# JISC DEVELOPMENT PROGRAMMES

## Project Document Cover Sheet

| | | | |
|---|---|---|---|
| **Project Acronym** | StORe | **Project ID** | |
| **Project Title** | Source-to-Output Repositories | | |
| **Start Date** | 1.9.2005 | **End Date** | 31.8.2007 |
| **Lead Institution** | University of Edinburgh | | |
| **Project Director** | John MacColl | | |
| **Project Manager & contact details** | Graham Pryor<br>graham.pryor@ed.ac.uk | | |
| **Partner Institutions** | • University of Edinburgh (lead)<br>• University of York, representing 'White Rose Partnership'<br>• University of Birmingham<br>• London School of Economics<br>• University of Manchester<br>• Imperial College London<br>• University College London<br>• UK Data Archive<br>• Johns Hopkins University | | |
| **Project Web URL** | http://jiscstore.jot.com/WikiHome | | |
| **Programme Name (and number)** | Digital Repositories | | |
| **Programme Manager** | Neil Jacobs | | |

## Document

| | | | |
|---|---|---|---|
| **Document Title** | Business Analysis | | |
| **Reporting Period** | | | |
| **Author(s) & project role** | Ken Miller (UK Data Archive partner) | | |
| **Date** | 15 December 2006 | **Filename** | StOReBusinessAnalysis |
| **URL** | | | |
| **Access** | ✓ Project and JISC internal | General dissemination | |

## Document History

| Version | Date | Comments |
|---|---|---|
| 1.0 | 04.09.06 | Draft |
| 1.1 | 22.09.06 | Draft (Board and Project Management Group) |
| 2.0 | 23.10.06 | First Revision (after Project Board meeting) |
| 2.1 | 07.11.06 | Second Revision (after Project Management Group meeting) |
| 2.2 | 15.12.06 | Final Revision |

**Project StORe**

# WORK PACKAGE 3: BUSINESS ANALYSIS

A generic specification for bi-directional links between source and output repositories, based on the required functional enhancements identified from the StORe survey of researchers (work package 2)

**Version: 2.2 (Final Revision)**
**Date: 15th December 2006**
**Author: Ken Miller (UKDA)**

# CONTENTS

## 1. Introduction to Project

The area of interaction between output (research publication) repositories and source (primary research data) repositories is the principal focus of the StORe project. The main aim of the project is to identify options for increasing the value of using both source and output repositories by improving the linkages between them, thereby increasing the potential from significantly enhanced information access and dissemination.

The functionality required by researchers in both types of repository was determined through a survey conducted using an on-line questionnaire and face-to-face interviews. The results from the survey, along with reviews of relevant publications and the experiences from similar projects, will determine the content, recommendations and conclusions of this business analysis.

## 2. Aims and Objectives of Business Analysis

The StORe project is multidisciplinary in scope, embracing the seven scientific domains of archaeology, astronomy, biochemistry, biosciences, chemistry, physics and the social sciences (originally described in the project plan as social policy and political science). The business analysis will formulate general principles for middleware development to link source and output repositories irrespective of discipline, but also outline domain-specific requirements. Consequent to this analysis a pilot demonstrator will be developed in the domain area of the social sciences. A full and extensive independent evaluation of the pilot demonstrator will be carried out in order to inform JISC of the best options for future development in this area.

The pilot demonstrator is a key deliverable from the StORe Project. It will consist of a set of middleware designed to demonstrate the function of bi-directional links between source and output repositories. This middleware will be developed to meet the specific needs of the social science e-research community, but based on the underlying general requirements as defined from the StORe survey of the behaviours of researchers within the seven scientific disciplines represented by the project.

This formal business analysis not only confirms the functional requirements identified from the survey, but addresses any perceived technical and organisational constraints to source/output interoperability. Conclusions drawn will provide the rationale for the technical specification and design of the pilot demonstrator, with the construction of linkages between the UK Data Archive (source repository) and multiple output repositories. The pilot will be used to demonstrate the implementation of enhanced functionality within a test environment and the potential for a generic solution across the UK's broader e-research community.

## 3. Background/State of the Art

The StORe survey, which ran from February through July 2006, comprised a description of source and output repositories from each target discipline, with an historical analysis of their use; an online questionnaire; and a series of structured interviews exploring the research lifecycle, the use of both source and output repositories, and opportunities and barriers to sharing source data.

To meet specific requirements of the business analysis, the StORe survey paid particular interest to the nature of workflows and norms for repository use, functional enhancements perceived to be desirable, and problems experienced with existing repositories. The report also identified common attributes across disciplines.

Within the JISC Digital Repositories Programme there have been several reviews, roadmaps and project reports that have outlined the present state of the art and direction that the area should and is developing. This business analysis will not duplicate those findings here, but list in the bibliography the relevant publication on which decisions have been based outside of the StORe survey findings.

The systems, standards, metadata and protocols developed and used in other JISC projects will be adopted where appropriate to ensure the widest possible interoperability. The following quote is taken from the 'Digital Repositories Review' published in February 2005:

*“Institutional repositories must be considered within the wider information environment. Creating small scale ‘silos’ of information within institutional repositories is not, on face value, a compelling information management strategy in the ‘Google age’.”*

Within existing institutional repositories open access and preservation have not always been a priority; however, many have been based on e-Prints, Dspace or Fedora architecture for managing and delivering their digital content with adherence to the Open Archival Information System Reference Model (OAIS).

Specially consideration will be taken of the eBank UK project which is concerned with enhancing scholarly communication by investigating ways to link e-prints and peer-reviewed articles to the primary research data upon which they are based, within the domain of combinatorial chemistry. The eBank demonstrator uses an e-Prints architecture and OAI Protocol for Metadata Harvesting and is underpinned by a data model and metadata schema for crystallography datasets.

StORe and the eBank UK projects share a common vision of an information environment where there is open access to both raw research data and publications, with a common technical infrastructure that allows interoperability between all types of repositories.

## 4. Generic Model

Analysis of the discipline-specific reports from the StORe survey, particularly material from the interviews, has revealed that there is some common ground between the disciplines that could form the basis of a common model. These are:-

1. That two-way links between data repositories and publications were thought to be useful by a majority in every discipline.
2. Each discipline identified barriers to the actual deposit of data/outputs, either because of time restraints, the bureaucracy imposed by repositories or constraints arising from their own or others' intellectual property rights.
3. The concept of sharing data was considered fundamental and important, although it is more likely to occur between known individuals than through deposit into a repository.
4. There was a perceived inconsistency across all repositories in terms of coverage, as well as in the standards and methods used for keywords, metadata and data formats.
5. The most common and preferred method of searching was a simple 'Google type' search.
6. Researchers from all disciplines seemed to favour self-reliance in data management and the use of repositories, as opposed to institutional, library or other support.
7. Members of all disciplines recognised the need for some common minimum metadata elements.

Based on the above level of consensus, the discipline-wide model for bi-directional linkages will take a Web 2.0 type approach, similar to existing FOAF (friend of a friend) web services such as Flickr, MySpace, etc., but incorporating a federation of institutional, source and output repositories rather than one central area where digital objects are deposited. The objects will be referenced by persistent identifiers that include domain identifiers.

Hence researchers can deposit digital material in various formats at their institutional repositories until the data and publications are made publicly available at linked source and output repositories. This focus on the institutional repository environment is predicted to have further value as the context for future implementations of asset-based research data repositories, where global services from 'big science' platforms are not provided and 'little science' is served by institutional data curation.

The central StORe portal will be an OAI-based aggregator service that harvests the contents of the federation's repositories and provides a simple search facility based on the centralised indexes. This simple search could then be enhanced for specific disciplines by the inclusion of domain ontologies. All digital objects will be title visible to all; however, access to non-public objects can be restricted by the researcher to their project-specific colleagues, their institutional colleagues, their personal colleagues or all of these colleagues. This is similar to the option for restricting access to family and/or friends for photos in Flickr rather than having them open to the public.

The central system would authenticate through Shibboleth and have a simple deposit interface that requests the minimum amount of mandatory metadata for each object, which group or individual it is accessible to and whether it is a candidate for public submission. The minimum metadata requirement for an individual item is a title, provided it is being associated with a usage/project that already has author, title, geography, time, keywords and abstract metadata elements.

The digital object would then be deposited in the institutional repository of the researcher, the metadata and access conditions would be stored centrally, and the search indexes would be built up from the centrally held metadata and a harvest of the actual objects. This harvesting could also be used for the creation of domain specific ontologies.

The source and output repositories of the federation would regularly trawl for potential acquisitions. If a publication or data are accepted then the repository would supply a public link in the system to a peer-reviewed, value-added version of the publication/data. Otherwise they would supply a reason for rejection, after which the researcher could apply to another federation source or output repository.

The proposed system combines informal networking and sharing of data with a public access system that supports stronger links between data sources and publications.

Once a user had come to the StORe generic model portal he would login to authenticate and the system would determine his organisation, preferences and known colleagues. The following options would then be available:-

1. To browse any new activity of colleagues.
2. To browse any objects available to present user (own and other colleagues).
3. To search either all, domain specific or repository specific objects with the ability to filter on a temporal basis.
4. To deposit an object.
5. To create a new project/usage.
6. To make an object available to other user.
7. To request that an object be made available to them.
8. To submit an object to output repository for publication.
9. To submit an object to source repository for preservation.
10. To download a repository object.
11. To edit, delete, organise or manage own objects.

However, there are several reasons why this approach might not be immediately accepted by academic researchers. Firstly there is a distinct distrust of Web 2.0 technologies amongst academics; secondly it will be hard to encourage use of a third party portal to deposit in a local institutional repository and thirdly the security and political objections to sharing sensitive data across institutions.

In the next section, recommendations are set out as to how these barriers might be broken down, in a stage by stage approach.

## 5. Recommendations

**Stage One:** Individual institutional repositories form a federation with the source and output repositories used by the researchers in their academic departments. Potential publications are made available via OAI for the domain-specific federated output repository to harvest for peer-review. If accepted, then part of the publishing process is that data has to be deposited or identified in the associated domain source repository. Any sharing of non-public data or documents is restricted to institutional colleagues. Object store in the institutional repository should be identifiable by domain, project, file type and format at least. Each object should have at least a title and each project at least Dublin Core metadata associated with it. Recommendations should be available in each domain regarding conventions, standards and metadata elements.

**Stage Two:** Each individual institutional repository then acts as a portal to itself and all the domain specific source and output repositories in its federation. As well as the assigned metadata, it would build indexes and ontologies by OAI harvesting the actual data, documents and publication from all the federated repositories. Additional Web 2.0, FOAF and Amazon-like features could then be added.

**Stage Three:** This stage would introduce the idea of a StORe subject or domain portal to the discipline specific elements of the federated repositories of distributed sites. Here Shibboleth authentication and registration would be required to allow sharing of non-public data and documents between colleagues from different institutions. One way around security issues would be the temporary copying of protected objects to the portal for download within a certain time period. This obviously offers wider coverage, more choice of source and output repositories and more scope for Web 2.0, FOAF and Amazon-like features. There could even be a common interface for deposit to individual institutional repositories. Additional features such as listings of forthcoming conferences, wikis, etc might encourage use.

**Stage Four:** The final stage would see the full generic solution implemented. Based on the model of the StORe subject portals, its coverage would be the entire federated institutional, source and output repositories that have adopted the recommendation and procedures outlined in Stage One. The aim is of course to encourage cross-disciplinary research, however, metadata mappings will have to be employed and even more additional features will have to be devised to encourage usage of such a universal portal.

## 6. Social Science Pilot Demonstrator

Based on the generic model outlined above, the pilot demonstrator will use a prototype federation comprising the UK Data Archive (source repository), the LSE's Research Articles Online (institutional output repository using ePrints) and a University of Essex prototype institutional output repository. Options for linking to a commercial publisher have also been explored but will not be included in the pilot.

The UKDA/University of Essex prototype institutional repository would be based on the Fedora open source digital repository software and will be compliant with the Reference Model for an Open Archival Information System (OAIS), having the ability to ingest and disseminate Submission Information Packages (SIPS) and Dissemination Information Packages (DIPS) in standard container formats such as METS and MPEG-DIDL.

The metadata assigned to objects will be a minimum set of elements from the Data Documentation Initiative (DDI) XML standard for technical documentation describing social science data.

Fedora repositories are also fully conformant, as is the GNU ePrints system at LSE, with the interoperability framework defined by the Open Archives Initiative Protocol for Metadata Harvesting. Hence the central indexing system will be able to harvest the institutional repositories of both the University of Essex and the LSE and build up indexes using Lucene, a high performance, scaleable, cross-platform search engine.

The deposit and access functions of the demonstrator system will require:-

1. Users to register in order to share digital objects.
2. Users, once authenticated, to be able to upload, modify or delete his/her digital objects.
3. The digital objects to be of a recognised format.
4. That metadata must be assigned before upload into his/her allocated space on his/her institutional repository is allowed.
5. Users to set permission to allow or restrict access to other organisations, groups or individuals.
6. A facility to submit the digital object to a source and/or output repository for their acceptance.

The search function of the system will require:-

1. A simple Google type search, with Boolean operators and wildcard functionality with the addition of advanced searches on selected areas of the metadata.
2. All digital objects to be listed in relevance order with a facility for sorting and filtering.
3. Users selecting a particular digital object to authenticate before access can be requested or granted.

The system will also allow reports to be generated, such as the total number of objects deposited or submitted in a period categorized by format. Certain information will be automatically added such as file size and dates of deposit.

The following diagram shows how the system will fit into the JISC information environment, with the Essex Fedora institutional repository acting as a portal to itself and its federated source and output repositories, as described in the generic recommendation Stages One and two above.

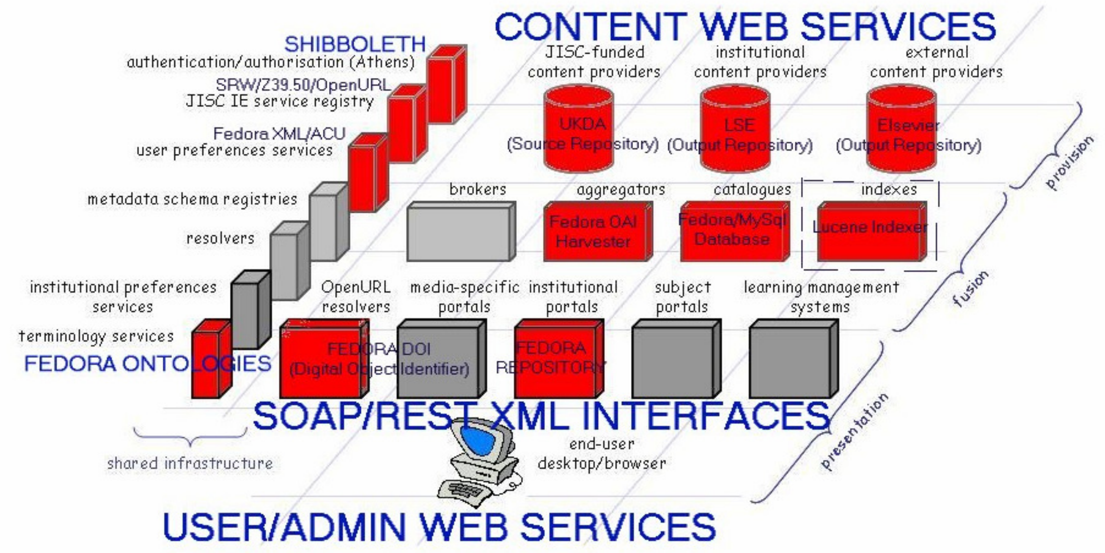

Figure 1: The pilot demonstrator within the JISC information environment

The following diagram shows the component features of the FEDORA system and how they utilises protocols, formats and standards that are used within the JISC information environment. The next release of Fedora should also contain a plug-in for the Lucene indexer that the prototype plans to adopt.

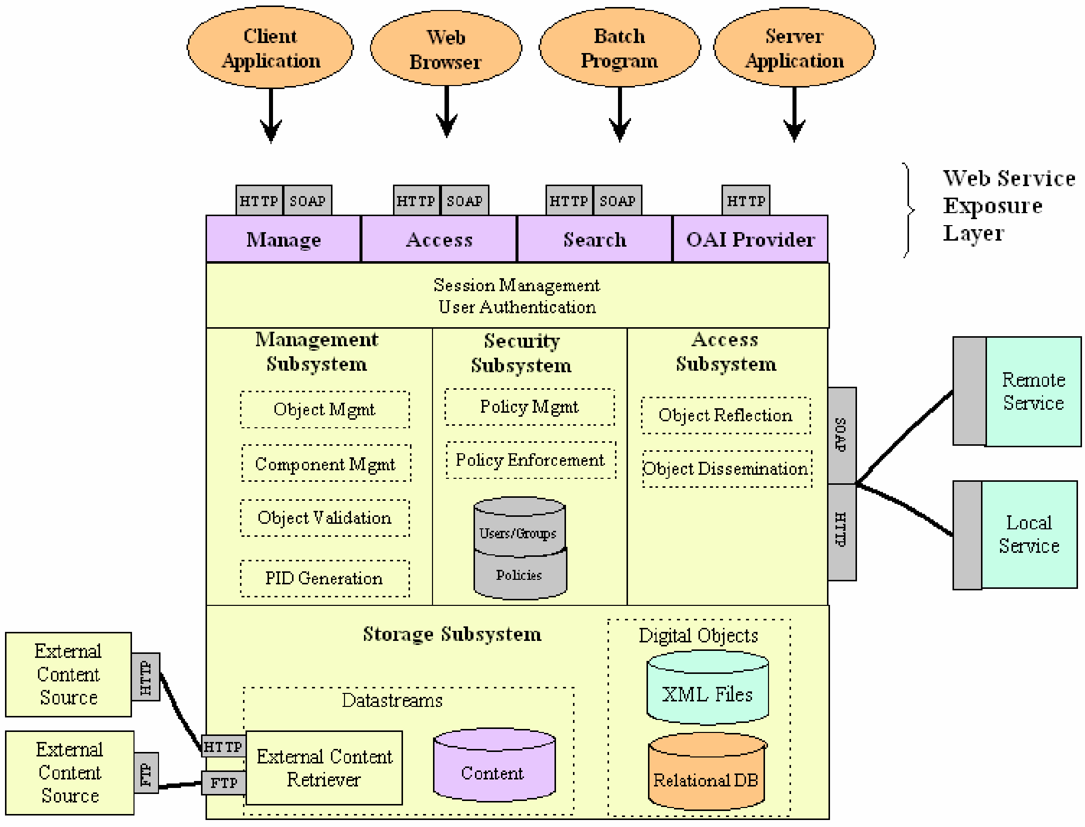

Figure 2: Features of the FEDORA system

## 7. Discipline Variations

A comparison review of the discipline-specific surveys is available at http://jiscstore.jot.com/BusinessAnalysis. It revealed that there is common ground in terms of a need for two way links between raw data repositories and academic publication repositories. Such links were considered useful by participants in the surveys and interviews across the disciplines and potential obstacles to sharing of data in such a way were also generally consistent. Noticeable variations in the way that data are gathered, formatted, allocated metadata and subsequently shared (both between disciplines and within disciplines) were noted, and this needs to be taken into consideration when establishing a Source to Output repository interface. It is likely that the discipline-specific requirements will result in a need for customisation of a generic Source to Output model. The disciplines investigated were Archaeology, Astronomy, Biochemistry, Biosciences, Chemistry, Physics and the Social Sciences. ***Note:*** *This section of the business analysis does not include biochemistry-specific reference as the individual report for that discipline has not been completed.*

There will be discipline variations to be managed at all four stages of the StORe portal development. Individual institutional repositories will have different federations, file types and formats, and will use different metadata standards. For example, FITS (Flexible Image Transport System) is the agreed standard for data analysis, transmission and archiving within the Virtual Observatory framework and is a vital part of the process for astronomical data storage and transmission. Within the biosciences the extensive capturing of metadata, such as the detailed description of the experimental conditions, is particularly important. However, this type of metadata is usually not provided in any standard format. In archaeology, where the use of standardized word lists and thesauri is common, the assignment of metadata is perceived as time consuming and complex.

There is a particularly wide range of data types and formats used in physics, and it is not uncommon for physicists to write their own programs to deal with these varying types. Also in chemistry, there are many variations in the data produced, and in its recording and storage, requiring the use of discipline specific software. Similarly in the biosciences source data might not be easily readable, but could instead require the use of a specific process or program to extract the information. This is because of the wide range of data produced and used, including spectrograms and videos, images, drawings/plots, raw data and gene/protein sequences, electrophoresis and micro array image data.

Another variation will be in the size of objects deposited. For certain domains, data interpretation, manipulation and methodology will be as, if not more, important than the raw data itself. Issues surrounding the actual mode of access to data also differ across the various disciplines.

Many physicists expressed a concern about the stage at which data should be made available, preferring towards the end of a particular analysis. In physics and archaeology some data is owned by a collaborative group rather than an individual, so there will be special requirements to be met when granting access to others. Archeologists tend to produce highly complex data sets, and these are often but not always linked into a GIS (Geographical Information System) which forms part of the way that the information is stored and presented. However, they also expressed concerns over the illegal looting of archaeological sites consequent upon the identification of geographical locations.

Although a simple search might cross disciplines, more advance searches would more likely be domain specific, with the resulting hit list, relevance ranking and sorting being different for each discipline. Consequently, both subject and global portals will require different incentives and Web 2.0 features for each domain. In addition, the access mode in certain disciplines should allow for the quick downloading of data sets so they can be processed with specific software.

Physicists liked the idea of linking from individual plots or figures in a publication to the data from which it was derived and recognised the benefits it would bring. In archaeology the linking of repositories was seen as a way of enabling more efficient scrutiny of the methodology applied and the research process. For researchers in astronomy usage figures obtained from the portal are likely to be useful in garnering additional funding or support.

The need for comprehensive and current resources in certain hard sciences was expressed with a demand that the whole process of deposit, linking, searching, sharing and access should be made as simple as possible. In the biosciences a registration process would be looked upon as a barrier and could prevent some researchers from using the portal. However, the submission of data to source repositories, such as GenBank, prior to the submission of a publication in a scientific journal, is already mandatory in this discipline.

It should be noted that the social science pilot demonstrator will represent one discipline variation within the overall generic solution and will also prove a useful starting point for all other disciplines.

## 8. Bibliography

‘StORe Project Web Site’ at http://jiscstore.jot.com/WikiHome for survey phase cross discipline and individual discipline reports.

Heery, R. and Anderson, S. 'Digital Repositories Review' JISC Digital Repositories Programme Report, February 2005.
Heery, R. and Powell, A. ‘Digital Repositories Roadmap: looking forward’ JISC Digital Repositories Programme Report, April 2006.

Swan, A. and Awre, C. ‘Linking UK Repositories: Technical and organisational models to support user-oriented services across institutional and other digital repositories’ JISC Scoping Study Report and Appendix, June 2006.
‘Draft JISC Strategy 2007-2009’, July 2006 at http://www.jisc.ac.uk/draft_strategy0709.html

‘JISC Information Environment Architecture’ at http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/

Lyon, L. ‘eBank UK: Building the links between research data, scholarly communication and learning’ Ariadne Issue 38, July 2003.

‘eBank UK Final Report’ at http://www.ukoln.ac.uk/projects/ebank-uk/docs/report2005/report.doc

‘Resource Discovery Network’ at http://www.rdn.ac.uk/

‘The Directory of Open Access Repositories – OpenDOAR’ at http://www.opendoar.org/index.html

‘The Bioliterature Project’ at http://www.biolit.org/

Staples, T., Wayland, R. and Payette, S. ‘The Fedora Project: An Open-source Digital Object Repository Management System’ D-Lib Vol 9 No 4, April 2003

‘The Repository Metadata and Management project (RepoMMan)’ at http://www.hull.ac.uk/esig/repomman/

O’reilly, T. ‘What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software’ at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, September 2005

Lewis, S. ‘Final Report’ JISC Repository Bridge Project, June 2006

‘Reference Model for an Open Archival Information System (OAIS)’ at http://public.ccsds.org/publications/archive/650x0b1.pdf

'METS (Metadata Encoding and Transmission Standard)' at http://www.loc.gov/standards/mets/

'OAI Protocol for Metadata Harvesting' at http://www.openarchives.org/

‘Lucene: open-source indexing and search software’ at http://lucene.apache.org/

‘An Introduction to Persistent Identifiers’ UKOLN QA Focus Document at http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-80/

‘The Data Documentation Initiative’ at http://www.icpsr.umich.edu/DDI/