff to undertake the discipline surveys, with a view to exploiting their knowledge and the effectiveness of their relationships with researchers 'on the ground'. The libraries responsible for the survey work and the disciplines they surveyed are shown in Table 1.

Surveying University Library	**Subject**
Edinburgh (lead) / Johns Hopkins	Astronomy
Birmingham	Physics
Imperial College	Chemistry
London School of Economics	Social Sciences
Manchester	Biosciences
University College London	Biochemistry
York (for the White Rose Partnership)	Archaeology

Table 1: Project partners & survey disciplines

Whilst it was important to the design of a relevant and appropriate solution that actual research working practices and environments would be identified and understood, the survey team's principal role was to address the requirements for new functionality within source and output repositories that would permit interoperability from the point of ingest, so that authors of papers could insert links to data and to published/unpublished papers, associating newly deposited publications with data held in data repositories. It was anticipated that a number of new operations could be supported within the two types of repository, both for academic submitters and for repository users, including automatic link creation, automatic embedding of source repository metadata, and a facility to run operations upon data. The desirability of these features was explored in depth during the interviews.

Upon completion of the survey, a business analysis of the survey reports was undertaken by staff at the UK Data Archive (UKDA).[7] This analysis was used as the foundation for a generic technical specification of the proposed bi-directional link, with the aim of translating real requests for 'missing' functionality into a structured technical architecture. The assumptions and deductions made in the business analysis were then tested with research active staff at the University of Essex and with library professionals from the London School of Economics (LSE), leading to further refinements to the specification.

In the final phase of this development process, the generic technical specification has provided the platform for the pilot implementation of a working bi-directional link. This has featured social sciences data and publications exclusively, using the *UKDA* as the source repository and the LSE's *Research Articles Online* as the test output repository, augmented by a further link to a prototype institutional repository at the University of Essex, which served as a control mechanism. It should be emphasised that limiting the pilot to only one of the original seven disciplines has been necessary to meet the logistical constraints of a test environment, but in building that environment the full set of requirements established by the survey of seven disciplines has been incorporated with a view to proving the middleware as a generic, non-discipline specific tool.

Throughout, the rationale of Project StORe has been to anchor technical and user aspirations to the pursuit of practical benefits. During the pilot implementation, a critical element of the process has been user testing, involving members of the original cohort who responded to the survey, and at its conclusion the pilot demonstrator has been subject to a rigorous, independent evaluation by the National Centre for e-Social Science,[8] which has depended for its legitimacy upon user participation in a series of workshops.[9]

3 Survey and Analysis

A majority (85%) of respondents to the StORe survey judged the provision of a bi-directional link as likely to prove advantageous to the research process, with a small preference overall for an output to source link. Key benefits were described as an opportunity to access the large data sets it is not possible to reproduce in an article;

and more specifically, an output to source connection would enable the comparison of results, thereby providing the means to authenticate claims made, which was deemed to be of particular value where claims are considered controversial.

By selecting from prepared lists, respondents were asked to identify the data types and their formats that might be generated during research, with the range of data types given in the lists appearing to satisfy the majority as being representative, and with no data type receiving a nil response. A further 32 *Other* types were also declared but were found to describe either a sub-type of items from the lists, the name of experimental equipment or process-specific data sets. Nonetheless, across and within the seven disciplines, the volume, range and diversity of data produced was confirmed as considerable. Whilst generic types such as drawings, plots, images and text-based files scored highly, each showing in excess of 150 responses, noteworthy scores were attributed to more specialised types such as radiographic data (11), remote sensing surveys (15) and gene/protein sequences (42).

In terms of data format, image files, spreadsheets and word processed files comprised the majority, with around 200 responses each. In the next tier, plain text, database files, portable document format and tables/catalogues all scored more than 100 responses. Of the 76 *Other* formats volunteered by respondents, those that were not species from the main selection list tended to be proprietary and linked to specific discipline processes or equipment. Of greater significance to the design and maintenance of links from publications to their source data is that almost three quarters of the survey's respondents were found to generate and use complex data sets (i.e. data produced and held in combinations of data formats and files).

All of the seven disciplines identified barriers to the deposit of data or publications in repositories, citing time constraints, the bureaucracy imposed by repository administration and structures, or constraints arising from their own or others' intellectual property rights. A perceived inconsistency across all repositories was also reported in terms of content coverage and in the standards and methods used for keywords, metadata and data formats. It was in this latter area that the most powerful consensus was found amongst the survey cohort, with the appropriate assignment of metadata being roundly acknowledged as critical and demanding, both intellectually and in the time required to do it well. Perversely, this consensus on the need for good metadata did not necessarily translate into good practice, there being a high level of self-assignment and with limited evidence that standard schema or thesauri were being employed. Perceived responsibilities for metadata assignment are illustrated by the following table from the StORe questionnaire.

I decide which terms to use and I assign them		**212**
Research colleagues assign metadata on the team's behalf		**55**
Research support staff assign metadata on the team's behalf		**22**
Metadata are assigned by library/information services staff		**4**
Metadata are assigned by the repository administrators		**37**
Metadata are generated automatically		**63**
It is not known who assigns metadata		**68**
Other (please specify)		**37**

Table 2: The Assignment of Metadata to Research Data

In order to establish whether there is a core set of metadata that might satisfy the needs of researchers in the seven disciplines, respondents were invited to identify key terms from a predefined list and to suggest their additional requirements. A large majority subscribed to the list as representing a functional generic suite of metadata, selecting such terms as project title, description and reference numbers, together with keywords, project and publication dates and format. Only 58 *Other* terms were suggested, and these were found to be highly discipline specific (e.g. archaeological period, celestial object, position and observation date, chemical entity, protein sequence).

As shown in Table 2, the subject of metadata provision revealed a broad spectrum of awareness and response amongst the survey cohort that was sustained when they were asked to indicate the point at which metadata are assigned. Assignment 'during file saving' attracted the highest score of 142, but there was insufficient evidence to deduce whether such a practice represented a properly structured activity or merely the casualty of afterthought. More reassuring were responses to the options 'Prior to data creation' (82), 'As part of the indexing

process' (98) and 'When submitting data to the repository' (89). Of some concern were the 35 respondents to this question who believed no metadata were being assigned to their research output, with a further 75 admitting they were not sure at which stage metadata are assigned.

The disjunction between aspiration and practice in the assignment of metadata is perhaps explained by tensions between the prevailing research culture and embedded attitudes towards the support services. It was made clear during the StORe survey that researchers from all disciplines favoured self-reliance in matters associated with data management and the use of repositories, as opposed to the provision of institutional support from the library or other areas of professional expertise. The inherent culture of self-sufficiency within research groups or programmes, where normal practice is to manage all aspects of the research lifecycle internally, was evident from statements submitted during the StORe survey. Whilst this culture has given rise to the development of some highly effective data repositories focused on serving specific disciplines, the general effectiveness of a self-sufficient approach to accessing, organising, promulgating and curating data was not demonstrated across the scientific research spectrum.

National and international strategies for data deposit and preservation are of course already emerging. One can point, for example, to the Wellcome Trust's flagship initiative to mandate the deposit of research publications in the biosciences, which mandate is anticipated will extend to the deposit of data; or to the astronomy community's Virtual Observatory, an initiative to make all the astronomy data in the world easy to access.[10] They are not isolated examples, but when one considers the research milieu as a whole their considerable progress was found not to be typical. At the level of the individual researcher, whether asked about metadata assignment in particular or data management in general, responses such as "it's my problem, I'll deal with it" were commonplace. Whilst libraries have conducted advocacy campaigns on behalf of open access publishing and repositories, in some cases providing technical expertise to support the use of repositories, researchers canvassed by the StORe survey in most cases perceived there was no support available, they had little confidence in what support was known to be provided, and they claimed sufficient familiarity with information technology to consider themselves self-reliant. Yet at the same time as declaring they would not normally associate the management of research data with librarians, and evincing little apparent demand for assistance in seeking and navigating information, there was evidence of a clear requirement for information intermediaries to assist not only in the construction and maintenance of metadata but also in the preservation and curation of data. This dichotomy was reflected in a further aspect of the survey, which concerned researchers' attitudes towards making data available, and would prove a singular force in the design of the StORe middleware.

With few exceptions, respondents to the survey supported the statement that it should be a requirement for data from publicly funded research to be made freely available, but generally with the caveat that access should be restricted until results are published in a paper, in order to prevent data scavenging. Others noted that whilst this might be a creditable aspiration, without a data administrator it represented a potentially large burden from editing, compiling and sanctioning the release of data. In fact, both the provision of access and the sharing of data were found to be constrained by a lack of confidence in processes, and it was difficult to conclude whether some practices were deliberately designed to frustrate accessibility. For example, the storage of unique and original research on PCs and laptops was found to be common practice, and the failure to take a more relaxed approach to access was influenced by a perceived absence of adequate protection in networked systems. As one respondent described his data management regime: "data is held on secured CDs in encrypted format with only an identifying code. The codebook is kept physically separate".

The StORe survey revealed a range of diversity in practice and attitude, both within and between the seven disciplines, but with a consistently firm body of consensus when it came to explaining fundamental needs. When searching for information, a universal preference for simple keyword searching was declared and browsing amongst library shelves appears to have been replaced by browsing within repositories and other online resources. This practice is of course only effective when enabled by the functional efficacy of application and metadata structures, designed by system and data experts to meet the clamour for a 'Google-type' approach to searching.

4 The Generic Model

The business analysis that followed the StORe survey revealed sufficient shared ground between the disciplines to suggest the basis for a common model. To recap, an examination of the discipline-specific reports produced a majority in every discipline favouring two-way links between data repositories and publications, but with barriers to the actual deposit of data or publications found to be a consequence of time constraints, organisational bureaucracy or concerns over intellectual property rights, although the concept of data sharing was considered

fundamental and important. A perceived inconsistency across all repositories in terms of coverage, standards and data formats was reported, with a simple 'Google type' approach to searching being preferred. Researchers from all disciplines also seemed to exhibit self-reliance in matters of data management and in the use of repositories, whilst recognising the need for assistance in the provision of some common minimum metadata.

Taking this level of consensus, the design of the model for a bi-directional link has adopted a Web 2.0 type approach, similar to existing FOAF (Friend Of A Friend) services such as Flickr or MySpace, but incorporating a federation of institutional, source and output repositories rather than one central area where digital objects are deposited. Articulation of a Web 2.0 rationale for the middleware has been a deliberate decision aimed at meeting cultural aspirations for self-determination and those individual anxieties concerning data ownership that were revealed during the survey, since it places control firmly in the hands of the researchers. In this model, objects deposited in federated repositories would be referenced by persistent identifiers that include domain identifiers, with researchers depositing digital material in various formats at their institutional repositories until the data and publications are ready to be made publicly available at linked source and output repositories. This focus on the institutional repository environment is predicted to have further value in providing a context for future implementations of asset-based research data repositories, in cases where global services from established discipline platforms such as astronomy's Virtual Observatory or the social sciences' UKDA are not provided, and discipline needs could be met instead by a regime of institutional data curation.

What may be described as the central StORe portal has been designed as an OAI-based aggregator service that will harvest the contents of a federation's repositories and provide a simple search facility based on centralised indexes. This basic level of searching can be enhanced for individual disciplines by the inclusion of domain ontologies, reflecting the need highlighted in the survey to enable discipline-specific terminology. All digital objects will be title visible to all, but researchers can restrict access to non-public objects to communities of project-specific colleagues, institutional colleagues, personal colleagues, or all of these. This is similar to the option for restricting access to family and/or friends in Flickr, in order to bar public access to private photographs, and is again a direct attempt to satisfy the demands of researchers to remain in command of their data.

Access management has proved to be a defining feature of the StORe middleware. Some data repositories are open to all enquirers, while others are password-protected, and in a scenario where users of open access research publications wish to view data in repositories to which access is normally controlled, a validation process will be required in order to allow temporary access rights. In this context we have investigated the authentication and authorisation issues involved with reference to the developing international work on Shibboleth, a federation-based architecture that enables organisations to build single sign-on environments for accessing Web-based resources.[11] Whilst it is not yet in place, it is planned that a production version of the central StORe portal will authenticate through Shibboleth, using a simple deposit interface to request the minimum amount of mandatory metadata for each object, identify the group or individual to which it is accessible and check whether it is a candidate for public submission. Until Shibboleth is adopted, we are applying a dummy Shibboleth mechanism for allocating user names and passwords. This will trigger an automatic process for setting up user accounts when legitimate users log in for the first time.

The minimum metadata required for any individual item is a title, provided the item is being associated with project data in a repository already assigned the metadata elements *author*, *title*, *geography*, *time*, *keywords* and *abstract*. The digital object will be deposited in the researcher's institutional repository, whilst the metadata and access conditions will be stored centrally; in turn, the search indexes will be built up from the centrally held metadata and harvesting from the objects themselves. This harvesting can also be used in the creation of the discipline-specific ontologies needed to satisfy metadata requirements that are not met by the generic core. Both source and output repositories in the federation will regularly trawl for potential acquisitions and, if a publication or data are accepted, the repository will supply a public link to a peer-reviewed version of the publication or to the data.

Hence, the generic model planned to be tested by the pilot demonstrator combines informal networking and sharing of data with a public access system that supports stronger links between data sources and publications. A user entering a StORe generic portal would log in to authenticate and the system will respond by determining his/her organisation, recorded preferences and known colleagues. Options would then be made available to browse any new activity of colleagues; to browse any objects available to the user (i.e. the user's own and other colleagues' objects); to search all discipline-specific or all repository-specific objects, with a further option to filter on a temporal basis; to deposit an object; to create a new project; to make an object available to another user; to request that an object be made available; to submit an object to an output repository for publication; to

submit an object to a source repository for preservation; to download a repository object; and to edit, delete, organise or manage the user's own objects.

It was clear from the outset that the success of this model would be determined by three factors. Researcher acceptance of Web 2.0 technologies was essential, and we have been actively encouraged here by the younger members of the user testing cohort who already work routinely within that environment. Persuading researchers to use a third party portal for deposit into their local institutional repository was also acknowledged to be challenging, whilst the third and possibly most difficult obstacle lay in the resolution of potential security and policy objections to sharing sensitive data across institutions. Eventually, it was decided that these barriers could be broken down in a stage by stage approach that would embed a federation in the established publishing process and restrict the sharing of non-public data to institutional colleagues. A demonstration of simplicity would be the key to stage one, with the objects stored required to be identifiable only by title, discipline, project, file type and format, employing minimum Dublin Core metadata elements. In the second stage, each individual institutional repository would act as a portal to itself and all the domain specific source and output repositories in its federation, thereby preserving familiarity of the working environment but allowing the addition of Web 2.0 and FOAF features. Only at stage three would the concept of a StORe subject or domain portal be openly introduced to the discipline-specific elements of the federated repositories. Here, one solution to security concerns would be the temporary copying of protected objects to the portal for download within a prescribed period.

Looking beyond the pilot environment, this approach offers wider coverage, more choice of source and output repositories and more scope for Web 2.0 service features. There could even be a common interface for deposit to individual institutional repositories, and it was envisaged that listing of forthcoming conferences, wikis, and other networking facilities might encourage use. The final stage would see the full generic solution implemented, comprising the entire federated institutional, source and output repositories that have adopted the approach outlined in stage one. This solution is well placed to encourage cross-disciplinary research, a key driver in the modern research environment, although metadata mappings will have to be employed and even more additional features devised to encourage the use of such a universal portal.

5 A Passage Through Project Store

StORe's pilot demonstrator was built for a test federation using the UKDA as source repository and the LSE's *Research Articles Online* as the output repository, complemented by a prototype institutional repository at the University of Essex.[12] Options for linking to a commercial publisher had also been explored but were considered logistically too ambitious for a pilot implementation. The pilot was designed and implemented between November 2006 and April 2007, and what follows is an abridged system walkthrough showing how items (data and publications) are managed.[13] This description is of a standalone system, but in a live working environment access could be initiated within an electronic article in an output repository or from a source repository having an association with the federation.

In the pilot, as in a working system, it is possible for an unregistered user to search or browse across all or specific research collections in the federation, but any titles marked as private will not function as a hyperlink to their content. Collection metadata can, however, be seen via a *View Collection* link. If the research project from which the target collection was generated involved the secondary analysis of existing data, a link to the underlying data will already exist, and will take the user to the relevant Web page of the supporting source repository. If the collection owner has agreed, then a further link will appear, allowing users to send an email requesting additional details or to be granted access to items in the collection.

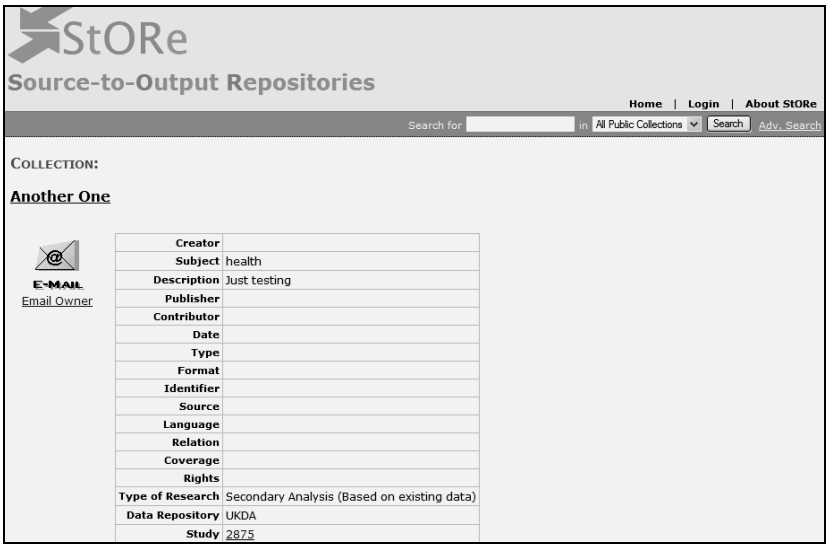

Figure 1: View Collection Metadata Screen

Registered users logging in to the pilot federation can view the content of all items in their public and private solely-owned collections. They can also see those items in public collaborative collections with the UKDA or *Research Articles Online* where they are a contributor, and may view public or private collaborative collections made with project colleagues or 'friends', where they are identified as either contributor or administrator. Collaborative collections are linked via a unique *LinkID*. In the example below, the user (identified as Forum) is a member of a private collaborative collection created by another researcher in order to share documents with Forum. Each collaborative collection is distinguished as a collection type, either source/archive, output/publisher or user/researcher. The logged-in and authenticated user has access to full functionality and can create private or public, solely-owned or collaborative collections, including an option to allocate other registered users to a collaboration.

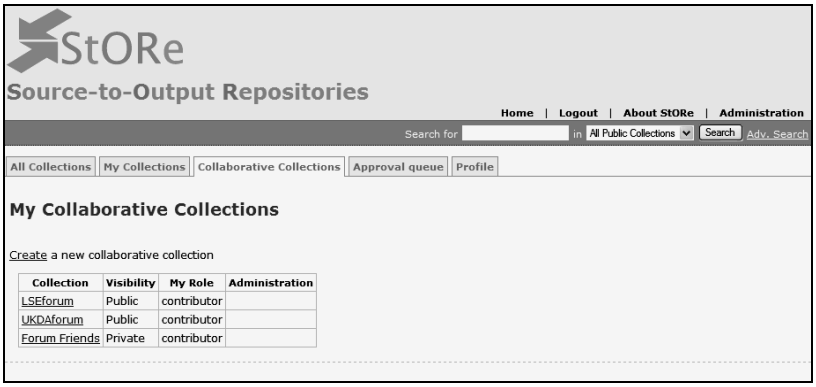

Figure 2: Collaborative Collections

Figure 5, overleaf, shows how the process of adding metadata to a publication has been kept simple. At collection level, apart from the collection name and description, only subject terms and the type of research and study (if secondary research) are mandatory. All other Dublin core fields are optional. The subject terms can be directly typed into the box or chosen from a list of tags displayed at the right-hand side of the page. The type of research is selected from a drop-down menu (Figure 3),

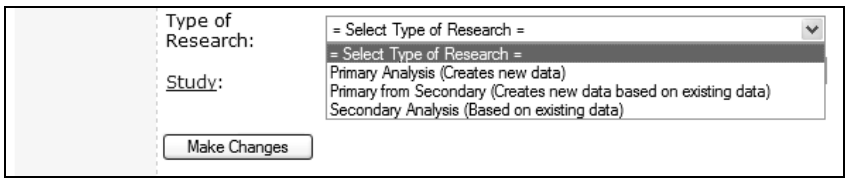

Figure 3: Allocating Research Type

and a study number corresponding to the number assigned to the corresponding data within the source repository is chosen from a further drop-down list (Figure 4).

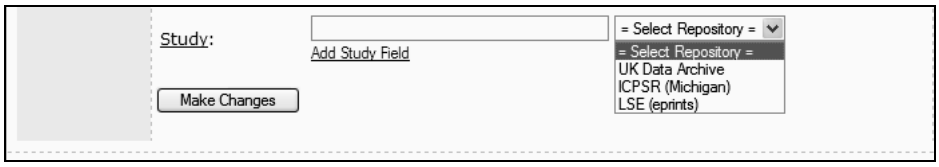

Figure 4: Selecting The Study Number

The study number then becomes a link to the appropriate page within the repository's web site.

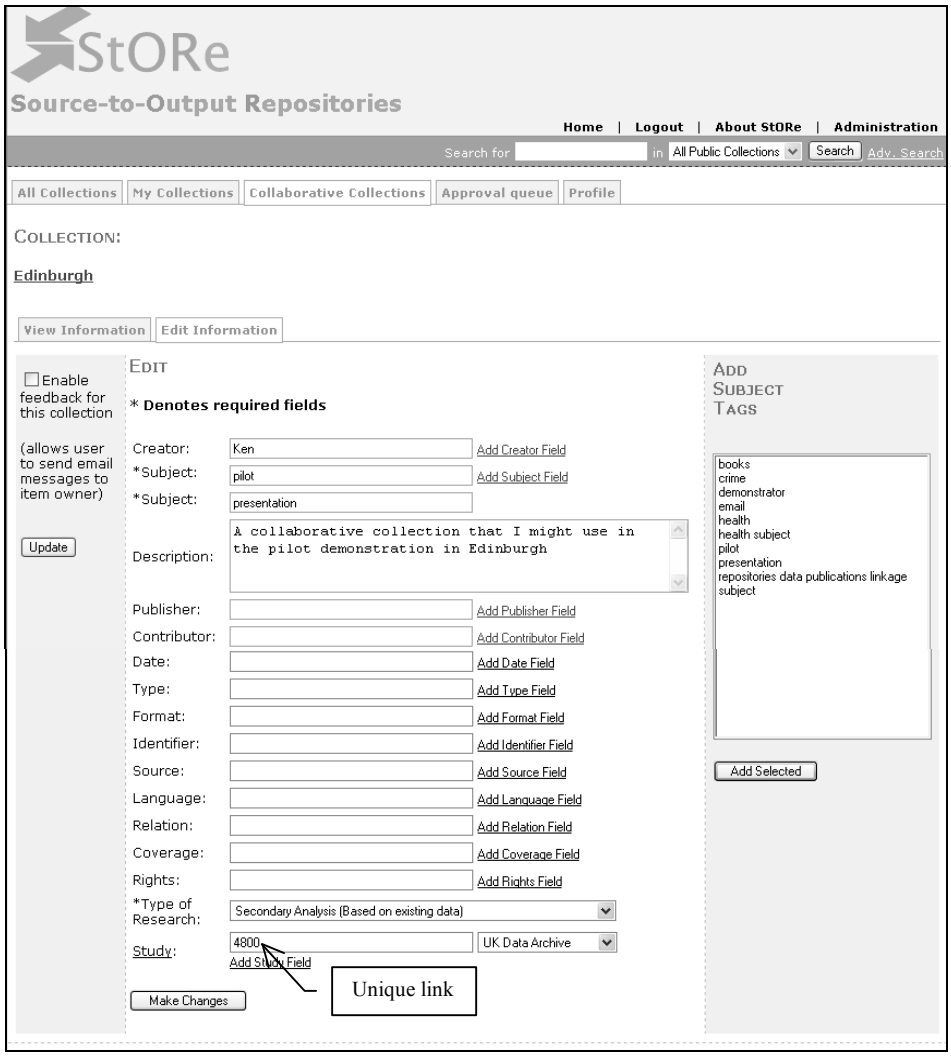

Figure 5: Assignment of Metadata

Moving a data item to the UKDA collaborative collection is a two-tier process. First, the data's identity is verified (which will enable publications based on this data to be moved and approved in *Research Articles Online*) and, where required, an embargo can be set by the data owner. Once verified, an acquisition number is assigned to the data item in the UKDA collaborative collection, which as already intimated will subsequently be assigned to any associated publications moved to a *Research Articles Online* collaborative collection. Upon approval by the UKDA this acquisition number is replaced by the actual research study number, which will function as a link to the data from its publications in *Research Articles Online*.

In StORe, individual items or folders are added to a collection either singly or bundled. Only the provision of an additional title and file name (or URL) is required, since each item adopts all the metadata associated with the

collection itself. Files in different formats (Word or PDF documents, URLs, image files, etc.) are associated to an item or folder, and the Dublin Core fields may be edited if required to produce a more specific metadata record. When a scientific paper ready for publication is moved from a researcher's institutional repository into a collaborative collection owned by *Research Articles Online*, all the metadata associated with it moves as well. Simultaneously, the middleware automatically assigns a metadata term to identify the collection of origin, and confirms that corresponding data exists in the UKDA collaborative collection. It also provides functionality enabling the addition of further files or URLs to the item, or to add additional metadata.

6 Conclusions

The StORe pilot has demonstrated the feasibility of a bi-directional link within the specific context of a single discipline. However, despite the level of consensus identified by the survey, discipline variations would need to be managed during export of the StORe model across other domains. Individual institutional repositories will also contain different file types and formats, and will apply different metadata standards. For certain disciplines data interpretation, manipulation and methodology are as, if not more significant than access to the raw data, and although a simple search might cross disciplines, more advanced discipline-specific searches would be more in demand, with the resulting hit lists, relevance ranking and sorting being different for each discipline. Consequently, both subject and global portals will require different Web 2.0 features for each discipline.

Recognising the key preferences and practices of researchers interviewed during the StORe survey, the solution developed showed that traditional practices for the informal networking and sharing of data could be combined with a public access system supporting stronger links between data sources and publications. The StORe solution gives researchers the means to manage a level of privacy and access defined by themselves, countering expressions of apprehension towards full open access, which some saw as a threat to data ownership. It also offers a simple Google type search, preferred amongst the majority of those surveyed, and viewed by many as an effective tool for replacing the option of browsing amongst shelves in a library, although Boolean operators and wildcard functionality are made available for more advanced searches. Using the StORe middleware, researchers can move seamlessly around the research data environment and its outputs, stepping from within an electronic publication directly to the data upon which its findings were based, or linking instantly to all the publications that have resulted from a particular research dataset. By intrinsically connecting the process of publishing scientific papers with the provision of their underlying data, StORe has also broadened the connotation of the terms *publish* and *publication*.

References

[1] http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories.aspx

[2] MILLER, G. *A Scientist's Nightmare: Software Problem Leads to Five Retractions*. Science, 22 December 2006, pp. 1856-1857

[3] An explanation of the Open Archives Initiative (OAI) and the OAI protocol for Metadata Harvesting may be found at http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/oai/

[4] The Dublin Core metadata element set is explained at http://www.ukoln.ac.uk/metadata/resources/dc/

[5] An OpenURL demonstrator can be found at http://www.ukoln.ac.uk/distributed-systems/openurl/

[6] The individual discipline reports, together with the survey overview, may be examined as documents within the Edinburgh Research Archive, http://www.era.lib.ed.ac.uk/handle/1842/1412 or at the project wiki, http://jiscstore.jot.com/SurveyPhase

[7] UK Data Archive at the University of Essex, http://www.data-archive.ac.uk

[8] http://www.ncess.ac.uk

[9] The NCeSS evaluation plan can be examined at http://jiscstore.jot.com/EvaluationOfPilot

[10] http://www.virtualobservatory.org/

[11] http://www.athensams.net/federations/shibboleth_intro.aspx

[12] Both institutional output repositories have been constructed using open source software: *ePrints* at the LSE and *Fedora* at Essex.

[13] A full walkthrough of the StORe middleware can be accessed at the project wiki, http://jiscstore.jot.com/PilotDemonstrator