



## JISC Final Report

### Project: SHERPA DP2

<b>Project Acronym</b>	SHERPA DP2	<b>Project ID</b>	
<b>Project Title</b>	SHERPA Digital Preservation 2		
<b>Start Date</b>	01/03/2007	<b>End Date</b>	28/03/2009
<b>Lead Institution</b>	Centre for e-Research, King's College London (originally Arts and Humanities Data Service)		
<b>Project Director</b>	Stephen Grace		
<b>Project Manager &amp; contact details</b>	Centre for e-Research King's College London 26 - 29 Drury Lane LONDON, WC2B 5RL Tel: 020 7848 1970 Fax: 020 7848 1989 Email: stephen dot grace at kcl dot ac dot uk		
<b>Partner Institutions</b>	None		
<b>Project Web URL</b>	<a href="http://www.sherpadp.org.uk">http://www.sherpadp.org.uk</a>		
<b>Programme Name (and number)</b>	Digital Preservation and Records Management Programme		
<b>Programme Manager</b>	Neil Grindley		

### *Document*

<b>Document Title</b>	Final Report		
<b>Author(s) &amp; project role</b>	Gareth Knight, Preservation Officer		
<b>Date</b>	23/Apr/2009	<b>Filename</b>	sherpadp2_finalreport_v1.doc
<b>URL</b>	N/A		
<b>Access</b>	<input checked="" type="checkbox"/> Project and JISC internal		<input type="checkbox"/> General dissemination

### *Document History*

Version	Date	Comments
0.1	31/Mar/2009	Initial draft
1	23/Apr/2009	First version



## **SHERPA DP2: Developing services for archiving and preservation in a distributed environment**

Author: Gareth Knight  
Contact: [gareth](mailto:gareth.knight@kcl.ac.uk) dot knight at kcl dot ac dot uk  
Date: 23 April 2009

### **Table of Contents**

Acknowledgements .....	2
Executive Summary .....	3
Background .....	3
Aims and Objectives.....	7
Methodology.....	7
Implementation.....	12
Outputs and Results.....	14
Outcomes.....	15
Conclusions.....	16
Implications .....	16
Recommendations .....	16
References.....	17

### **Acknowledgements**

SHERPA DP2 (<http://www.sherpadp.org.uk/>) was funded by the JISC under the Digital Preservation and Records Management Programme. We would like to thank the programme manager, Neil Grindley, for his support during the project, and the various project partners for their support during the work. In particular, we would like to thank Colin Watt and Martin Velecky of the University of Edinburgh, Jacqueline Cooke at Goldsmiths College, Chris Awre, Richard Green and Simon Lamb at the University of Hull, Tamara Lopez, Richard Palmer and Paul Spence at the Centre for Computing in the Humanities (CCH) and several staff members at CERN. We would also like to acknowledge Sebastien Francois of the University of Southampton for his advice on the SWORD plug-in for EPrints.

## Executive Summary

The SHERPA DP2 project (2007-2009) was a two year project funded by the JISC under the Digital Preservation and Records Management Programme. The project was led by the Centre for e-Research at King's College London (formerly the Executive of the Arts and Humanities Data Service), which is working with several institutions to develop a preservation service that will cater for the requirements of a diverse range of digital resources and web-based resources. In summary, the project has the following objectives:

1. Extend and refine the OAIS-based Shared Services model created for the initial SHERPA DP project to accommodate the requirements of different Content Providers and varied collaborative methods.
2. Produce a set of services that will assist with the capture and return of research data stored in distributed locations, building upon existing software tools.
3. Expand upon the work processes and software tools developed for SHERPA DP(1) and SOAPI to cater for the curation and preservation of increasingly diverse resource types.

The project investigated the curation and preservation requirements of research data that is encoded as many different content types and made available using many different technologies in disparate locations. It is built upon the premise that an organization or department that has been tasked with the preservation of the collective of an institution must consider a range of different publication methods and technologies. The project builds upon the methodology developed during the first SHERPA DP project (2005-2007). The project defined a disaggregated service model, in which key components of the OAIS Reference Model are distributed between two parties:

1. A *Content Provider* that accepts research data and publishes it for use by academics and
2. A *Preservation Service Provider* that is responsible for normalisation and other activities necessary to ensure continued information access.

While the first project considered the curation of academic research papers and electronic theses and dissertations published through DSpace and EPrints, SHERPA DP2 extends the curatorial service in three ways:

1. It interacts with a large number of different technical systems operated by Content Providers, including non-repository based systems;
2. It defines different types of interaction between a Content Provider and Service Provider, utilising a combination of internal and web accessible systems;
3. It is built upon a preservation strategy that caters to a larger range of content types;

An overview of the over-arching methodology adopted for the project is outlined within this report. This is supported by a series of reports and several case studies that indicate the approach taken for each project partner.

The provision of a preservation service, offered by a third-party organisation or a central department within the institution represents an effective method to curate and preserve research data stored by an institution. The SHERPA DP2 project offers a practical example of the type of services that a Preservation Service Provider may offer to a Content Provider and the scenarios in which they may be provided.

## Background

Research data represents any type of digital information created in the process of undertaking research. It is produced as a result of individual or group investigation to achieve a specific objective. The digital research being created is increasingly varied in its scope and scale, due to the increasing ambition of research and developing capabilities of contemporary creation tools. Rather than being represented simply as one or more research papers published in a journal and archived in a digital

repository, storage and publication systems may contain research data rendered in several different forms - text, datasets, still images, moving images, sound and interactive resources - that each represents specific and often unique knowledge.

In recent years there has been increasing emphasis upon the need to curate and preserve digital research, through the establishment of the Open Data and Research Data Management agenda. The reasons for maintaining digital research were summarised succinctly by the recent JISC-funded Keeping Research Data Safe report (Beagrie et-al, 2008) into four key objectives. The desire to: [1] protect earlier investment in research; [2] preserve opportunities to use the data for future research; [3] promote the work of the institution and the researcher; and [4] support the research and learning process. Although it is well recognised that research data has continued value, efforts to curate and preserve it have so far been limited and fragmented within the institutional environment. Further work is necessary to develop methods to archive, curate and preserve research data that work in collaboration with existing systems for its creation and publication.

### Data creation and publication process

To understand how research data is published and distributed, it was considered beneficial to understand the process of its creation. The research creation process has been touched upon by a number of studies (Martinez-Urbe, 2008; Beagrie et-al, 2008) that have examined the topic from a data lifecycle and information lifecycle perspective. These recognise that the process of research development progresses through several phases through its lifetime, from creation to (potential) deletion. Additionally, many of these studies recognise that the process of research creation in a digital environment is diverse and complex, varying between different types of research data. On the basis of the review of secondary sources, it was recognised that researchers performed several key activities and stored research data in a variety of sources, as indicated by figure 1.

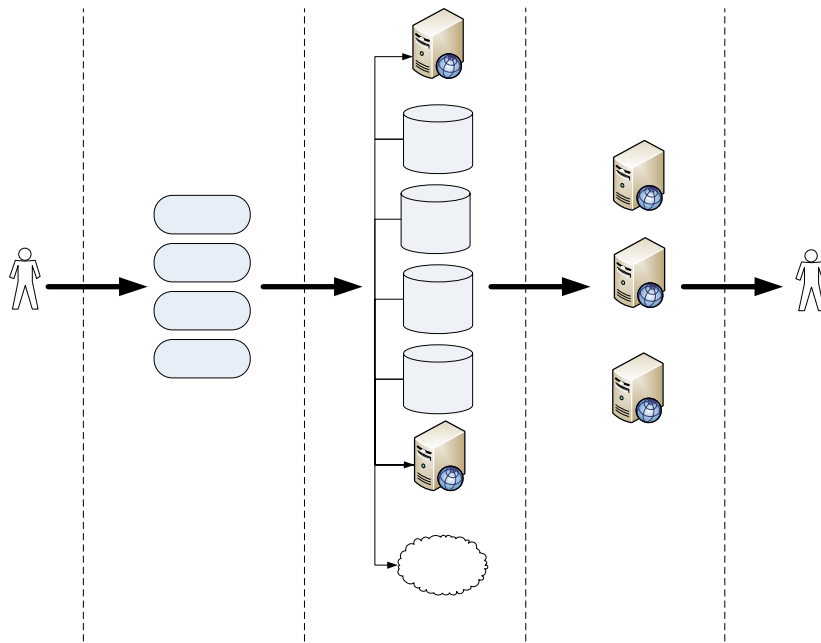


Figure 1: Process of data creation and management using different technologies and services

As an outcome of the investigation, the project team recognised three factors that influence the influence the method in which data is stored and published and, by extension the activities that must be performed by the project to capture data:

1. Digital research is varied in its scope and scale, due to the increasing ambition of research and developing capabilities of contemporary creation tools
2. Digital research may be created and manipulated using tools that are operated on the client's host machine, within the institution's local system, or through Internet accessible services that may have some influence upon its of storage and publication.

3. Digital research is increasingly being published using diverse services that provide facilities for handling specific content types (e.g. audiovisual, still images, documents), access methods, or other distinctions.
4. Digital research is often created in fulfilment of research grants allocated by funding bodies, which may impose specific requirements on how data is stored and published;

In recent years, the technological environment has moved away from the concept of a single 'monolithic' storage system (e.g. a single institutional web site) in which all research data is stored and published to a disparate approach, in which data may be stored in several locations. Research data may be stored on a systems may be maintained by an academic department of which the researcher is a member (e.g. a departmental repository), a central service provided by the institution for all its researchers (e.g. an institutional repository), a third-party service offered to researchers within a specific subject community (e.g. arxiv.org), as well as content-specific services that are available to all types of user (e.g. YouTube, Flickr). As an outcome, components of a research project may be stored by different organisations on physically distributed systems that possess diverse policies and procedures on data management requirements.

### **Locating research data**

The need to ensure that research data continues to be locatable on internal systems and public systems has been recognised by several research projects and institutions as a key requirement in ensuring its long-term accessible. The Data Audit Framework study in its analysis of data management risks notes that information loss may occur if staff members are unable to locate data located on different storage systems. Similarly, the recent Keeping Research Data Safe report (Beagrie et-al, 2008) indicates that the maintenance of a "*complete and accurate scholarly record*" (p16) is essential to maintain the value of research data over time. It must be possible to obtain a copy of data in its original form in the short-term, in order to perform further actions to ensure access to research data in the long-term. Institutions, it may be observed have taken several approaches to ensuring that research data is locatable. The DISC-UK DataShare has established data repositories within their organisation and worked with departments to move their data into the managed system. Other institutions, such as CERN and Goldsmiths College create a record in their digital repository that references the location of research data stored and distributed through third-party systems located internal to the institution and externally. As a result, the network structure of many institutional repositories has begun to exhibit characteristics of a web portal: providing a pathway to diverse types of content that is stored on many different systems<sup>1</sup>.

The reference to third-party storage represents an effective strategy for institutions to publicise research data that is stored in the researcher's working environment and/or cannot be managed in the repository itself for a variety of reasons. However, the institution is reliant upon the third-party to continue to maintain and deliver the resource, which may represent a risk in the long-term. Although research data may in its entirety form a larger research project, components stored on systems managed by different organisations may be subject to differing data management strategies. As a result, there is the risk that data may undergo conversion that results in loss or damage to one or significant properties that a specific Designated Community require to use the data. Alternatively, if data located on a third-party system is moved to a different location without updating the identifier or is removed entirely, the relationship between components of research data may be broken and the scholarly record is rendered incomplete.

### **Archiving and preserving research data**

To ensure that research data remains accessible in the long-term, it is necessary to perform curatorial and preservation action. This may prove to be problematic for many institutions. As noted in the final report for the first SHERPA DP project (Knight & Anderson, 2007), repository staff are frequently required to focus upon the immediate priorities of embedding their digital repository in the wider institutional infrastructure and repository population and may not have sufficient time or resources to make long-term data management decisions. Additionally, they will be unable to take an active role in the curation and preservation of research data stored on external systems.

---

<sup>1</sup> It should be noted that digital repositories do not fulfil all of the characteristics of a web portal: they do not perform any type of standardisation or present information from multiple sources through a single interface.

The SHERPA DP2 project builds upon earlier work in the field of research management by establishing a method for institutions to take a uniform approach to the management of research data across technical systems and institutions. It advances the research data management agenda by asking and addressing the following questions:

1. What services are required to capture research data of an institution that is stored and published through disparate technical systems?
2. What services are required to curate and preserve research data to ensure it remains accessible?

The project is built upon the premise that the *data management requirements of an institution extend beyond the confines of a digital repository*. Instead, preservation services must be able to interoperate with diverse types of technical systems and curate a wide variety of content types. To address the stated questions, the project team has developed a means to maintain a complete (or semi-complete<sup>2</sup>) copy of research data published by a researcher published through different services and made available using different technologies. The process is composed of a semi-automated workflow and combines the capabilities of several data capture tools to create a manifestation of data obtained from third-party systems on a local storage facility. The captured data is managed in a trusted environment operated by a third-party Preservation Service Provider, but which could equally be maintained by a centralised Archives unit or other department within an institution, where it undergoes curatorial and preservation activities, as required.

The approach has a number of benefits for an institution:

1. It maintains a record of research outputs of an institution, department, or other subsidiary that is not reliant upon a third-party that has no direct investment in maintaining the research data or is stored on an external web accessible resource that is beyond the control of an institution;
2. It enables a uniform approach to curation and preservation of data that takes into account the significant properties of research data, beyond the lifecycle of existing encoding formats and content delivery technologies;
3. It provides an alternative method to the population of a preservation repository with research data that avoids disruption to existing practices of research creation, which may result from manual deposit of frequently updated data,

The project team worked with several partners that store and publish real data: CERN, based in Switzerland; the Centre for Computing in the Humanities (CCH) at King's College London; Goldsmiths College, the University of Edinburgh; and the University of Hull which publish data through a combination of digital repository and content management systems and also frequently refer to third-party systems. The institutions use one or more of several technical systems, including CDS Invenio, DSpace, EPrints, Fedora Commons, the Subversion versioning system, as well as general publication methods, such as MySQL-driven dynamic web sites and static web sites. However, the methodology may be extended by the Centre for e-Research and other projects to consider other types of publication systems.

### **How the project builds upon existing work**

The SHERPA DP2 project builds and complements the work performed in several studies, most notably the Data Audit Framework (Jones, Ross & Ruusalepp, 2009) and the Keeping Research Data Safe report (Beagrie et-al, 2008). We have also been able to learn from other projects, including the JISC-PoWR project which was funded to examine the management requirements of web-based resources and the JISC Source project.

As part of the JISC Digital Preservation and Records Management Programme the project has worked closely with the SOAPi project (Service Oriented Architecture for Preservation and Ingest of Digital Objects) to integrate and extend the ingest framework to the requirements of the SHERPA

---

<sup>2</sup> Data may be captured if it has been published using specific publication methods (e.g. a static web site, digital repository) and the URI can be located and accessed by the capture tools.

DP2. The project has also liaised with staff working on PRESERV2 at the University of Southampton and University of Oxford to understand the use of OAI-ORE to transfer data between repositories.

## Aims and Objectives

The SHERPA DP2 proposal notes the following five objectives:

1. Extend the SHERPA DP OAIS based distributed preservation model to accommodate different types of institutional repositories and different collaboration methods, and investigate other options for provision of distributed preservation services.
2. Investigate and develop tools to transform repository content (digital objects) as base64 encoded bitstreams for placing inside METS packages. The tool will also create the basic METS package and ensure that the encoded bitstream is appropriately located within the package.
3. Investigate and assess other methods for connecting to digital repositories and downloading repository metadata and content.
4. Refine the Sherpa DP set of protocols and software in order to interact with institutional repositories using a wider range of repository software applications and with a broad range of digital object types.
5. Amend, update and expand as appropriate the Digital Preservation User Guide produced by the original Sherpa DP project to take account of the outcomes and lessons of the Sherpa DP2 project.

In the early stages of the project it was identified that the encoding of digital objects as base64 would not provide any benefit to the project and the component was dropped. Although the use of base64 enables a digital repository to store data and metadata as a single file for capture, it dramatically increases the data size, which may increase the time required to capture each object. At the same time, objectives 1, 3 & 4 were also revised and refined to provide a more accurate description of the work being undertaken:

1. Extend and refine the OAIS-based Shared Services model created for the initial SHERPA DP project to accommodate the requirements of different Content Providers and varied collaborative methods.
2. Produce a set of services that will assist with the capture and return of research data stored in distributed locations, building upon existing software tools.
3. Expand upon the work processes and software tools developed for SHERPA DP(1) and SOAPI to cater for the curation and preservation of increasingly diverse resource types.

## Methodology

The methodology for the SHERPA DP2 project is built upon the disaggregated service model developed for use in the first SHERPA DP project (2005-2007). The model establishes a co-operating archive relationship, in which the requirements for OAIS compliance in a digital repository are fulfilled by multiple parties that, working in co-ordination take responsibility for different activities within the OAIS workflow. The model identifies two key players:

1. A *Content Provider* that accepts research data and publishes it for access and use in the short-term;
2. A *Preservation Service Provider* that performs activities necessary to ensure that digital resources remain accessible and usable over the long-term.

The SHERPA DP2 project extends and enhances the disaggregated service model in two ways: First, it recognised that a Content Provider may be technologically diverse, hosting digital resources on a combination of physical and virtual machines that is geographically distributed and delivered through one of several diverse interfaces, including digital repository software, content management systems, dynamic database systems and static web sites.

Second, it recognised that a Content Provider and Preservation Service Provider may themselves obtain support from other Service Providers to fulfil its aims and objectives.

Due to the influence of these factors, the co-operating archive model has evolved from a dual node system in which a Content Provider and Preservation Service Provider is represented to a web-like system that encapsulates different storage environments. To illustrate the interaction, figure 2 indicates the potential relationships established between technical systems operated by a Content Provider and figure 3 indicates the interaction between the Content Provider systems and a Preservation Service Provider.

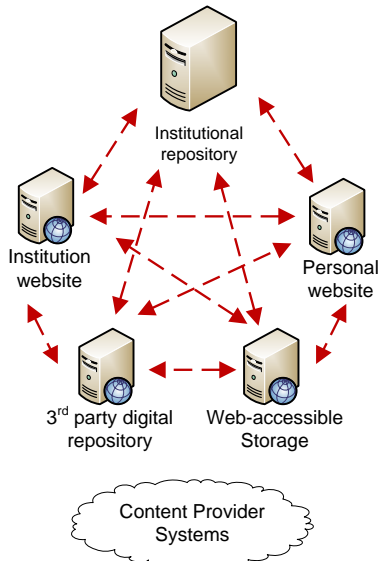


Figure 2: potential relationships established between technical systems operated by a Content Provider

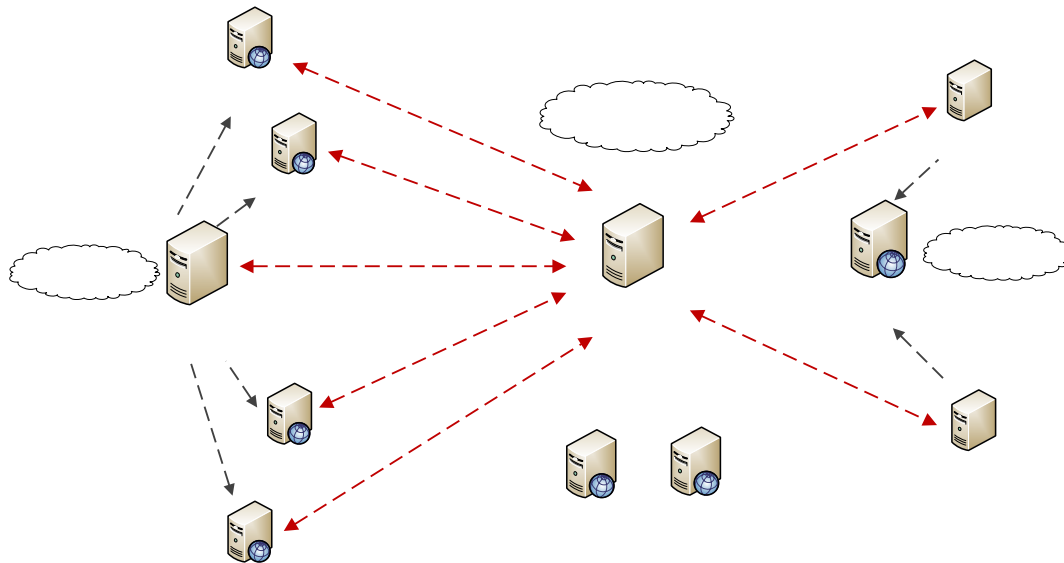


Figure 3: Interaction between the Content Provider systems and a Preservation Service Provider

While the first project considered the curation of academic research papers and electronic theses and dissertations published through DSpace and EPrints, SHERPA DP2 extends the curatorial service by interacting with a large number of different technical systems operated by Content Providers, including non-repository based systems and defining different types of interaction between a Content Provider and Service Provider, utilising a combination of internal and web accessible systems;

The investigation undertaken by the project encapsulated three broad activities:

1. Enumeration of services that a Preservation Service Provider might provide to a Content Provider built upon a scenario-based approach. The analysis resulted in the specification of three services that the SHERPA DP2 PSP would provide.
2. Analysis of research data stored or referenced by Content Providers, which led to the development of a preservation strategy;
3. Analysis of technical systems to determine the methods available to monitor and capture data from a Content Provider and subsequently resubmit data it at a later date.

Although the activities are described in the order stated above in this report, they were performed in reverse order during performance of the project to ensure that the development of a Preservation Service Provider would be practical.

### **1. Scenarios and services**

The first component of the investigation was concerned with the definition of a set of services that a Preservation Service Provider may provide to a Content Provider. These were expressed as a set of distinct scenarios in which a PSP was requested to perform data management activities in response to specific needs and requirements. Eight scenarios were examined:

1. *Storage failure scenario*: The scenario is based upon the premise that a Content Provider has previously established a contract with a PSP to provide an off-site backup and, as a result of a data storage failure has lost data. As a result, it has requested that the Preservation Service Provider provide a copy of the most recent version of the data collection, in part or in its entirety.
2. *Data replacement scenario*: The scenario is based upon the premise that a Content Provider has previously established a contract with PSP to provide an off-site backup and wishes to obtain an earlier version of research data that is stored by the PSP, but has been removed by the Content Provider themselves. In response, it has requested that the PSP provide a copy of the data that was captured at a set date and time. To perform the scenario, the contract between the CP and PSP must indicate the number of versions of frequently changing data that the PSP will store.
3. *Data audit scenario*: The scenario is based upon the premise that a Content Provider requires a third-party to perform an audit of the data objects that it stores and/or publishes on different servers, e.g. to establish conformance of data to format specifications, web standards, and so on.
4. *System switch scenario*: The scenario is based upon the premise that a Content Provider is planning to switch from System A to System B and requires data to be reformatted into a different form, e.g. an institutional repository wishes to move between DSpace, EPrints and Fedora. Potential changes may include metadata packaging (e.g. METS or MPEG21), metadata crosswalk (e.g. MODS to MarcXML), creation of additional metadata (e.g. RDF or PREMIS) or format conversion. The scenario may be initiated through a short-term contract to perform specific activities, or as part of long-term contract (e.g. a 5 year contract to perform specific data management services).
5. *Data enhancement scenario*: The scenario is based upon the premise that the Content Provider requires data to be enhanced in some way. For example, the Preservation Service Provider is contracted to improve the quality of metadata or creation additional data. The scenario may be initiated through a short-term contract to perform specific activities, or as part of long-term contract (e.g. a 5 year contract to perform specific data management services).
6. *Curation scenario*: The scenario is based upon the premise that a Content Provider requires preservation metadata to be generated for some or all of the data objects for which it is responsible. The scenario may be initiated through a short-term contract to perform specific activities, or as part of long-term contract (e.g. a 5 year contract to perform specific data management services).

7. *Preservation scenario*: The scenario is based upon the premise that a Content Provider requires data to be converted into a form that may be more easily managed, e.g. through normalization to a set of common encoding formats and creation of preservation metadata. The scenario may be initiated through a short-term contract to perform specific activities, or as part of long-term contract (e.g. a 5 year contract to perform specific data management services).
8. *Migration scenario*: The scenario is based upon the premise that the requirements of a Designated Community has changed over time and existing dissemination formats are at-risk of becoming inaccessible or unusable in future software. The Preservation Service Provider is contracted to produce new manifestations of each object suitable for access and use by the Content Provider's Designated Community. The scenario may be initiated through a short-term contract to perform specific activities, or as part of long-term contract (e.g. a 5 year contract to perform specific data management services).

The requirements of each scenario are diverse and varied; the ability of a Service Provider to meet requirements for each scenario is influenced by the existing technical services in place and those that can be developed for use. Although each requirement is represented as a distinct scenario, there is the potential that a Content Provider may wish to perform activities from two or more scenarios. Due to the short timescale of the project, the project team selected four scenarios (1, 2, 6, 7 and 8) and examined the requirements of each scenario. The cross-analysis resulted in the definition of three types of services that could be provided in the SHERPA DP2 project:

1. *Archiving service*: The Preservation Service Provider stores a complete or partial backup of data in an offsite location. The backup would cover different versions of data and metadata stored by the Content Provider. The time period in which each version would be stored would require negotiation between the Content Provider and Preservation Service Provider.
2. *Migration service*: The Preservation Service Provider stores a complete or partial backup of data in an offsite location and generates replacement Dissemination copies of data objects in the event that the current manifestations is at-risk of being rendered inaccessible or unusable in future software, as a result of changes in the Designated Community or the computing industry in its entirety.
3. *Preservation Service*: The Preservation Service Provider stores a complete or partial backup of data in an offsite location and produces manifestations of a digital object that can be more easily managed, e.g. through normalization to a set of common encoding formats and creation of preservation metadata.

Similar to the previously outlined scenarios, the service types are not mutually exclusive; a Content Provider may require a combination of different activities to be performed.

## **2. Analysis of research data and development of preservation strategy**

The second component of the investigation was concerned with the development of a preservation strategy to manage research data stored by Content Providers. To achieve the objective, the project team examined the type of data that could be obtained from the technical system and spoke with each project partner to identify notable data types (Knight, 2008). The format in which data is encoded and stored was recognised as a key factor that could enable or inhibit access to digital information in the long-term. They define the rules used by application software to convert bits (the fundamental unit of digital data) into meaningful information that can be viewed and manipulated by a user.

In the distributed environment outlined in the disaggregated service model, it was recognised that the type of data that the Preservation Service Provider would receive to curate and preserve would be influenced by a number of factors:

1. The collection remit of the Content Provider, in terms of the content types (e.g. research papers, still images, moving images) and data formats (e.g. PDF, MS Word) that it accepts;

2. The management activities that Content Provider staff performs between the initial deposit and publication stage of the workflow (e.g. convert MS Word to PDF, convert TIFF to JPEG).
3. The form in which data is made available (e.g. the use of Fedora disseminators to produce specific data output) which may, for some case studies influence the data types that were captured by the Preservation Service Provider

To assess the potential longevity of data formats the project performed a risk assessment. The methodology developed by Arms & Fleischhauer (2005) for the Library of Congress was considered to be the most effective at the time that the assessment was performed, although the project would likely have utilised the PLANETS Objective Tree methodology (Kulovits, 2008) in preference if it had been available. The methodology specifies seven criterion by which the suitability of 15 encoding formats and variants were evaluated. On the basis of the analysis, the project classified each data type and variant into one of three categories – preservation of the bitstream, information content or experience – based upon the type of activity it could perform and specified the preservation formats in which it would store data.

### **3. Technologies**

The third area of investigation which was performed in the early stages of the project was to analyse the technical systems in use by Content Providers to store and deliver digital resources. The technical review examined the overall functionality offered by each technical system - Fedora, CDS Invenio, DSpace, EPrints, static web sites and dynamic web sites - with specific focus upon three aspects:

1. Facilities available to monitor one or more capture targets provided by Content Providers;
2. Technology required to capture digital resources from each capture target;
3. Facilities available to submit data to the Content Provider;

As a result of the investigation of technical systems, the project team enumerated a list of data monitoring, capture and submission technologies that it would be beneficial to support. The project investigated each of the technologies, but was able to provide only limited support for six of the seven technologies (see implementation)

### **Questions that must be addressed by a Preservation Service Provider**

The methodology utilised for the project was reviewed and revised at several stages throughout the project, extending it to consider new scenarios and provide a more granular level of analysis. On completion of the project the project team had identified a set of questions that a Service Provider must address prior to an agreement being made with a Content Provider to archive, curate and preserve research data:

#### *Strategy:*

1. What services does the Content Provider require and what are the associated activities that must be performed?

#### *Monitoring:*

2. What facilities are available to monitor a resource for updates or changes?
3. What type of updates or changes should be monitored?
4. How frequently should monitoring activities be performed?

#### *Capture*

5. Where is the data for capture located?
6. How is it distinguished from data that should not be captured?
7. What tools, services or other functionality are required to capture data?
8. What information is required to access data for capture (e.g. passwords, VPN access)?
9. Can the captured data be validated for unexpected changes that occur during transfer (e.g. checksums)?
10. Are any restrictions established on the ability to capture a specific type or amount data (e.g. IP address ban)?

#### *Curation and preservation:*

11. What type of data must be curated and preserved?

12. What type of metadata exists to support the data?
13. Are any restrictions established on the ability to perform curation or preservation activity (e.g. technical restrictions, such as password protection, legal restrictions that prohibit actions, such as format conversion)

*Submission:*

14. What are the scenarios for which original or newly created data must be provided to the Content Provider?
15. How should data be transferred to the remote system?
16. What are the requirements to submit data to the remote system (e.g. usernames, passwords, use of specific protocols, specific packaging formats that must be used for submission)?
17. Are any restrictions established on the ability to transfer specific type or amount data (e.g. a deposit client is able to deposit specific data types; a digital repository is able to store only limited types of specific metadata formats)?

Practical experience on issues that have been encountered may be found in the case studies performed for each project partner (Knight 2009a-e).

## Implementation

To implement the technical architecture necessary to support the project, the project team began by defining a basic workflow of the activities that must be performed at key stages to capture, ingest and submit. The work was seen as particularly important for the Centre for e-Research (or AHDS Executive at the time of initial funding) – the department was undertaking several Fedora-based projects that had preservation-related components. The process specified how components developed by one project would relate and influence components developed by other projects, which was seen as valuable to ensure that each project had a clear remit in its development objectives and avoid duplication of effort. Although the work was initially quite simple, it built upon earlier work in SHERPA DP (Bodhmaghe & Hedges, 2006), culminating into a workflow modelled using a subset of the Business Process Modelling language (BPM). The language provided particularly useful, enabling the team to recognise the triggers to initiate each step, the requisite pre-conditions necessary to perform the task and the post-conditions (Knight, 2009f). The project team attempted to build upon the workflow management investigation performed for the SOAPI project, utilising jBPM to design the capture and submission workflows. However, similar to the SOAPI project, we encountered a number of low-level errors that delayed the implementation. After further investigation, it was agreed with the SOAPI project that jBPM would not be used and the workflow was re-implemented in Java code.

The management workflow for the SHERPA DP2 project specified five key stages: [1] capture, [2] ingest, [3] obsolescence monitoring [4] format conversion and [5] submission, of which the software developer implemented four stages (1,2,4 & 5) in a semi-automated workflow for selected types of data types and technical system. The following section outlines the work performed for each stages of the management workflow.

### Capture

The Capture event encapsulates actions associated with the transfer of data from the Content Provider to the Service Provider. It is a component of 'Pre-ingest', an informal term initially used in the CEDARS project (Stone & Day, 1999) and subsequently referenced in the Trusted Digital Repositories report (RLG, 2002) that refers to activities that take place prior to Ingest in the OAIS Reference Model. The technologies supported by digital repositories has grown increasingly diverse during the project funding period, providing diverse methods to monitor and capture research data at an increasing level of granularity. Each Content Provider provided different methods of publishing data, utilising different technologies that had to be considered and handled appropriately. The project team evaluated several capture methods:

1. *OAI-PMH*: The OAI Protocol for Metadata Harvesting provides a standard format for querying a digital repository at the system-level for new and updated items. OAI-PMH has been widely adopted in the content management and library science world for use in digital repositories, but has only limited acceptance in other technological systems. The digital repositories examined in the SHERPA DP2 project utilise OAI-PMH to publish metadata conformant to several formats, including simple Dublin Core, qualified Dublin Core (DCTerms), UKETD (a Dublin Core

application profile for electronic theses and dissertations), METS and MPEG21-DIDL (although the latter is primarily a container for simple DC elements).

2. *Web feed*: A web feed is a data format created for providing users with frequently updated content. A Content Provider publishes a web feed, allowing users to subscribe and receive notification of updates. A Content Provider may publish a web feed in one of several broadly similar, but incompatible formats, such as RSS 1.0, RSS 2.0 and Atom. Web feeds are supported by a subset of digital repositories and web sites.
3. *Web crawl*: A web crawl refers to the process of gathering information from a web site and recreating it on a local storage area. A crawler is useful for gathering specific types of content from a web site that does not provide other publication methods.
4. *Database backup*: Database backup refers to the process of writing a copy of an active database – the table structures and contents – to a static file. A database backup may be used as the basis on which data enhancement or preservation action is performed on frequently changing content and may be restored in the event of data loss.
5. *Version Control check in/out*: A version control system tracks changes to a set of files. Many version control systems are designed around an client-server architecture – a server stores the current (and often previous) version of a data file which is “checked out” by a client. The client modifies one or files in the project and “checks in” the modified data. A VCS may contain the digital master of a resource from which a dissemination manifestation is produced. The Centre for Computing in the Humanities (CCH), one of the project partners uses the Subversion versioning system.

The project utilised OAI-PMH, web crawl, database backup and version check-in/out within the capture workflow and performed some experimentation with the use of web feed capture. However, it was unable to implement support for OAI-ORE or other technologies during the allocated time period.

To capture research data located on different systems it was necessary to adopt a combination of capture methods (e.g. OAI-PMH and web crawl for the CERN Document Server, database backup and version control check-out for the CCH resources). The experience of capturing third-party content for SHERPA DP2 may be contrasted against that gathered during the earlier funding period: the first SHERPA DP project was able to capture the majority of digital assets referenced by a digital repository using OAI-PMH and direct HTTP transfer. In contrast, SHERPA DP2 found that digital repositories would reference data on several systems located within the institutional domain and externally. The referenced resources were more diverse and complex in comparison to data assets captured in the earlier project. A record may contain multiple files that represent different components of the research (e.g. a research paper and a set of images) or manifestations of the research in different formats. The diverse form of digital research data encountered in SHERPA DP2 may be indicative of the changing role of a digital repository within an institution - rather than represent a monolithic storage system to store and publish the research output of an institution, many institutional repositories are demonstrating characteristics similar to a web portal, providing a gateway to other systems through which research data is delivered.

### **Ingest**

The ingest event encapsulates activities associated with the preparation and ingest of data collections obtained from a Content Provider into the digital repository. The objective of the Ingest workflow, expressed using OAIS RM terminology is to take the Submission Information Package (SIP) and produce an Archival Information Package (AIP) and Dissemination Information Package (DIP). It encapsulates activities such as virus scanning, checksum generation and validation, format identification, characterisation and format conversion (for preservation and dissemination).

The technical implementation of Ingest activities was developed for the JISC-funded SOAPI project and subsequently adopted and extended for use within SHERPA DP2. Further information on the workflow may be found in the SOAPI Final Report (Hedges, 2009)

### **Submission**

The third key event in the Preservation Service model is the process of submitting data collections that the Preservation Service Provider has curated and preserved to the Content Provider for subsequent re-ingest into their workflow. As noted, a request to resubmit data may be initiated for several reasons in different scenarios. In total, the project tested the re-deposit of data in accordance with four scenarios.

1. *Original data and original metadata scenario*: The data files and metadata remain as-is, in the format that it was captured from the institutional repository.
2. *Original data and reformatted metadata*: Metadata is reformatted as a METS package containing Dublin Core descriptive metadata, PREMIS Object technical metadata, PREMIS event metadata and RDF relationship metadata; data files remain as-is, in the format that it was captured.
3. *Normalised data and original metadata*: Data files are converted into a format appropriate for preservation and metadata remains as-is. In the absence of appropriate tools to normalise PDF to PDF/A, the project exported PDFs as a text file and set of PNG images for diagrams and a set of TIFFs for each document page.
4. *Normalised data and reformatted metadata*: Data files are converted into a format suitable for preservation and dissemination and metadata is enhanced and stored in a packaging format. Similar to the third scenario, PDF documents were resaved as a text file and set of PNG images for diagrams and each page converted to a set of page TIFFs; metadata was exported as a METS package containing Dublin Core descriptive metadata, PREMIS Object technical metadata, PREMIS event metadata and RDF relationship metadata.

For each scenario, the project team selected five item-level records from each Content Provider on which to perform the data conversion (20 items in total). The items were subsequently submitted to the institutional repository from which they were originally obtained for ingest, storage and publication.

The project utilised several technologies to transmit data to the Content Provider's storage system:

1. *SWORD*: SWORD (Simple Web-service Offering Repository Deposit) is an application profile of the Atom Publishing Protocol that is intended to provide a method for submitting data in one location to another. The project has used the SWORD protocol to submit data into the DSpace, EPrints and Fedora-based repositories participating in the project.
2. *Version Control check-in*: As noted, the CVMA and H3FRP projects maintained by the Centre for Computing in the Humanities (CCH) use the Subversion versioning system. The project used the SVNKit (<http://svnkit.com/index.html>) – an open source, Java software library that provides API access to manipulate Subversion repositories over http(s), svn(+ssh) and file:// protocols – to automate the process of data check-in.
3. *File Transfer Protocol (FTP)*: The File Transfer Protocol was used as a data transfer mechanism of last resort, if alternative submission mechanisms do not exist.

## Outputs and Results

SHERPA DP2 built upon the groundwork laid in the first project to develop a set of modular services that are capable of capturing diverse types of data from varied technical systems, performing curatorial and preservation activities and re-submitting the data to a third-party.

### Application of disaggregated Service Model to a wider variety of content providers

The first SHERPA DP project produced a high-level OAIS-compliant model for disaggregated services around which the discussion of a service provider infrastructure might be discussed. The SHERPA DP2 project extended the model to consider Content Providers that exhibit portal-like behaviour, storing and/or referencing data in disparate locations using several different technologies.

### Extended monitoring and capture services

The technical analysis of each project partner highlighted several methods of monitoring and capture at different levels of granularity (repository-wide, academic unit, specific academics). We were able to utilise a range of technologies, including OAI-PMH, web feeds, web crawl, database backup and

version control check in/out to the capture of data stored by Content Providers on internal networks and public servers.

### **Requirements analysis**

The approach was underpinned by a better understanding of the expectations that staff located at Content Providers expect and require from a Preservation Service Provider. The project team consulted the institutions that were participating in the project, reanalysed requirements gathered in the first SHERPA DP project (Bodhmagé, 2006; Knight, 2005) and performed secondary research of other sources to develop a set of requirements. The outcome of the requirements analysis was the identification of three user types (Preservation Administrator, Repository Administrator and Repository Viewer), each of whom would have different needs and require different privileges to perform and monitor various types of curatorial action. These may be separated into three high-level categories: [1] the ability to request the performance of a curatorial activity (e.g. format conversion, metadata generation); [2] the ability to perform a curatorial activity; and [3] the ability to obtain information about curatorial activities.

### **Continued development of preservation metadata standards**

The first SHERPA DP project produced a PREMIS-compliant metadata schema for the description of e-prints stored by a Preservation Service Provider. For SHERPA DP2, the project team worked with the SOAPI project to extend the vocabulary to cater for the disparate requirements of content types of technical systems operated by DP2 partners and identification further work that is necessary to support the PREMIS 2.0 standard (Hedges & Knight, 2008).

### **Outcomes**

The embedding of preservation functionality within the digital curation workflow has been a key area for development for several years. Projects such as PRESERV and SHERPA have examined the preservation requirements of research papers and electronic theses that, until recently represented a large percentage of the data stored by institutional repositories. However, as a representation of the research output produced by an academic institution, there are many other content types and delivery systems that have been previously ignored.

The SHERPA DP2 project complements the work performed by other projects in developing the research data management agenda. The project was built on the premise that the data management requirements of an institution extend beyond the confines of a digital repository. Instead, preservation services must be able to interoperate with diverse types of technical systems and curate a wide variety of content types. Indeed, the research output stored external to the managed repository environment may be at most risk of becoming inaccessible, as a result of data loss or corruption. By building upon earlier work in recognising the requirements of research data, it advances the research data management agenda by asking the following question: what services can be built that enable an institution to capture, curate and preserve research data, without disrupting the existing workflow of the researcher?

By working with several partner institutions, the project gained a greater understanding of the range of systems in use within the research community to store and deliver digital outputs. More importantly, it was also able to develop a processing workflow and integrate third-party tools that allow communication with several types of technical system (DSpace, EPrints, Fedora, Invenio, Subversion, MySQL and other types of static web site) for the purpose of the monitoring and capture of research data for subsequent curation and preservation in a trusted preservation environment. Case studies that describe the approach taken for each project partner may be found on the project web site (<http://www.sherpadp.org.uk/sherpadp2.html>)

At the outset of the first SHERPA DP project, an objective was to develop a third-party preservation service that would enable institutional repositories to become OAIS-compliant and comply with the requirements of the Trusted Digital repository specification by accepting responsibility for long-term preservation, even if they did not have the funding or staff allocation to perform appropriate activities in-house. The preservation service may be operated by an external organisation or a central department located within the institution itself. The work performed by the SHERPA DP2 project enables curation and preservation services to be provided to a range of technical architectures and systems. The implications for institutional systems may be extensive – by using a service provider to

undertake preservation activities and operate as the 'glue' to tie it with other services such as characterisation registries, it may be feasible for institutions to certify a range of systems against criteria for a trusted digital repository.

## Conclusions

The general conclusion of the SHERPA DP2 project mirror those that were drawn for the first project, there remains no out-of-the-box solution to preservation. Although extensive progress has been made in the development of standards, specifications and tools to support the process, the performance of activities necessary to initiate and monitor curatorial and preservation activity requires strategic decision making that, to date can only be made by staff able to consider the needs and expectations of long-term access.

The provision of a preservation service, offered by a third-party organisation or a central department within the institution represents an effective method to curate and preserve research data stored by an institution. The primary requirement is that appropriate expertise and services exist to support the machine-to-machine transfer, curation and preservation and submission activities (see requirements list in methodology) required for a Content Provider and a Service Provider to interact.

The SHERPA DP2 project provides a practical example of the type of services that a Preservation Service Provider may offer to a Content Provider and the scenarios in which they may be provided. More interesting, it highlights potential issues that a Service Provider must address when negotiating a contract with a Content Provider. It may be difficult to provide preservation services for many Content Providers, due to the type of data that they store and the scale of services required.

## Implications

The development of an operational Service Provider for Content Providers is a practical solution to preservation. The research community may support a number of preservation services that operate in and are tailored to the requirements of different research communities and different organizational models. For example, a preservation service may be established to cater for the needs of different repositories in an institution, or different repositories that share a common research theme. However, the project, by definition, has considered preservation services in the context of Content Providers that implement a limited number of technical systems and require certain types of functionality. Further questions that may be addressed are: What are the implications for Content Providers that require different types of service and interact in different ways using different technologies?

## Recommendations

- It is recommended that repository developers, particularly those responsible for DSpace, EPrints & Fedora collaborate to simplify the process of exchanging complex metadata formats between different platforms. Although efforts such as the Scholarly Works Application Profile (SWAP) have made efforts to document esoteric elements in an application profile, it remains difficult to import metadata formats (e.g. the full EPrints application profile, Fedora-like METS and preservation metadata formats) exported from Repository A into Repository B or Repository C. It is suggested that an appropriate body, such as the Repository Support Project (RSP) or Digital Curation Centre (DCC) is allocated the task of organising meetings or other event that focus upon encouraging collaboration between repository developers that result in the creation of new import scripts for each repository system to handle metadata exported by other repository systems.
- The SWORD plug-in was a key component of the Submission workflow for repository-to-repository transfer. However, it is evident that further development is necessary. As an extension of the above, it is recommended that further work is performed to enable each repository to accept and process a wider range of metadata application profiles.
- Recommend that further investigation of the activities required to capture data from web sites and other technical systems and import it into a digital repository environment is performed. In particular, work should be funded to look into the requirements for curation of WARC and ARC objects in a digital repository.

Project Acronym: SHERPA DP2  
Version:  
Contact: Gareth Knight  
Date:

## References

Arms, C.R. & Fleischhauer, C. (2005). Digital Formats: Factors for Sustainability, Functionality, and Quality. IS&T Archiving 2005 Conference, Washington, D.C.  
<http://www.digitalpreservation.gov/formats/intro/papers.shtml>

Beagrie, N. Chruszcz, J. & Lavoie, B (2008). Keeping Research Data Safe.  
<http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx>

Bodhmag, K. & Hedges, M. (2006). Technical Specification - SHERPA DP  
<http://www.sherpadp.org.uk/sherpadp.html>

Jones, S. Ross, S. & Ruusalepp, R. (2009). Data Audit Framework Methodology. [http://www.data-audit.eu/DAF\\_Methodology.pdf](http://www.data-audit.eu/DAF_Methodology.pdf)

Knight, G. & Anderson, S. (2007). Sherpa DP Final Report. <http://www.sherpadp.org.uk/sherpadp.html>

Hedges, M & Knight, G. (2008). Preservation Metadata Framework. Preservation Metadata Framework. <http://www.sherpadp.org.uk/sherpadp2.html>

Hedges, M. (2009) SOAPI final report. Awaiting publication

Knight, G. (2008) Report of file types supported by IRs participating in the SHERPA DP2 project.  
<http://www.sherpadp.org.uk/sherpadp2.html>

Knight, G. (2009a). SHERPA DP2 – Edinburgh Research Archive Case Study.  
<http://www.sherpadp.org.uk/sherpadp2.html>

Knight, G. (2009b). SHERPA DP2 – Goldsmiths College Case Study.  
<http://www.sherpadp.org.uk/sherpadp2.html>

Knight, G. (2009c). SHERPA DP2 – University of Hull Case Study.  
<http://www.sherpadp.org.uk/sherpadp2.html>

Knight, G. (2009d). SHERPA DP2 - CVMA Case Study.  
<http://www.sherpadp.org.uk/sherpadp2.html>

Knight, G. (2009e). SHERPA DP2 – Henry III Fine Rolls Case Study  
<http://www.sherpadp.org.uk/sherpadp2.html>

Kulovits, H. et al (2008). Planets Planning Tool (version 2)  
[http://www.planets-project.eu/docs/reports/Planets\\_PP4-D4\\_PlanetsPlanningTool.pdf](http://www.planets-project.eu/docs/reports/Planets_PP4-D4_PlanetsPlanningTool.pdf)

Martinez-Urbe, L. (2008). Research data management services: Findings of the consultation with service providers scoping digital repository services for research data management.  
[www.ict.ox.ac.uk/odit/projects/digitalrepository/](http://www.ict.ox.ac.uk/odit/projects/digitalrepository/)

Research Libraries Group (2002). Trusted Digital Repositories: Attributes and Responsibilities.  
<http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf>

Knight, G. (2009f) Workflow Analysis of management activity within the Preservation Service Provider.  
<http://www.sherpadp.org.uk/sherpadp2.html>

## Links

- Digital Preservation and Records Management Programme  
<http://www.jisc.ac.uk/preservation/>
- DISC-UK DataShare project <http://www.disc-uk.org/datashare.html>
- PoWR project <http://jiscpowr.jiscinvolve.org/>

Project Acronym: SHERPA DP2

Version:

Contact: Gareth Knight

Date:

- SOAPI (Service Oriented Architecture for Preservation and Ingest of Digital Objects). <http://www.kcl.ac.uk/iss/cerch/projects/portfolio/soapi.html>
- PRESERV Project. <http://preserv.eprints.org/>
- SHERPA DP. <http://www.sherpadp.org.uk/sherpadp.html>
- SOURCE Project. <http://www.jisc.ac.uk/whatwedo/programmes/reppres/tools/source>