



Final best practice guidelines and recommendations

For Large-scale long-term repository migration;
For Preservation of research data;
For Bit preservation

Authors


Large-scale long-term repository migration: Miguel Ferreira, José Carlos Ramalho, Rui Castro, Luís Faria, Vitor Fernandes, Luís Miguel Ferros, Hélder Silva, (KEEP Solutions), Ivan Vujic (Microsoft Research), Opher Kutner (Ex Libris), Lynne Chivers (British Library), Kåre Fiedler Christiansen (State & University Library of Denmark), Stanislav Barton (Internet Memory)

Preservation of research data: Catherine Jones, Simon Lambert (Science and Technology Facilities Council), Barbara Sierman (National Library of the Netherlands), Kirnn Kaur (British Library)

Bit preservation: Lieke Ploeger, Barbara Sierman (National Library of the Netherlands), Catherine Jones (Science and Technology Facilities Council), Bram van der Werf (Open Planets Foundation), Ivan Vujic (Microsoft Research), Opher Kutner (Ex Libris), Kirnn Kaur (British Library)

February 2014

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

This work is licensed under a CC-BY-SA International License 



Introduction to the best practice guidelines and recommendations

The SCAPE project¹ aims to enhance the state of the art in digital preservation with a particular emphasis on the scalability of its solutions: that is, their capacity to handle digital objects that may be very numerous, individually very large, heterogeneous or complex. The motivating force of the SCAPE project is scalability, interpreted in several dimensions: number of objects, size of objects, complexity of objects, and heterogeneity of collections.

The best practice guidelines and recommendations cover three areas of digital preservation. These are: *Large-scale long-term repository migration*, *Preservation of research data* and *Bit preservation*.

Large-scale long-term repository migration

This document provides guidelines to support the migration from legacy repository systems by describing the stages, activities and associated risks that comprise this type of endeavour. The presented guidelines are based on a combination of 13 existing methodologies that have been surveyed and unified into a comprehensive multistep methodology.

This document includes implementation guidelines, examples of practice and expected outputs of each activity in the migration process. Also, a quick implementation checklist is included to aid IT professionals in double checking that all the angles have been covered during the preparation, execution and post-operational stages of a repository migration project.

Preservation of research data

It is clear that much research data has some of the SCAPE characteristics of scale. Even in domains where the sheer data volumes are not so large, the data is likely to have complex semantics and to have undergone processing which might need to be recorded in order that future users may understand the provenance of the data. In this document guidelines and recommendations for the preservation of research data are based on a broad literature review. These have been further enhanced by the experiences of the SCAPE partners and lessons learnt within the project.

Bit preservation

Charting the border between digital and bit preservation can somewhat be evasive. In this document, the following working definition is used: Bit preservation is concerned with the *persistence* of the file over time, while digital preservation is concerned with the *accessibility* of the file over time in terms of format and/or application obsolescence. That is to say, bit preservation is a precondition of digital preservation.

¹ www.scape-project.eu



The aim of the SCAPE project was to investigate the issues that large scale collections bring to the subject of digital preservation. It can be straight forward to manage tools and activities to support small amounts of material, but are these scalable to real life collections? Some of the SCAPE partners are content holding institutions and their expertise has been used to write this report. Real-life experiences are shared through the use of two case studies.



Large-scale long-term
repository migration

Best practice guidelines and
recommendations

Table of Contents

1	Introduction	4
1.1	Scope	5
1.2	Audience	7
1.3	Approach	7
2	Repository migration best practice guidelines	9
2.1	Analysis and consultation	9
2.1.1	Characterisation of legacy environment	10
2.1.2	Characterisation of target environment	12
2.1.3	Data analysis	14
2.1.4	Strategic planning	16
2.1.5	Definition of requirements	17
2.2	Planning & design	19
2.2.1	Project planning	19
2.2.2	Design of migration routines	21
2.2.3	Design of test plan	23
2.3	Development	25
2.3.1	Development of migration routines	25
2.3.2	Development of testing routines	27
2.4	Setup & testing	28
2.4.1	Target environment provisioning	28
2.4.2	Rehearsal & testing	29
2.5	Execution	30
2.5.1	Execution of migration routines	31
2.6	Validation	32
2.6.1	Execution of testing routines	33
2.6.2	Reporting	34
2.6.3	Cut-over	35
2.7	Wrap up	37
2.7.1	Training	37
2.7.2	Documenting	38
2.7.3	Supporting	39
3	Conclusions	41
4	Implementation checklist	42
5	References	45

1 Introduction

Several institutions around the world are currently running long-term digital repositories that have now been in operation for many years. Some of these systems are approaching the end of their life spans and will soon be replaced by the next-generation of long-term large-scale repository systems. This will unavoidably imply the migration of metadata records, millions of files, and terabytes of data from the legacy repository system to the newly adopted one. Because of the large scale of this operation, this procedure should entail careful planning, validation and support.

There are many reasons why organisations might decide to migrate to new a repository system. For example:

- Repository system does not cope well with current business needs (e.g. lacks desired characteristics like functionality, performance, capacity, interoperability, usability or others);
- Budget cuts mandate that a new, more financially sustainable repository is adopted;
- The repository vendor or supporter ceased to exist (i.e. the repository is no longer supported);
- Repository vendor or supporter does not provide a satisfactory level of support services;
- The technological environment needed by the repository system is no longer supported (e.g. security updates are no longer available for the supporting operating system).

Several scenarios can be considered examples of legacy repository migration projects, for example [1]:

- Migration from a relatively simple system into another system;
- Upgrading a system to a new version of the same system;
- Converge multiple systems into a single composite system;
- Critical system migration that requires the migration to be rolled out over a period of time without interruption of operations;
- Multiple concurrent systems migrations and consolidation efforts.

Systems (or repository) migration is often referred to as "IT Transformation". However, this term usually applies more to the overall changing of an organisation's technology systems and usually implies a significant business change consequent to the technology change.

It might appear that any two systems that maintain the same sort of information must be doing very similar things beneath the surface and, therefore, should map from one to another with ease. However, this is hardly ever the case. Legacy systems have historically proven to be far too lenient with respect to enforcing integrity at the data level [9]. Fields that typically should be populated from a controlled list of values tend not to be validated by the system, and therefore the database ends up with unexpected values that require exceptional handling during migration. Another common problem has to do with the theoretical design differences between hierarchical and relational systems. Two of the cornerstones of hierarchical systems, namely de-normalization and redundant storage are strategies that make the relational purist cringe. [9]

Additional difficulties may be encountered while migrating information from one system to the other, for example:

- Extracting information from the legacy system can be extremely complex, especially in the case where the functionality to export information does not exist, technical support is unable

to provide the necessary help, documentation is scarce or incomplete, the organisation does not have the necessary system level credentials to gain actual access to the data;

- Mapping between the previous semantic structures to the ones of the new repository might be difficult to attain, or impossible when these structures are highly incompatible (i.e. data loss will take place);
- The process of transforming and/or cleansing data during the migration process is prone to errors caused by incorrect settings or bad programming;
- The necessary validation procedures can be extremely difficult to design or automate.

Although migrating data can be a fairly time-consuming and risky process, the benefits can be worth the cost, as legacy systems do not need to be maintained any longer. Although migrating from legacy systems is a major research and business issue, there are few comprehensive approaches to migration. Given the bewildering array of legacy information systems in operation and the problems they pose, it seems unlikely that a single generic migration method would be suitable for all systems. However, a set of comprehensive guidelines to drive migration is essential [2].

For the reasons mentioned above, there exists an obvious need for a sound, methodological approach by which organisations can steer themselves to accomplish a successful repository migration. This document provides guidelines to support the migration from legacy repository systems by describing the stages, activities and associated risks that comprise such an endeavour.

These guidelines are structured as follows: section 1 constitutes an introduction to the guidelines, describing the motivation behind the development of this document, the expected audience, and the approach that was followed. Section 2 details the repository migration best practice guidelines. This section is organized in multiple subsections, each of these depicting a stage or an activity that comprise these guidelines. Finally, a conclusions and future work section is included in section 3.

1.1 Scope

In the context of Information Technologies, the term migration may mean a lot of different things. Considering the context of digital preservation alone, concepts such as file migration, media migration, format migration, repository migration, data or metadata migration/conversion are commonly found in the specialized literature. However, all of these terms mean very different things and may comprise very distinct approaches in the way they are conducted. In the context of these guidelines, we will make use of the term “migration” as in “systems migration”, i.e. the set of activities necessary to replace an existing Information Technology system (or platform) by a new one. This necessarily entails the establishment of a new IT environment and the moving of all relevant information managed by the original system to the new one.

In the context of digital preservation, the term “repository” is often used with two different meanings:

- 1) The “repository” as the set of policies, standards, and technology infrastructure that provides the framework for doing digital preservation [12], and
- 2) The “repository” as an Information Management system, i.e., a system of software and hardware that can be relied upon to manage digital information that follows certain rules [12].

In the context of these guidelines, we will make use of the second definition of repository. A repository is, therefore, a system composed of software and hardware that is set up to follow certain rules or policies and that is responsible for safekeeping and managing digital information.

A repository usually entails several types of digital information, for instance:

- **Data** – usually the most important asset managed by the repository, i.e., the actual information that users are interested in and expect to be kept safe and accessible (e.g. images, audio, video, documents, datasets, 3D objects, etc.);
- **Metadata** – information about data managed by the repository. Metadata fulfils many goals, e.g. supports data discovery, ensures authenticity and provenance, provides characterisation and technical information about the data, etc.;
- **System specific information** – information that is highly dependent of the information system, often automatically generated and intrinsically necessary for the system to function (e.g. configurations, logs, indexes, user information, branding and styling, etc.).

“Repository migration” is the process of transferring (and/or transforming) digital information between two or more information systems, whether this be data, metadata or any other kind of information considered to be relevant to the continuity of the organisation or individual that relies on that information. This activity can broadly be called “IT transformation”, however, in the context of these guidelines we will refer to this process as “repository migration”. This allows us to build guidelines that are more focused on the actual preservation problem that is to make sure data survives the replacement of its host system. Moreover, from a preservation planning perspective, the replacement of the repository can be seen as a preservation action, and therefore the term “repository migration” is well applied in this context.

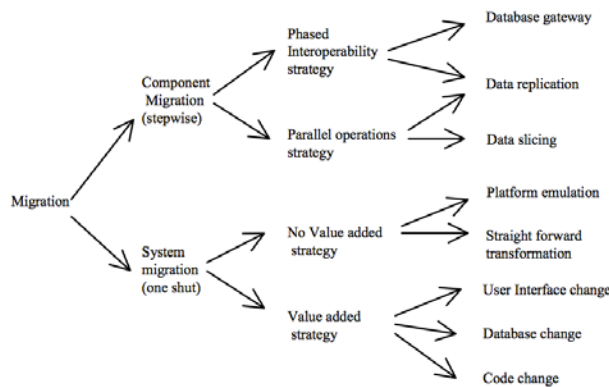


Figure 1 - Classification of repository migration approaches.

In Figure 1, two classes for repository migration approaches can be seen. The first class is “component migration” in which the legacy systems are broken down into independent components and each component is migrated separately. There will be a period of transition where both legacy and the new platform have to be online and to work together. Two strategies will arise, “phased interoperability” and “parallel operations”. Both of these need the data to be shared via “database gateways”, replicated on the two platforms, or sliced into separate independent domains to be migrated gradually to a new platform [20].

The second class of migration approaches is the “system migration” approach in which the whole legacy system and the data are transferred to the new platform in a single step. There are two subclasses to this approach: “no value added” (in which the system remains practically the same,

either by emulation or simple upgrade) and “value added” (where the system acquires a considerable number of new capabilities and/or functionality) [20].

This last approach, value added migration, may lead to changes in the user interface, the database and the program logic. Although, migration may be more complex in this situation, its long-term benefits will be much greater. It may offer more flexibility, better system understanding, easier maintenance, and reduced costs [20].

1.2 Audience

These guidelines are meant for those who work in or are responsible for digital repositories seeking guidance on how to design a reliable process of information migration from a legacy system(s) to a new repository system. Some institutions may also choose to use these guidelines when they are not themselves involved in the design of the migration process, but are outsourcing it to an external supplier. The document will help these institutions to better understand all the underlying steps involved in a repository migration process and enable them to better quantify the resources needed, the checkpoints that should be created and the monitoring and validation procedures to be installed.

The guidelines are also expected to be of interest to a wider community of IT specialists, programmers, project managers, researchers and practitioners in general. However, organisations planning to upgrade or adopt new digital repository systems are the ones most likely to be interested in reading these guidelines.

1.3 Approach

In order to develop these guidelines, we adopted the following five-step procedure:

Step 1 - Survey existing best practices documentation

This step consisted of collecting existing best practice materials from recent years. The survey revealed a significantly mature research field in which repository migration is already well framed, i.e. the Legacy Information System Evolution [2, 3, 9, 14–16]. Additionally, several white papers, communications and technical reports from the IT industry were also surveyed and taken into consideration as they provide valuable hands-on information necessary to support these guidelines [1, 4, 8, 11, 16–19, 21].

Step 2 - Identification of repository migration methodologies

Several of the research and technical documents collected in the previous step depicted methodologies for how to perform legacy information systems migration. Some of these methodologies were very simple and generic [4, 11, 19], while others were very detailed and domain specific [2, 14, 15]. This step consisted of the analysis and evaluation of all the previously collected documentation and the selection of those that included well-grounded information on how to perform repository migrations.

Step 3 - Comparison of repository migration methodologies

Some of the methodologies identified in the previous step were very simplistic, being composed of merely four generic steps that could very well be applied to any software development process, while others were extremely detailed in the tasks they depicted (down to the product name and version number). In order to compare all of these approaches systematically we have created a

comparison matrix in which all the common steps described in each of the repository migration methodologies are mapped for easy comparison (Table 1). The numbers in the top row of Table 1 are references to the system migration methodologies (see Chapter 5, References).

Table 1 - Comparison matrix of legacy information system migration methodologies.

Activity	Sub activity	System migration methodologies												
		[8]	[9]	[19]	[1]	[4]	[16]	[17]	[11]	[21]	[18]	[15]	[14]	[3]
Analysis & consultation	Characterisation of legacy environment		•			•			•	•	•	•		•
	Characterisation of target environment										•	•	•	
	Data analysis		•			•		•		•	•	•	•	•
	Strategic planning		•		•			•		•	•			
	Definition of requirements	•		•	•	•			•				•	
Planning & design	Project planning	•		•	•		•	•	•	•				
	Design of migration routines	•	•		•	•	•	•			•	•	•	•
	Design test plan								•		•			
Development	Development of migration routines	•		•	•				•	•		•	•	
	Development of testing routines													
Setup and testing	Target environment provisioning				•			•	•			•	•	•
	Rehearsal & testing		•				•	•	•	•		•	•	
Execute	Execution of migration routines	•		•	•		•	•	•	•				
Validate	Execution of testing routines	•	•		•	•	•	•	•		•	•	•	•
	Reporting								•		•			
	Cut-over	•		•	•				•	•		•	•	
Wrap up	Training													
	Documenting				•			•	•			•	•	•
	Supporting		•				•	•	•	•		•	•	

Step 4 - Creation of a unified repository migration methodology

The next step in the creation of these guidelines was the classification and generalisation of all the activities found in the surveyed methodologies. This process allowed us to combine all the approaches into a single unified methodology that comprised the steps included in all the other surveyed methodologies. We also made small adjustments to the terminology to make it more compatible with the terminology used in the digital preservation domain.

2 Repository migration best practice guidelines

These guidelines are, above all, the result of a formalisation exercise that aims to identify and describe the most important steps in a repository migration process. They do not intend to be prescriptive or even complete but instead they aspire to provide enough information to any vendor, customer or IT specialist to crosscheck that the most important steps have been addressed during all stages of a repository migration project.

Figure 2 depicts the unified methodology for a successful legacy repository migration. The methodology comprises 7 stages, each covering a series of activities.

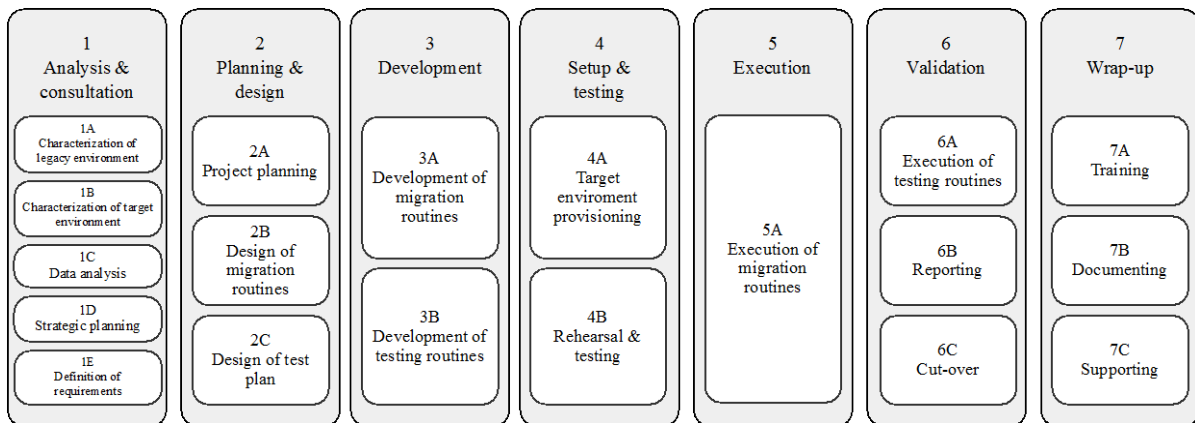


Figure 2 – Unified methodology for legacy repository migration.

It is important to note that the activities included in this methodology are presented in a sequential manner that reflects the natural flow of events that take place in a repository migration process, however, often is the case that some of these activities are executed in a different order or even revisited during the overall migration process.

The following sections describe each of the stages and activities included in the unified methodology. For each activity in the methodology we provide a table depicting the ID of the activity, its name, a brief description, implementation guidelines, expected outputs, examples of practice and other relevant information.

2.1 Analysis and consultation

ID	1
Activity	Analysis & consultation
Description	The first step in a repository migration process is to gain insight into the needs of all interested parties in order to define the most appropriate migration strategy and formalise all the requirements. This includes a deep analysis of both legacy and target systems, the characteristics of data to be migrated, and the business needs that the project/migration expects to meet. The quality of the analysis stage will influence the degree of success of the overall migration project.

	<p>This step can be interpreted as a preservation planning activity. The business goals and requirements of the repository should be defined and migration alternatives should be tested (or, at the very least, brainstormed) in order to determine the best migration approach and destination platform. Tools such as Plato - the Preservation Planning Tool - may play an important role in this activity as these can guide the project manager through the series of well-defined steps in the decision making process.</p>
<p>Expected outputs</p>	<ul style="list-style-type: none"> • A set of documents that provide insight on the business processes and rules already established in the organisation that have influence on the repository; • A set of documents that provide an overall technical description of the legacy repository system(s). The documentation should be as comprehensive as possible (more is better than less). • A set of documents that thoroughly describe the target repository environment. The documentation should be technical and as vast as possible (more is better than less). • High-level definitions and agreements about the data entities that exist in the legacy system, which should be migrated, and what entities or periods of time can be discarded. A deeper analysis of the data follows later in the process; • Signed non-disclosure agreements with individual team members (if necessary) and definition of a screening process for each team member (if necessary); • Minutes recording the outcomes of all conversations held during the strategic planning stage; • Definition of migration requirements; • Definition of the migration strategy to follow. <p>All of these outcomes may result in a well documented preservation plan, including the business goals and preservation requirements, the alternatives that have been evaluated and the results of the decision process.</p>
<p>Other relevant information</p>	<p>Plato - the Preservation Planning tool is accessible at http://plato.ifs.tuwien.ac.at:8080/plato.</p>

2.1.1 Characterisation of legacy environment

<p>ID</p>	<p>1A</p>
<p>Activity</p>	<p>Characterisation of legacy environment</p>
<p>Description</p>	<p>The first step in preparing a repository migration should be the assessment of the legacy system(s) technology environment. Repository migration requires a complete understanding of all the involved technologies, including hardware, networks, software, interfaces, programming languages, data structures, services, servers and time</p>

	<p>requirements and acceptable service levels (e.g. availability of the system).</p> <p>To be able to size and plan the migration process, as well as setting accurate budgets and timelines, one must understand the complexity, relationships, quality, and volume of the legacy system and its data. This will enable the definition of appropriate requirements such as replication needs, project schedule, system response times, vendors that need to be contacted, and the hardware configuration necessary to host legacy and upcoming data.</p> <p>The management costs of the legacy technological environment should also be determined. This offers the best opportunity to define the benefits of migration and to help narrowing down migration strategy options, e.g. keeping the legacy system running in parallel, might be too expensive.</p>
<p>Implementation guidelines</p>	<p>Collect all relevant documentation on the legacy repository environment. This should include hardware architecture, software architecture, white papers, technical reports, database schemes, relevant configuration files, firewall rules, network diagrams, metadata schemas, repository policies, user manuals/guides, etc. Also, do not forget to look at interfaces with other systems, for example resource discovery and access.</p> <p>Sometimes documentation does not exist and information will have to be obtained by other means. Talk to IT personnel and backoffice-users. They are a valuable source of information, e.g. in long lasting repository systems it is common to find distinct metadata eras. This means that the rules or policies used to create/manage data and metadata have evolved over time. Most of the time these policies do not exist or are not well documented.</p> <p>This collection of information will help the project team to identify and locate all relevant data and metadata to be migrated to the new environment.</p> <p>Keep all information well organised. This will be used later for planning and documenting the overall migration process.</p>
<p>Expected outputs</p>	<p>A set of well organised documents that provides a comprehensive technical description of the legacy repository system. The documentation should be as vast as possible (more is better than less). Examples of such documents are:</p> <ul style="list-style-type: none"> • Information about the location of data and metadata, and detailed instructions on how to extract them; • Network diagrams, firewall rules, relevant network addresses, etc.; • Credentials to all components and data stores of the legacy system;

	<ul style="list-style-type: none"> • Data dictionaries, metadata schemas and folder structures; • Documentation about rules or policies for data management and data/metadata creation that have been in practice over time (talk to users if necessary); • Hardware/software architecture; • Business/data architecture; • White papers or technical reports about the legacy system; • Relevant configuration files; • Relevant data standards.
Examples of practice	<p><i>“During the migration of around 3 million descriptive metadata records of an Archival legacy system, we found that there were clearly distinct eras in terms of metadata creation policies. This resulted in more complex mapping rules and validation procedure.”</i>, KEEP SOLUTIONS</p> <p><i>“While analysing data containers in a legacy Web archive, we found that file organization naming schemes changed over time, not retaining backward compatibility. Moreover minor changes in the data writer routines did not yield the most recent reader routines making a small fraction of the stored data currently unreadable/inaccessible by the query engine.”</i>, Internet Memory Foundation</p>

2.1.2 Characterisation of target environment

ID	1B
Activity	Characterisation of target environment
Description	<p>When planning a migration project it is also important to understand the capabilities and architecture of the target technological environment. Knowing what users want from the new repository (or disliked about the old one) and understanding its architecture will guide the development of the data migration routines, including mappings, data selection, time behaviour, etc.</p> <p>Collecting documentation on the target system is naturally simpler than collecting documentation on the legacy system. Nevertheless it is very important to collect and organise this information as it will be extremely important during this and subsequent repository migration projects.</p> <p>This step assumes that the target environment/system has already been chosen and is ready to be implemented.</p>
Implementation guidelines	<p>Collect all relevant documentation of the target system(s). This should include hardware architecture, software architecture, requirements for both software and hardware, white papers, technical reports, database schemes, configuration files, firewall rules and network diagrams, metadata schemas, repository policies, etc.</p>

	<p>Keep all information well organised. This will be used later for documenting the overall migration project and to define the mapping rules for the migration procedure.</p>
<p>Expected outputs</p>	<p>A set of documents that thoroughly describe the target repository environment. The documentation should be technical and as vast as possible (more is better than less). Examples of such documents are:</p> <ul style="list-style-type: none"> • Information about the location of data and metadata, and detailed instructions on how to create it; • Network diagrams, firewall rules, relevant network addresses, etc.; • Credentials to all components and data stores; • Data dictionaries, metadata schemas and folder structures; • Hardware/software architecture; • Business/data architecture; • White papers or technical reports about the new system.
<p>Examples of practice</p>	<p><i>“The new repository system may require some organisational processes to change in order to cope with the new offered functionality and mode of operation, e.g. perhaps repository ingest and access processes may need to be changed. It is important to gain user support, especially if major changes in the established processes are expected to happen. Include in the planning steps sufficient user engagement and communication. Start this early on.”, British Library</i></p> <p><i>“To better determine the characteristics of the target environment it is sometimes necessary to do a pilot implementation of the new repository in order to assess the performance of the data storage as well as the performance of the migration process.”, Internet Memory Foundation</i></p> <p><i>“Attention should be given to system performance – specifically loading duration – in order to provide estimates for the expected duration of the full migration. This is especially important to determine when changes in interfacing systems will need to occur.”, Ex Libris</i></p>

2.1.3 Data analysis

ID	1C
Activity	Data analysis
Description	<p>A legacy repository often comprises a wide range of distinct information, including structured and unstructured data. In order to migrate the myriad of information into the new repository, it must first be located, examined, defined and delimited.</p> <p>The aim of the data analysis step is to identify the information sources and information entities that have to be transported into the new system. Information sources include all types of data stored, managed or generated by the legacy system (e.g. digital objects, metadata, logs, user information and configurations). Identifying the ways that data was used is highly important.</p> <p>One may assume that not all data is relevant to be preserved, meaning that some of it can be discarded during the migration process. In order to get a better sense, it is helpful to look at the applications, databases and talk to users to understand exactly what information items are relevant to be migrated. You may find that the overall cost of migration is prohibitive relative to the volume of data that needs to be moved and that a compromise on which data is to be migrated must be done. Examples of legacy data that might not be considered necessary to include in the migration process are old collections that are no longer relevant from a legal and/or business perspective, system dependent data such as internal indexes or system settings, etc.</p> <p>Data classification, i.e. the conditions for data access, retention requirements and security measures such as encryption, should also be addressed in this step. Often repositories hold classified information whose access is highly restricted. One may have to identify the needs of the IT environment and ways in which data may be segregated and protected from members of the migration team. Members of the project team may have to be screened and required to sign non-disclosure agreements. Even a limited set of classifications will have tremendous impact on the way the migration project may be conducted.</p>
Implementation guidelines	<p>This activity is still very much high-level. The outputs of this activity do not consist of mapping rules or source code. These will be devised later on in the project in step 2B.</p> <p>In this step the project manager must collect information about all the data entities that exist in the legacy system, which of these are expected to be migrated and which can be discarded. The project manager should also record the level of privacy and protection associated to each data entity and add notes about any apparent data transformations that will have to be created later on.</p> <p>This step is also an ideal opportunity to assess possible data quality</p>

	improvements that could be included in the migration process.
Expected outputs	<p>The expected outputs are mostly high-level definitions and agreements. For example:</p> <ul style="list-style-type: none"> • A report on the data entities existing in the legacy system, which should be migrated, and what entities or periods of time can be discarded. For migrations from multiple legacy systems, it's important to know if duplicated information exists. For the selected entities, the report should include what are the access constraints and data quality improvement suggestions; • Signed non-disclosure agreements with individual team members (if necessary); • Documentation about the screening process that each team member should be submitted to (if necessary).
Examples of practice	<p><i>“The consumers (or representatives) may play an important role in this step. Make sure they are consulted. In the case of multiple legacy repositories, bear in mind that there may be duplicated data across the systems. This will need to be managed carefully in the migration. Check whether any Authority files need to be migrated.”</i>, British Library</p> <p><i>“Data analysis may comprise a confirmation of the ability to access/read all the archived data. Due to evolutions in third party software that produces container files for a Web archive, we find out that with the same third party reader libraries it was no longer possible to read all the archived files. To address this this problem, it was necessary to tweak the Web archive file reader. The output was a report addressing each issue that were encountered.”</i>, Internet Memory Foundation</p> <p><i>“In long lasting repository systems it is common to find distinct metadata eras. This means that the rules or policies used to create/manage data and metadata have evolved over time. Sometimes this even changes depending on the operator of the system. In systems with very little data constraints this means that different migration rules or mappings will be used to migrate distinct eras (e.g. it is common to find systems where dates have been entered in very distinct ways, e.g. 1986-02-01, 86-01-02, 2nd Jan. 1986).”</i>, KEEP SOLUTIONS</p> <p><i>“As part of preparation for the migration we needed to get details of the IDA content. The repository contained about 40TB of data at initial stage of migration planning in late 2011, which become 48TB during 2012 as data was ingested into the IDA until December 2012 at the point the new digital repository was switched on.”</i>, Archives New Zealand [10].</p>

Other relevant information	ISO 27001 is a good source of information on how confidential data should be handled.
----------------------------	---

2.1.4 Strategic planning

ID	1D
Activity	Strategic planning
Description	<p>The objective of the strategic planning step is to identify the business and operational requirements that impact the migration process. Various stakeholders within the institution should be consulted to ensure that their requirements are factored into the migration plan.</p> <p>This step takes into consideration the information collected in the previous steps and defines the migration strategy to be adopted in the following steps. Migration strategies depend on the size, complexity and business requirements of the repository system. For example:</p> <ul style="list-style-type: none"> • <u>Light migration scenario</u>: it typically involves loading data from a single source into a single target. Few changes are required in terms of data quality improvement; mapping is relatively simple as is the application functionality to be enabled. This will likely be a once-off, "big-bang" procedure, i.e. "one shut system migration" [11, 20]. • <u>Medium migration scenario</u>: may involve loading data from a single source into a single target or to multiple systems. Data migration may be performed through multiple iterations, transformation issues may be significant and integration into a common data model is typically complex, i.e. "phased interoperability strategy" [11, 20]. It is common in upgrade projects, that migrations between multiple versions of the same system are done in order to bring data from the legacy system to the most recent version of the product; • <u>Heavy migration scenario</u>: typically involves providing a solution for application co-existence that allows multiple systems to be run in parallel. The integration framework is formulated so the current repository and future repository can work together, i.e. "parallel operations strategy" [11, 20]. This is usually the case when the amount of data to be migrated is so big that the whole migration process would impose an unbearable downtime of the service or data unavailability. Large-scale migration processes can easily take several months just for copying data between repositories.
Implementation guidelines	<p>By now the project manager should have a good idea about the complexity of the repository migration. Conversations should be held with all the stakeholders of the repository in order to define the migration strategy to adopt. The conversation should revolve around topics such as:</p> <ul style="list-style-type: none"> • Operational constraints – can the legacy system be stopped for a given period of time in order to complete the migration process?

	<p>Consider the service level agreements for the legacy systems, especially if the repository is providing a service to users external to the organisation;</p> <ul style="list-style-type: none"> • Time constraints – how much time can the legacy system be unavailable for consultation and/or data insertion? • Immovable deadlines, e.g. contract expiry; • Window of opportunity – when will be the best time to do the migration (e.g. over the weekend, during the holidays period, at night)? • Hardware constraints – do we have the appropriate hardware resources to implement a given migration strategy (servers, storage space, storage speed, etc.)? • Network constraints – is the network adequate to implement the migration strategy or is it necessary to create a separate network just for the migration? • Legal constraints – Are there any legal impediments that may influence the way the migration process can be conducted?
<p>Expected outputs</p>	<p>The main output of this activity is a definition of the migration strategy to follow.</p> <p>The outcomes of all conversations held during this activity and all resulting decisions should be recorded into minutes for later accountability and documentation. These will also be used as requirements in the project plan.</p> <p>Don't forget to consider the costs of the different strategy options, including staff costs. These may play an important role in the decision process.</p>
<p>Examples of practice</p>	<p><i>“Using a pilot implementation of the target repository system, we have come up with time/resource estimations necessary to migrate. This helped us to anticipate the impact on the quality of service of the current solution and take actions to minimize it.”, Internet Memory Foundation</i></p>

2.1.5 Definition of requirements

<p>ID</p>	<p>1E</p>
<p>Activity</p>	<p>Definition of requirements</p>
<p>Description</p>	<p>All the previous steps enable the project manager to estimate the resources that are needed to perform a successful repository migration. In this step, the project manager will define the success criteria for the overall migration project. These may include service-level agreements, expectations for the new storage infrastructure, and objectives such as reduced management costs, reduced storage expenditures, greater insight into expenditure, a simplified vendor model or greater technical flexibility or stability [4].</p>

	<p>In this step, the high-level requirements for migration, including the data to be migrated, performance requirements should be considered to devise an appropriate contingency plan in case anything does not run according to plan.</p>
<p>Implementation guidelines</p>	<p>Based on all the information collected so far, the project manager is expected to devise a series of success criteria that will be used to measure the operational success of the migration process as well as the overall success of the repository migration.</p> <p>A contingency plan should also be created in order to rollback all operations during migration in case the operational criteria of success are not being met.</p> <p>The <i>cut-over</i> strategy should also be defined in this step. This means defining how the legacy system will be abandoned and how all operations should be moved to the new system. There are essentially three ways of accomplishing this: 1) the “big-bang” strategy; 2) phased interoperability strategy; and 3) the parallel operations strategy. More details on these strategies are available on section 2.6.3.</p> <p>Preservation metadata requirements should also be defined. The legacy and target digital preservation system are expected to entail some sort of preservation metadata (e.g. PREMIS). The requirements of how this metadata should be transported and inserted into the new system should be specified in this step.</p> <p>Requirements should also take into account other particularities of the legacy system. For example, if the legacy system generates/manages persistent identifiers, these should be kept resolvable in the new system.</p>
<p>Expected outputs</p>	<p>The expected outputs of this activity consist of a list of all system migration requirements collected so far. The list should include, but is not restricted to, the following outputs:</p> <ul style="list-style-type: none"> • Migration strategy; • List of measurable success criteria; • Contingency plan; • Cut-over strategy; • Data and metadata requirements, including relevant standards; • Other specific system requirements.
<p>Examples of practice</p>	<p><i>“17 library catalogue systems had to be migrated into a new system and the strategy chosen was the “big-bang”. This migration involved millions of records.”</i>, British Library</p> <p><i>“Exported objects should contain legacy system unique IDs. This may</i></p>

	<i>include persistent identifiers (DOI, URN, Handle etc.), or where non-existent, the legacy system's generated unique ID. Objects should be retrievable in the target system by this field, so that search APIs are able to allow for creating automatic redirection of requests to objects in legacy system. Provided migration tool enriched the exported objects with PREMIS objectIdentifier fields with the legacy system name (as objectIdentifierType) and unique ID (objectIdentifierValue).”, Ex Libris</i>
Other relevant information	For more information about PREMIS and how this can be used to record metadata about the migration process consult [5, 6, 13].

2.2 Planning & design

ID	2
Activity	Planning & design
Description	<p>The planning and design stage follows the definition of requirements. In this stage the project manager is capable of building a project plan and designing all the specifications necessary to drive the development of migration routines and testing procedures.</p> <p>The project manager should take into consideration all the information collected during the previous steps and devise an appropriate migration plan that includes:</p> <ul style="list-style-type: none"> • All project requirements (including time and data requirements); • Success criteria; • Test plan; • Contingency plan; • Technical specifications of the migration process; • Resources (including human resources); • Project tasks, assignments and duration; • Project scheduling.
Expected outputs	A detailed project plan that sets the requirements for the success of the migration process. This plan should be sufficient for the development team to create all the necessary routines to complete the migration process and include aspects such as: project requirements, success criteria, test plan, contingency plan, technical specifications for data migration and validation, resources, task descriptions and schedule

2.2.1 Project planning

ID	2A
Activity	Project planning
Description	After the analysis & consultation stage we are ready to devise an appropriate project plan. In the plan one should describe the strategy

	<p>and approach, delineate the scope of migration, define a schedule, identify the necessary resources (human and other kinds), define technical and business requirements, customer expectations (goals), project deliverables, and create a detailed execution plan.</p> <p>Creating an effective migration plan is often quite challenging. Different types of data or components may require different migration approaches, and comprise different business and operational requirements, e.g. the downtime window may require creative ways of moving the data.</p> <p>The project plan, which is the end deliverable of the planning phase, will function as the blueprint for the migration implementation.</p>
<p>Implementation guidelines</p>	<p>The project manager should take into consideration all the information collected in the previous steps and devise an appropriate migration plan that includes:</p> <ul style="list-style-type: none"> • All project requirements (including time and data requirements); • Success criteria; • Test plan; • Contingency plan; • Technical specifications of the migration process; • Resources (including human resources); • Project tasks, assignments and duration; • Project scheduling.
<p>Expected outputs</p>	<p>The output of this task is a project plan that sets the requirements for the whole repository migration process. The plan should be improved with the specifications that result from steps 2.2.2 and 2.2.3.</p>
<p>Examples of practice</p>	<p><i>“An in-depth comparative study of the two systems is always undertaken. Among the questions asked are: 1) Do objects in legacy system conform to target system data model? If not, what steps need to be taken, and when? 2) What metadata schemas are supported? Does the legacy system contain preservation-relevant metadata (events), and will this need to be migrated? Will metadata transformation be necessary? 3) How are collections created and structured in the target system? Can this be migrated? If not, what steps need to be taken to retain necessary information to rebuild the collections in the target system?”</i>, Ex Libris</p> <p><i>“Planning is key. One cannot start too early, and it will take longer than one might think. Make it a formal project. You need on-going business commitment. Remember that the migration is not just a technical IT task.”</i>, British Library</p> <p><i>“A recommendation is to refrain, if possible, from migrating live collections. Instead, wait until they are completely loaded into the legacy system. Where this is not possible, the workflows of ongoing ingest</i></p>

	<p><i>projects that generate material for loading into the legacy system should also be migrated to the target-system-ready. In this case, it is preferable that these modified workflows be tested and set into action before the actual repository migration, so that migrated objects can be compared to objects created by the modified workflow during the testing phase. Detailed and clearly defined procedures and timelines on a per-collection basis, indicating which collections are complete as opposed to ongoing, disseminated to all parties involved, are the key to ensuring all data is migrated and to avoiding data duplication.”, Ex Libris</i></p>
--	---

2.2.2 Design of migration routines

ID	2B
Activity	Design of migration routines
Description	<p>During the data analysis step, the project team has already decided upon which information entities and data sources should be migrated. However, it’s in the design of migration routines phase that the actual mappings between the legacy semantic elements and the new semantic elements will take place.</p> <p>A migration project is the perfect opportunity for some cleanup. Repository owners are encouraged to sift and sort through information, removing out-dated or redundant information, thus reducing the volume of information to be moved. Data cleansing tools can be useful as they allow information to be brought up to current standards and its quality to be measured. However, the effort put into cleansing content should be dependent on the business impact if the content value is incorrect [18].</p> <p>Bear in mind that data cleansing is highly dangerous in the context of digital preservation without a thorough analysis. Deleting, for example, logs that describe transformations a digital object has undergone could endanger trust, if one is no longer able to track the sources of the data or metadata. One should be very careful what data is safe to delete and those reasons well documented and approved by the relevant stakeholders.</p>
Implementation guidelines	<p>This is the first real low-level task in the overall migration methodology. This step requires high performance technical skills in order to devise appropriate transformations and mapping specifications that will guide the developers through the processes of creating the actual software routines that will perform the data migration. Involve the people who really understand the data, how it is created and used.</p> <p>In some cases, these specifications can automatically be transformed into actionable routines that perform the entire data migration (or a</p>

	<p>large portion of it). Examples of these are the Extract/Transform/Load (ETL)² routines included in some Database Management Systems. In other contexts, specially designed software will have to be developed to tackle the transfer of data between the legacy and the target system.</p>
<p>Expected outputs</p>	<p>The expected outputs of this step consist of:</p> <ul style="list-style-type: none"> • Detailed entity and data mappings; • Detailed data transformation specifications; • Data cleansing specifications; • Other relevant data transformation specifications.
<p>Examples of practice</p>	<p><i>“In this step it is important to involve the people who really understand the data and how it is created and used. Ensure that the provenance/authenticity of the digital objects to be transferred is not compromised by the migration. Consider whether any reformatting of the digital objects in the repository is appropriate, for example: for normalisation (e.g. if migrating legacy data to an existing repository or migrating multiple diverse repositories to a single one); for reducing storage costs (e.g. by using compressed file formats, lossy or non-lossy). If considering reformatting of the objects, involve and get support of curators and end-users (the people for whom the repository is there ultimately).”, British Library</i></p> <p><i>“General recommendation is to limit migration to master copies and re-create derivative/access copies in the target system. Considerations are as: 1) Auditing – the target system is expected to be able to audit the generation of the derivative copies (date, tool); 2) technology changes - since the creation of the original derivative copy, pertaining to the tool and/or format and/or parameters may dictate that new derivative copies be generated; 3) simplicity.”, Ex Libris</i></p> <p><i>“We found it was necessary to develop a tool that, during export, could restructure legacy-system objects according to the data model of the target system, while allowing for metadata transformation (MARCXML to Dublin Core) and enrichment of the metadata to allow for collection reconstruction in the target system.”, Ex Libris</i></p>
<p>Other relevant information</p>	<p>Because organizations may have concerns about the cost and risk for database migration, Microsoft provides a tool, SQL Server Migration Assistant (SSMA), to automate the migration process. The latest SSMA</p>

² ETL is short for Extract, Transform, Load, three database functions that are combined into one tool to pull data out of one database and place it into another database.

	<p>v.5.2 supports migration from Oracle, Sybase, MySQL and Access databases to SQL Server. SSMA can be used to ease organization database-migration project. For more information visit http://www.microsoft.com/en-us/download/details.aspx?id=28763.</p> <p>The SCAPE project has produced a Digital Preservation Toolkit that comprises tens of open-source characterisation, conversion and quality assurance tools for various media types. These tools can be used to assist in the development stages of the migration process if file format migrations are expected to take place.</p>
--	---

2.2.3 Design of test plan

ID	2C
Activity	Design of test plan
Description	<p>After all the activities related to analysis are concluded, one should have all the information needed to devise an appropriate test plan. This should entail all the steps necessary to make sure that the migration has met all the requirements previously identified.</p>
Implementation guidelines	<p>The test plan should be as complete and specific as possible, i.e., it should be able to provide answers to questions such as:</p> <ul style="list-style-type: none"> • Does the target system contain the same number of metadata records of the legacy system? • Have all user-defined attributes been migrated? • Are there any unresolved encoding issues? • Was any file corrupted during the copying process? • Has the authenticity of the data in the new system been retained? • Do the original system invariants still hold in the new data model? <p>The test plan can be implemented entirely by scripts and automatic routines, manually or by a combination of both. In any case, humans ultimately check if the migration has been accomplished successfully, so in practice all test plans usually end up being a combination of automatic and manual checks.</p> <p>One must keep in mind that the migrated information has been restructured for the new system and that context has changed, hence it might be difficult to compare with the legacy system. However, success criteria and metrics should always be possible to be devised.</p> <p>The person responsible for the repository should also be consulted in order to assess their opinion on the thoroughness of the quality assurance plan. This testing plan can, of course, be revised during the</p>

	<p>following stages of the project.</p>
<p>Expected outputs</p>	<p>The outcomes of this step consist of:</p> <ul style="list-style-type: none"> • A detailed specification of what is expected to be assessed and how; and what the expected results are. (What indicates success?) • The identification of the team that will be responsible for assessing the quality of the migration (technical and non technical personnel); • A checklist or test plan to be used by the human evaluators (may be enhanced in later stages).
<p>Examples of practice</p>	<p><i>“Testing and quality assurance is key, and this can be challenging at times. Internal experts on the data might not have the appropriate testing skills (e.g. writing and executing comprehensive testing scripts), and external testers do not have detailed knowledge of the data. Sometimes it can occur that the expert knowledge is not available (left the organisation) and the “system” becomes the data expert. One needs ongoing business commitment for quality assurance and testing. This is not just a technical task.”</i>, British Library</p> <p><i>“In very simple repository migration processes, the absolute minimum quality assurance assessment routine is determining the number of metadata records in the legacy system and comparing it with the number of records in the target system moments after migration.</i></p> <p><i>Unless transformation of object files has taken place during transfer, checksum comparison should be done for all digital objects. One does not want corruption of files to have been introduced during the migration and gone undetected.”</i>, KEEP SOLUTIONS</p> <p><i>“Testing should include all interfaces with external systems or components. See if discovery systems are able to harvest and present data properly and with no regression to end-user experience. New viewers, if these exist, should be tested to confirm files are delivered as expected. Plugins such as technical metadata extractors should be tested as well.”</i>, Ex Libris</p> <p><i>“We have identified a set of statistics to gather on the migrated data, we have then compared these figures with similar stats or anticipated stats on the legacy system. Using this approach we have discovered few corner cases that had to be treated individually (e.g., URL canonicalization mapping several entities to one key).”</i>, Internet Memory Foundation</p>

	<p><i>“As a rule, a fixity check should be run every time a file is copied. This is likely to include: export from legacy storage location to staging location and copy from staging location to target system storage. Since checksum values already existed in the legacy system, they were exported with the objects’ metadata and verified by the target system.”, Ex Libris</i></p>
Other relevant information	<p>SCAPE has produced quality assurance tools for various media types such as image, video, text and audio. These tools can be used to assess the correctness of a file format migration process.</p>

2.3 Development

ID	3
Activity	Development
Description	<p>After analysis and planning stages, we have all the necessary elements to begin the development of all migration routines and testing procedures. This stage is where migration tools are actually going to be built (or configured) according to the specifications created in the previous stages.</p> <p>Development is divided into two categories: 1) development of migration routines that handle the transformation and transference of data between the two systems and 2) development of quality assurance routines based on the test plan previously designed.</p>
Expected outputs	<p>Examples of expected outputs are:</p> <ul style="list-style-type: none"> • Software source-code and/or binaries; • Configuration files; • Software documentation (user manual and source documentation); • ETL routines; • XSLT; • Scripts; • Migration checklists for manual assessment; • Implementation checklists to assist implementers in making sure that the necessary environment to run the migration routines is in place (i.e. installation instructions).

2.3.1 Development of migration routines

ID	3A
Activity	Development of migration routines
Description	<p>This step is where the migration developers come in and implement the routines previously designed. This may consist of building ETL (Extract,</p>

	<p>Transform and Load) jobs, specialised programs or scripts that implement the mappings and specifications created in the design phase. All mappings, quality rules, and field validations should be built into the migration routines.</p> <p>Keep in mind that one may have to return to this step as many times as necessary to drive migration errors down to zero. Revisiting the development stages for six, seven or eight times is not unheard of [21].</p>
<p>Implementation guidelines</p>	<p>The tasks included in this step are mainly related to software development, i.e. writing source-code, testing and documenting. Software engineering and project management methodologies should be used to ensure the quality of the resulting product and that the schedule is respected.</p> <p>Based on the specifications created during the planning stages, developers make use of their skills to implement those specifications on the most appropriate technology. The selection of technology depends on a variety of factors such as:</p> <ul style="list-style-type: none"> • The type of underlying platform that supports the legacy and target systems - e.g. certain programming languages are more capable of running in certain operating systems. For instance, one will not be able to run ETL routines on systems that are not RDBMS-centric; • The type of migration strategy - e.g. depending on the migration strategy, a one-go migration procedure might not be possible; • The type of data to be migrated - e.g. XML data is easier to process using XLST transformations; • The size of data - e.g. for large XML files a DOM-based parser³ might not be possible to use due to memory constraints; • Specific project requirements – e.g. certain programming languages or coding techniques might not have the performance necessary to meet the time constraints of the project. <p>Migration routines might interconnect the legacy and the target system directly or might be based on the extraction of data to an intermediate format that will subsequently be transformed and ingested on the target system. This should not be regarded as a standard ingest process where Submission Information Packages (SIP) are prepared and fed to a repository system. Data from the legacy system is expected to include more information than a standard SIP is capable of carrying (e.g. technical metadata, several representations of the same content coupled with event information, other relevant preservation metadata, etc.) so</p>

³ The Document Object Model (DOM) is a cross-platform and language-independent convention for representing and interacting with objects in HTML, XHTML and XML documents. DOM-based parsers typically load the entire XML file into RAM before the developer has access to its methods. For large-sized XML files this strategy might not be possible due to large amount of RAM that would be necessary to load the file.

	<p>this is in fact a special ingest procedure that instead of SIPs is expecting to receive Archival Information Packages (AIP).</p> <p>Include a facility for logging the migration. The migration may fail and if so, it will be important to know when and why this happened.</p>
Expected outputs	<p>Examples of expected outputs consist of:</p> <ul style="list-style-type: none"> • Software source-code and binaries; • Configuration files; • Software documentation (user manual and source documentation); • ETL routines; • XSLT; • Scripts.
Examples of practice	<p><i>“An iterative migration process works best. Run a test data migration, produce reports (based on test criteria) and review. Make changes, repeat. Make sure the reporting mechanism following a migration run is good enough to support identification and fixing of errors. Its relatively easy to get 90% of the migration free of errors, then it becomes more and more difficult as you approach 100% correctness. You need to know when to stop.”</i>, British Library</p> <p><i>“In the migration process of our Web archive, distributed copy scripts and MapReduce migration job have been developed to improve the performance of the transfer”</i>, Internet Memory Foundation</p>
Other relevant information	For more information about SIPs, AIP and ingest procedures, read [7]

2.3.2 Development of testing routines

ID	3B
Activity	Development of testing routines
Description	<p>This step consists in building the test routines that will validate the success of the migration. The deliverables that come out of this step may include validation checklists, testing scripts or dedicated programs that report any anomaly in the migration process execution.</p>
Implementation guidelines	<p>This step also consists of developing software, but in this case it is software dedicated to make sure that the repository migration routines run according to the specifications. The development should be guided by the success criteria defined on the project plan and be agnostic in terms of technology and methods used for the development of the migration routines.</p> <p>It is advisable that the testing routines are developed by a different team than the one that developed the migration routines. Source-code between both teams should not be shared as this increases the chance</p>

	of mistakes being propagated between both projects and errors going through undetected.
Expected outputs	<p>Examples of expected outputs are:</p> <ul style="list-style-type: none"> • Testing software in source-code and binaries; • Configuration files; • Software documentation (user manual and source documentation); • Migration checklists for manual assessment; • Implementation checklists to assist implementers in making sure that the necessary environment to run the migration routines is in place.
Examples of practice	<i>“We have developed MapReduce jobs producing stats about the migrated data for enhanced performance”, Internet Memory Foundation</i>

2.4 Setup & testing

ID	4
Activity	Setup & testing
Description	The setup and testing stage consists of creating the infrastructure where the target repository system and the migration solution are going to work. For the migration to be effective, one should prepare the infrastructure for full-scale trials of the target system against migrated data. If the specifications are thorough and accurate, this phase should be routine and predictable. A strong technical background and documentation will greatly simplify the provisioning effort [19].
Expected outputs	<ul style="list-style-type: none"> • A platform to run the testing implementation of the target repository system and the migration routines; • Testing implementation of the target repository system deployed; • Migration and testing routines deployed; • Automatic testing reports; • Target system with complete or partial data that allows for human inspection; • Manual checklists application results; • Test plan report that focuses on how well the rehearsal met the project requirements and success criteria.

2.4.1 Target environment provisioning

ID	4A
Activity	Target environment provisioning

Description	<p>During the target environment provisioning phase, the destination infrastructure and software is prepared for the data transfer. This includes setting up hardware, operating systems, configuring storage volumes, installing the new repository system and configuring it to accommodate migrated information and business rules.</p> <p>Provisioning for a no value-added strategy is usually simple but for a completely new system it may be more complex. However, using information generated from the analysis and design steps, it will be possible to automate many of the provisioning tasks [17].</p>
Implementation guidelines	<p>This task consists mainly in provisioning enough resources (memory, CPU, storage, etc.) to execute the target repository system in production mode but also to accommodate the migration process.</p> <p>The migration process might require additional resources in order to comply with the project time constraints. An “elastic” infrastructure plays an important role here as one may easily expand the available resources during migration and testing, and then shrink it to the right amount of resources for production mode.</p>
Expected outputs	<p>The expected outcomes of this step consist of:</p> <ul style="list-style-type: none"> • A platform to run the target repository system and the migration routines. This may be a temporary testing facility; • Target repository system deployed; • Migration and testing routines deployed.

2.4.2 Rehearsal & testing

ID	4B
Activity	Rehearsal & testing
Description	<p>After the migration routines have been fully developed and before the definitive migration is executed one should perform a series of migration rehearsals in order to make sure that all the requirements have been correctly implemented.</p>
Implementation guidelines	<p>Rehearsal migrations may be partial or complete. A complete end-to-end migration in the pilot environment is of course desirable. However, depending on the amount of information to be moved, this may not be possible due to time constraints or even due to the stress that this may cause on the production repository.</p> <p>After each rehearsal, one should run the entire test plan. The output of this activity will dictate whether one can move on to the definitive migration or should go back to the drawing board and revise mappings, routines or even the project plan. For example, if testing shows that</p>

	<p>allowable downtime would probably be exceeded, the migration methodology will need to be revisited [17].</p> <p>The decisive test is to provide the populated target system to the users who assisted in the analysis and design of the migration project. Invariably, users will begin to identify historical data elements that must be migrated that were not apparent to them during the analysis/design sessions [9].</p> <p>Depending on the results of this testing step, one may move on to the definitive migration or go back to the drawing board and revise the migration routines or even the success criteria and project requirements (e.g. if meeting the success criteria would imply a massive investment in additional infrastructure, it may be more advantageous to ease the project requirements).</p>
<p>Expected outputs</p>	<p>Examples of expected outputs from this step are:</p> <ul style="list-style-type: none"> • Automatic testing reports; • Target system with complete or partial data that allows for human inspection; • Manual checklists to support human inspections; • Test plan report that focuses on how well the rehearsal met the project requirements and success criteria.
<p>Examples of practice</p>	<p><i>“A dedicated test environment is required. Separate from the production environment. This needs to be planned for and costed.”</i>, British Library</p> <p><i>“Several iterations of rehearsal on a representative sample of data to migrate have been done. Rehearsals replayed until all related issues were solved.”</i>, Internet Memory Foundation</p> <p><i>“The following considerations have been taken into account while determining a representative sample data set: 1) diversity, i.e. samples should include representatives of data with all structures, formats, sizes etc.; and 2) resource allocation, i.e. tools and manpower required to analyse and review the data.”</i>, Ex Libris</p>
<p>Other relevant information</p>	<p>For more information about cost models monitor the results of the 4C Project at http://4cproject.net.</p>

2.5 Execution

<p>ID</p>	<p>5</p>
<p>Activity</p>	<p>Execution</p>
<p>Description</p>	<p>If the migration trials run without issues, one may move on to the</p>

	<p>execution stage. This stage consists of the execution of the previously developed migration routines and migrate all the information from the legacy system onto the new target system.</p> <p>This is where all the effort invested so far is going to be put into practice and any glitch in the process might mean that the contingency plan will have to be employed (see Section 2.2.1).</p>
Expected outputs	<p>A new repository system installed, with all the necessary data transferred into it, testing and quality assurance procedures have accepted the migration as successful, and the whole system is ready to be put into use as the production system.</p>

2.5.1 Execution of migration routines

ID	5A
Activity	Execution of migration routines
Description	<p>The execution step consists of running the migration routines developed in step 3A. This will be the “actual” migration.</p> <p>Before proceeding with the migration, it is important to review all the guidance and best practices of the previous steps. Ensure that the objectives are being met and contingency plans are in place [18].</p> <p>Additionally, it is important to keep in mind that any data ingested during the rehearsal steps must be erased before the actual migration, that is, if the rehearsal ingests have been done in the target production system.</p>
Implementation guidelines	<p>During execution one should have the contingency plan ready to be used in case the migration execution does not run as expected. This means that the entire migration process should be monitored on a frequent basis as one would not like to wait for several days (or months) for a migration to finish in order to come to the conclusion that an important piece of information was not copied correctly or that the progress is not moving as timely as expected. Keep in mind that rehearsals are like to be run on just a portion of the data, so errors can still occur.</p> <p>Keeping a running copy of the original system ready in case one needs to go back is always a good strategy. It is also common practice to keep the migration team ready for action in case anything goes wrong.</p>
Expected outputs	<p>A new repository system installed, with all the necessary data transferred into it, testing and quality assurance procedures have accepted the migration as successful, and the whole system is ready to be put into use as the production system.</p>
Examples of practice	<p>On large-scale repository systems, the migration process can take several</p>

	<p>months. For example, the Archives New Zealand have decided to move digital content from Fedora Commons to Ex Libris' Rosetta Digital Preservation System. This process took roughly 12 months to migrate 40TB of data. Data was ingested into the new repository during the night and validation steps would occur during the day [10].</p> <p><i>“Thorough surveillance of the migration process is key. The test criteria from previous processes help to assess the validity of the data being migrated. The time/resource estimations are re-validated during the process”, Internet Memory Foundation</i></p> <p><i>“Institutions should refrain from migrating from a live productive system, due to the following considerations: 1) Performance - large data exports are likely to have a negative effect on users' experience; 2) Data Manipulation - Data massage is typically needed to accommodate differences between data models. In some cases it may be advantageous to perform this pre-export. Production environment's integrity should be preserved by performing all manipulations in a cloned environment. To this end, a cloned migratory environment is recommended if the resources exist to do so. In order to reduce time and cost, it should be possible to point the migration temporary environment to the legacy environment's storage. The migration server should be given read-only access to storage. A further measure to reduce migration time and cost is to migrate only the metadata with links to legacy storage. This can be done either on a metadata basis (METS filesec) or an operating system basis (shortcuts or symbolic links). In such cases target system should have (read-only) access to legacy storage. This will also relieve the need of a fixity check, since it eliminates a filestream copy.”, Ex Libris</i></p> <p><i>“There is an ongoing need for communication across all the stakeholders, especially during lengthy and complex migration projects. Keeps everyone engaged and reassured. Helps flag issues and changed requirements/constraints.”, British Library</i></p>
Other relevant information	For more information about the Archives New Zealand repository migration process, read [10]

2.6 Validation

ID	6
Activity	Validation
Description	After the full migration, a complete run of tests should be executed on the target platform. This will ensure that the process has run according to plan and that no errors have taken place. This stage also includes the

	creation of reporting materials to document the overall process and the cut-over to the new system.
Implementation guidelines	Validation consists of running the test routines on the target system during or after the definitive execution of migration routines.
Expected outputs	<ul style="list-style-type: none"> • Automatic testing reports; • Manual checklists application results; • Migration report that focuses on how the project met the requirements and success criteria; • A list of issues that need to be fixed in the migration routines (if any); • A well organised collection of validation reports and other relevant evidences of the success of the migration; • A new repository system in production mode.
Examples of practice	
Other relevant information	

2.6.1 Execution of testing routines

ID	6A
Activity	Execution of testing routines
Description	As in the migration rehearsal phase, this step consists of rerunning the entire test plan against the new populated system to make sure that everything went according to plan.
Implementation guidelines	<p>The test plan may consist of set of automatic and manual verifications. Part of the testing routines could be embedded in the migration software thus allowing for real-time information about the quality of the overall migration process.</p> <p>If any inconsistency is detected either by the testing routines or users, the contingency plan might have to be put to practice and the migration process repeated after fixing the uncovered issues.</p> <p>In some cases, quick fixes can be made on the running system without having to go through a completely new migration. One might just re-import some data without having to reboot the whole migration process.</p>
Expected outputs	<p>Examples of expected outputs from this step are:</p> <ul style="list-style-type: none"> • Automatic testing reports; • Manual checklists to support human inspections; • Migration report that focus on how the project met the requirements and success criteria; • A list of issues that need to be fixed on the migration routines (if

	any).
Examples of practice	<p><i>“The verification phase should acknowledge what is expected to be constant in the target system, as opposed to what is expected to differ, i.e. checksums - all migrated objects’ checksums’ values are expected to conform to respective values in the legacy system. If target system supports loading with exported checksum values and validating these, this should be an automatic process; Object count - count of migrated and imported objects should be identical. Count should exclude derivative copies, if these are not part of the export and technical Metadata as nuanced format identification tools and new technical metadata extractors used by the target system are likely result in different technical metadata than that in the legacy system. Data analysis and reporting tools should be utilized to confirm deviations between legacy and target systems appear only where expected and not otherwise.”</i>, Ex Libris</p> <p><i>“If an MD5 or SHA-1 checksum was present for a file in Fedora, it was re-calculated after the file was extracted and compared against the stored value. Warnings were produced for: missing checksums, unsupported checksum types, failed checksum checks resulting in a failure for the item.”</i>, Archives New Zealand [10].</p> <p><i>“Our final audit was done via the [the legacy] Archway database, where all items are stored. It has been compared with the original list of item IDs stored in IDA repository and then with the [target] Rosetta Oracle database of item IDs in our production environment. To simplify matters, if an IDA item has an associated Rosetta item ID, we can say that it has been synchronised via the Rosetta publishing process with Archway and therefore successfully migrated. We have only identified two duplicate items ingested during the entire operation.”</i>, Archives New Zealand [10].</p> <p><i>“In the process of migrating data from the [legacy repository] into Rosetta, we chose an approach more suitable for large amounts of data - we haven’t tried to solve all issues in the Rosetta Technical Analyst workbench, rather moved all SIPs caught in technical analyst workbench to our own quarantine location. There the digital objects were analysed, fixed in bulk with an agreed solution, and the whole SIP re-submitted into Rosetta.”</i>, Archives New Zealand [10].</p>

2.6.2 Reporting

ID	6B
Activity	Reporting
Description	The reporting step is run side by side with the execution of testing

	<p>routines. This step basically consists of collecting all the evidence and reports produced by the testing routines in order to document and finalise the validation phase. This constitutes proof that the migration was a success and may very well prevent future legal annoyances or disputes.</p> <p>An additional step is to save and archive all the migration routines. Data migration is often a one-time exercise, however, with the right tools, protocols and mappings, migration routines can be reused in future projects within the organisation or in other organisations. A documented report of the migration process will serve as a reference guide and may also help to diagnose and fix post-migration issues [18].</p>
Implementation guidelines	<p>Collect and organise all the reports produced during the course of migration process. Store those reports on a safe place as they will provide valuable evidence and documentation about the migration process.</p> <p>For example, during audits or inspections it is common to look for evidence that certain activities have taken place within an organisation.</p> <p>Additionally, if something is found to be wrong long after the system goes into production, one may resort to the documentation collected in this step to find out the causes of the recently found issue and develop an advised corrective action for it.</p>
Expected outputs	A well organised collection of validation reports and other relevant evidences of the success of the migration.
Examples of practice	Remember that the reporting information should also be preserved, so it might be a good idea to ingest it into the target repository to insure its long-term preservation.

2.6.3 Cut-over

ID	6C
Activity	Cut-over
Description	<p>Once the target repository has been built up and all the legacy information has been migrated, the new system is then ready to run.</p> <p>Based on the cut-over strategy defined in the planning stages of the project, one should move forward as quickly as possible to put the new repository system in production mode.</p>
Implementation guidelines	<p>There are three main strategies to accomplish the transition between the legacy and the new repository system [3]:</p> <ol style="list-style-type: none"> 1. The big-bang strategy consists of switching off the legacy repository system and start using the new replacement system; 2. In a phased interoperability strategy, the cut-over is performed in

	<p>small, incremental steps, each of these replacing a few components (applications or data) of the legacy system(s);</p> <p>3. In the parallel operations strategy, the legacy repository and the target system operate simultaneously, with both systems performing all operations. During this period, the target system is continually tested; once it is fully trusted, the legacy system is retired.</p> <p>The big-bang strategy is, in many cases, too idealistic because of the risk of cutting over to the new system in a single step putting the whole organisation’s information flow in an untested and thus untrusted system.</p> <p>On the other hand, the phased interoperability strategy is potentially very complex. To be successful, this method requires the migration team to split legacy system applications into functionally separate modules or to separate the data into portions that can be independently migrated. The monolithic and unstructured nature of most legacy systems makes such an approach difficult, if not impossible. A concrete transition strategy for a particular migration project would probably involve a combination of these approaches, applied to different repository components [3]. Nonetheless, this approach may be most appropriate if migrating multiple legacy systems to one. Each legacy system might be a phase, for instance.</p> <p>In parallel operations both systems should to be synchronized at all time, enabling the migration team to assess that the new system is operating as expected. Synchronization of updates in the legacy repository and the target it is by no means trivial and should be well planned.</p> <p>Switching to the new system may also involve the change of environmental variables and external systems such as networks addresses, firewall rules, DNS, persistent identifier registry updates, new links on the organisation’s Website to direct users to the new system, etc. In parallel operations scenarios, particular care should be paid to interfaces with other systems.</p> <p>Such updates to the environment should also be carefully planned so that they can be implemented as quickly as possible to minimise downtime.</p>
<p>Expected outputs</p>	<ul style="list-style-type: none"> • The output of this step consists of the new repository system in production mode with all the necessary environmental changes to enable it to run appropriately; • The environments on which the legacy and target repositories function should be updated to support on-going running operations.
<p>Examples of practice</p>	<p><i>“In a project that involved the migration of a library catalogue from a commercial to an open-source system, both systems were operated</i></p>

	<p><i>simultaneously by the library staff over the course of a month. Reports from both systems were produced and compared every day to make sure that circulation records and updates to the catalogue were perfectly synchronised between both systems. It was only after this confirmation that the legacy system was disabled.”, KEEP SOLUTIONS</i></p> <p><i>“External systems, such as harvesters or gateways, should be updated to request data from the target system, and their data (e.g. delivery URLs) should be updated as necessary.”, Ex Libris</i></p>
--	--

2.7 Wrap up

ID	7
Activity	Wrap up
Description	<p>After the new repository has gone into production, there are a few activities that one should consider. These include training users and repository managers to use the new system, collect, build and archive all the project documentation and deliverables, and provide maintenance and support to new system in case of an emergency or if any tuning is necessary.</p> <p>Implementation of this activity consists of training users to be proficient on the new system, writing the final reports of the project and providing helpdesk to end-users as well as technical support to the running platform.</p>
Expected outputs	<ul style="list-style-type: none"> • End-users trained to work with the new system; • All textual and non-textual deliverables of the project stored in a discoverable and safe archival environment; • Issues submitted by end-users solved by a support technical team.

2.7.1 Training

ID	7A
Activity	Training
Description	<p>No system adoption is complete without training of its end-users. Through their insightful questions, one will quickly learn how the target system should be reconfigured or enhanced, both crucial inputs for this and future migration projects.</p> <p>As training is known for having a short lifespan, it is normal to postpone training until the end of the project. However, training key end-users may be done earlier in the project to assist in the configuration of the system [19].</p>

Implementation guidelines	<p>Follow as much as possible good practice on training. These include planning training sessions, creating adequate training materials and future reference documentation, evaluate the effectiveness of the training session and include hands-on exercises for users to be able to practice.</p> <p>It is important to point out that hands-on training might have to be done on a replica of the production system (eventually on a demo site) so that the system in production does not end up with test data that on a preservation environment might be difficult (or even impossible) to erase.</p>
Expected outputs	<p>Expected output of the training activity are:</p> <ul style="list-style-type: none"> • Training plan; • Training materials and reference documentation; • Hands-on exercises; • End-users able to work with the new system.
Examples of practice	<p><i>“Plan and schedule training well before the repository has gone into production, and ensure that everyone who must be able to use the production system on the “go-live” day is able to so do. Consider how many users there will be. They will need appropriate training and possibly new equipment that will need to be budgeted for and procured. Don’t forget procurement of replica/demo site if necessary. If the number of users is large, then this needs to be considered carefully during the planning stages.”</i>, British Library</p> <p><i>“An additional activity post-implementation is a project review, involving the users. This should, hopefully, help flush out any remaining issues, keeps users confident that it hasn’t been a “cut-and-run” implementation, and will identify any lessons learned for future repository migration projects. Involving the users makes sure that the lessons learned do not just reflect the technical team’s point of view. This review will then provide input to the documenting stage 2.7.2.”</i>, British Library</p>

2.7.2 Documenting

ID	7B
Activity	Documenting
Description	<p>After the migration has been completed, the project team should compile all the source-code, migration statistics, test reports, designs and plans and prepare a report to highlight what worked, what didn’t work and lessons learned. The report should be shared with all members of the migration team. These types of reports are critical in building a repeatable and consistent process through continuous improvement</p>

	[11].
Implementation guidelines	<p>Collect all the deliverables that resulted from the migration project and use them to write one final report that documents the entire migration project. Lessons learned should be an important chapter of this report, as it will improve future migration endeavours.</p> <p>Attach all the additional information that you find relevant to document the decision process (e.g. data mappings).</p> <p>All of the resulting documentation and non-textual deliverables of the project should be organised and classified (including source-code). One may consider ingesting those materials into the new repository system for long-term preservation.</p>
Expected outputs	This activity results in the archival of all textual and non-textual deliverables of the project into a safe archival environment.
Examples of practice	<i>“Each issue has been thoroughly documented and documentation saved in the Archives content management system (CMS). CMS IDs of the documentation files were then added into metadata of the corrected digital objects. The documentation consists of the problem description with links to relevant file format documentation. There is a list of options for dealing with the problem and finally the decision about the preferred solution. Another part of documentation is about how the solution was tested. Custom scripts are also stored in the organization CMS. The idea behind this is that all changes to files have to be documented and referenceable from the item metadata so that future users can understand what was done and why.”, Archives New Zealand [10].</i>

2.7.3 Supporting

ID	7C
Activity	Supporting
Description	The supporting phase consists of keeping a team of technicians ready to assist users with any question or operational difficulty.
Implementation guidelines	<p>Post-migration issues may be of informational nature (e.g. information missing, bad mappings, etc.) which, in this phase, can usually be fixed directly on the production system; or system nature (e.g. bad configuration, bad tuning, among others). It could also consist of questions submitted by end-users about topics that were not covered during the training sessions or were eventually forgotten.</p> <p>The goal of this activity is to provide prompt technical assistance to any of these issues so that day-to-day operations do not get affected in a significant way.</p>

Expected outputs	<p>This activity consists of solving issues submitted by end-users or correct system and environmental problems that are detected on the production system.</p>
Examples of practice	<p><i>“Issues related to migration are expected to be reported during the first few months of the new system going live. When the number of reported issues falls bellow a certain level, the migration team can be officially released from the project, and normal operations and support will then take place.”</i>, KEEP SOLUTIONS</p>

3 Conclusions

Repository migration is an inevitable process which any institution that hosts or manages a digital repository will have to go through. It is just a matter of time before the repository becomes incompatible with current technologies, inadequate to serve the business needs of its institution or no longer being able to meet the expectations of its users.

Although a complex and risky process, with the proper guidance and preparation, these risks can be minimised.

A comprehensive methodology, a well prepared team and clearly defined project goals are always a good recipe for success. Even so, repository migration processes often cause major disruptions as a result of downtime or performance issues, which can have a truly negative impact on users' perception of system quality, trustworthiness and future productivity.

Major risks to the repository include: data loss, corrupt data, changed meaning of data, loss of service (availability/downtime), loss of functionality, etc.; additional project risks might include: running over budget, running over schedule, scope-creep, etc. To prevent these problems, organisations need a consistent and reliable methodology that enables them to analyse, plan, design, develop, migrate and validate the migration process. Potential pitfalls can be avoided by following the best practices presented in this document.

4 Implementation checklist

1. Analysis & consultation	Yes	N/A	Notes
Has a comprehensive set of technical documents that describe the legacy system been collected, organised and classified?			
Has a comprehensive set of technical documents that describe the target system been collected, organised?			
Have conversations about the data entities that exist in the legacy system, which should be migrated, and which entities or periods of time can be discarded been held and have the resulting decisions been recorded into minutes?			
Have the security levels of each data entity been determined?			
Have project staff members signed individual non-disclosure agreements?			
Has the migration strategy been set?			
Have the high-level requirements of the migration been set?			
2. Planning & design	Yes	N/A	Notes
Have the project low-level requirements been established?			
Have the success criteria for the project been defined and measurable metrics been identified?			
Has a validation/test plan been devised?			
Has a contingency plan been devised?			
Have the detailed technical specifications for the data migration and validation routines been created?			
Have the human and material resources necessary for the success of the project been identified and quantified?			
Have task descriptions, their sequence, assignments, duration and scheduling been defined, i.e., do we have a proper project plan?			
3. Development	Yes	N/A	Notes
Have migration routines been developed according to the specifications?			
Have testing routines been developed according to the specifications?			
Is the source-code well documented?			
Have the user manuals been written and revised?			
Have checklists for manual assessment of the migration been devised?			
Have checklists or instructions been created to aid implementers in the deployment and setup			

migration and test routines?			
Are all the code and binaries stored in a safe place and versioned (e.g. CSV, SVN, Git or other)?			
	Yes	N/A	Notes
4. Setup & testing			
Has the platform necessary to run the target system been adequately provisioned?			
Has the target repository been deployed and configured?			
Has the testing facility been prepared (if not using the target production system for testing purposes)?			
Have the migration and testing routines been deployed?			
Has the test plan been executed in rehearsal mode?			
Have manual verifications been done making use of the created checklists?			
Have the results of testing (manual and automatic) been analysed and archived?			
5. Execution	Yes	N/A	Notes
Have the final migration and validation routines been run?			
Did it all go according to plan, i.e. have the success criteria been met?			
6. Validation	Yes	N/A	Notes
Has the automatic test routines been executed after the final migration?			
Have manual verifications been done on the final migration, making use of the created checklists?			
Have the results of testing (manual and automatic) been analysed and archived?			
Has a final migration report been written focusing on the fulfilment of the success criteria?			
Have you accounted for all the environmental changes that need to be done to make the new system go into production mode (e.g. change DNS settings, opening firewall ports)?			
Has the new system gone into production mode?			
Has access to the legacy system been restricted?			
7. Wrap up	Yes	N/A	Notes
Have end-users been trained appropriately to			

operate the new system?			
Have all the deliverables (textual and non-textual) been archived?			
Have you assigned a support team to accompany operations during the weeks/months after the system gone into production?			

5 References

- [1] Bearing Point Inc. 2008. Data Migration through an Information Development Approach - A management overview.
- [2] Bisbal, J. et al. 1999. Legacy information systems: issues and directions. *IEEE Software*. 16, 5 (1999), 103–111.
- [3] Brodie, M.L. and Stonebraker, M. 1995. *Migrating legacy systems: gateways, interfaces & the incremental approach*. Morgan Kaufmann Publishers Inc.
- [4] Burry, C. and Mancusi, D. 2004. How to plan for data migration. *Computer World*.
- [5] Caplan, P. 2009. Understanding PREMIS. *Context*. 26, (2009), 26.
- [6] Committee, P.E. 2008. PREMIS Data Dictionary for Preservation Metadata. *Preservation*. 37, March (2008), 224.
- [7] Consultative Committee for Space Data Systems 2002. *Reference Model for an Open Archival Information System (OAIS) - Blue Book*. National Aeronautics and Space Administration.
- [8] Harris, L. 2010. IPM 11g Migration Best Practices. *Nexus'10* (2010), 27.
- [9] Hudicka, J.R. 1998. An Overview of Data Migration Methodology. *Select Magazine*. (1998), 5.
- [10] Hutař, J. 2013. Archives New Zealand Migration from Fedora Commons to the Rosetta Digital Preservation System. *iPRES2013 Proceedings* (Lisbon, Portugal, 2013).
- [11] IBM Global Technology Services 2007. *Best practices for data migration - Methodologies for planning, designing, migrating and validating data migration*.
- [12] Jantz, R. 2005. Digital Preservation: Architecture and Technology for Trusted Digital Repositories. *DLib Magazine*. (2005), 1–17.
- [13] Lavoie, B.F. 2008. PREMIS With a Fresh Coat of Paint: Highlights from the Revision of the PREMIS Data Dictionary for Preservation Metadata. *DLib Magazine*. 14, 5/6 (2008), 1–12.
- [14] Lawless, D. et al. 1997. The Butterfly Methodology : A Gateway-free Approach for Migrating Legacy Information Systems. *Proceedings Third IEEE International Conference on Engineering of Complex Computer Systems Cat No97TB100168*. (1997), 200–205.
- [15] De Lucia, A. et al. 2008. Developing legacy system migration methods and tools for technology transfer. *Software Practice and Experience*. 38, 13 (2008), 1333–1364.

- [16] Mohanty, S. 2004. Data Migration Strategies, Part 1. *Information Management*.
- [17] Network Appliance Inc. 2006. *Data migration best practices*.
- [18] Open Text Corporation 2009. *Top 10 best practices in content migration*.
- [19] Pick, B.G. 2001. Data Migration Concepts & Challenges.
- [20] Rahgozar, M. and Oroumchian, F. 2003. An effective strategy for legacy systems evolution. *Journal of Software Maintenance and Evolution Research and Practice*. 15, 5 (2003), 325–344.
- [21] Utopia Inc. 2009. *Data migration management - A methodology: Sustaining data integrity after the go live and beyond*.



Preservation of research data

Best practice guidelines and
recommendations



Table of Contents

1	Introduction	5
1.1	Audience.....	5
1.2	Approach	5
1.3	Scope	6
2	Background Context	8
2.1	Organisations.....	8
2.1.1	Models for storing and curating research data	8
2.2	Repository management.....	9
2.2.1	Trustworthy repositories of data.....	9
2.3	Data creators	10
2.3.1	Research Data Lifecycle/process	10
2.3.2	Perceived threats to data access and reusability	11
3	Guidelines and Recommendations	12
3.1	Organisation	13
3.2	Repository Management.....	19
3.2.1	Repository & content set-up	20
3.2.2	Preservation activities	26
3.3	Data concerns.....	29
4	Conclusions	37
5	Bibliography and useful websites	38
5.1	Useful bodies and conferences	38
5.2	Policy and strategic: High level view	38
5.3	Costs and benefits	39
5.4	Practical Guidance.....	39
5.5	Repository Management & Infrastructure.....	40
5.6	Data citation	41
5.7	SCAPE related papers	41



6 Introduction

The SCAPE project⁴ aims to enhance the state of the art in digital preservation with a particular emphasis on the scalability of its solutions: that is, their capacity to handle digital objects that may be very numerous, individually very large, heterogeneous or complex. It is clear that much research data has some of these characteristics of scale. Even in domains where the sheer data volumes are not so large, the data is likely to have complex semantics and to have undergone processing which might need to be recorded in order that future users may understand the provenance of the data.

The motivating force of the SCAPE project is scalability, interpreted in several dimensions: number of objects, size of objects, complexity of objects, and heterogeneity of collections. More specifically, the project aims to enhance the state of the art of digital preservation in three ways: by developing infrastructure and tools for scalable preservation actions; by providing a framework for automated, quality-assured preservation workflows and by integrating these components with a policy-based preservation planning and watch system.

6.1 Audience

This guide and associated recommendations has two audiences: those who manage and curate data by providing data centres, repositories or archives for others to use (see sections 3.1 and 3.2) and those who create and deposit data within those repositories (see section 3.3). It has three foci for discussion: organisations; repository management and data.

The guidance is not explicitly differentiated for these two audiences, because they must necessarily work together and have a common understanding of the issues in preservation approach. However it is expected that working researchers are likely to access this material through summaries to be provided through other project dissemination methods.

6.2 Approach

Based on a broad literature review we have collated guidelines and recommendations for the preservation of research data. These have been further enhanced by the experiences of the SCAPE partners and lessons learnt within the project.

Addressing our two key audiences the document is split into three areas: organisational level considerations; repository management level considerations and those relating to specific data concerns. We are using the term “repository” in its broadest sense, as a location of collections of information, rather than to indicate a preference to particular software.

⁴ www.scape-project.eu

We expect the first two to be of interest to those who are responsible for the management and curation of research data and the third to be of interest to researchers who are creating data which will be deposited in a repository.

See the diagram below for a pictorial representation of this structure.

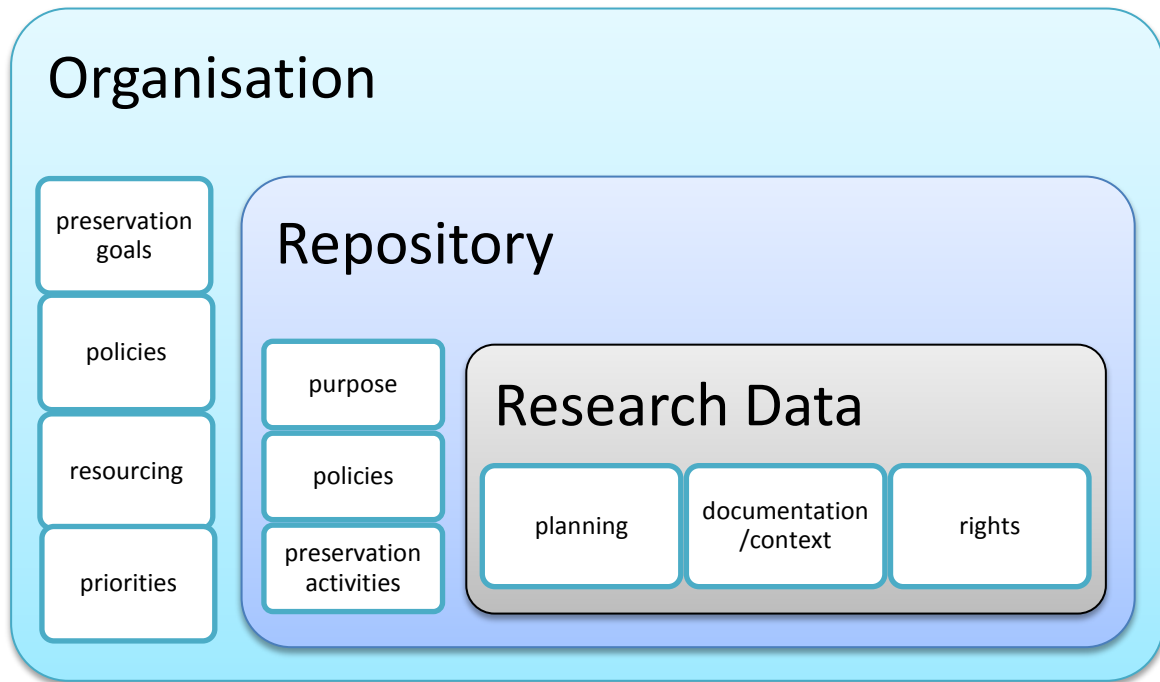


Figure 3 Diagram relating the different levels of interest

6.3 Scope

It would not be possible for a single document to cover best practice for such a variety of disciplines covered by the term “research data”; this section explains the scope of the guidelines contained within this document.

The OECD Principles and Guidelines for access to research data from public funding [12] defines **research data** as “*factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings*”. These factual records may have been generated directly for research purposes in academic projects or are being used for research purposes having been generated for other reasons.

The collection, use and preservation of research data is greatly determined by the domain of study. Primary research data in most domains is gathered as part of the process of experimentation, observation or analysis of existing sources. In some areas of research this will be done automatically using instrumentation. The primary data collected may undergo many transformations or analysis steps before it is used to reach a conclusion. There are different methods of working; different



standards and expectations. To some a document is not data, but to someone who is interested in text mining or discovering mentions of comets in twelfth century diaries then it is data. This means that one cannot use a broad brush approach to defining research data using file formats or document type as a basis.



7 Background Context

Organisations and the people who belong to them do not operate in a vacuum; external pressures are exerted from policy makers, funders, changes in researchers' expectations and technological developments. This section identifies key factors and drivers which are making an impact on the management and preservation of research data. These factors and drivers are grouped into organisation, repository management and data areas which reflect the best practice and guidance given in section 3.

7.1 Organisations

There are growing expectations from governments and funding bodies both for greater openness and preservation of data generated through public funding. This is driving the need for organisations who provide data centres/archives/repositories to consider preservation and thus to provide infrastructure to support these requirements. Depending on the type of organisation providing this infrastructure, preservation may already be a key driver but for others more focused on access to data preservation is a new requirement.

7.1.1 Models for storing and curating research data

In the academic and research environment there are four common models for the storage, and curation, of research data:

- **Research groups or project collaborations:** Those who create data are ultimately responsible for the decisions on storage and curation of that data. Depending on expectations and norms in their subject domain they may choose, or be expected, to interact with services elsewhere. Very large project collaborations, such as Large Hadron Collider at CERN, may develop and support large computing infrastructure for that project.
- **Institutions that are responsible for both data creation and curation:** These are institutions such as National Libraries, archives and Scientific Facilities who are responsible for the long-term preservation of content and also create digital content through initiatives such as digitisation of print material or the provision of large scale scientific equipment. Many of the SCAPE partners such as the British Library, the KB-NL and the Science and Technology Facilities Council fall into this category.
- **Third-party archives and institutions:** The third party archives are mainly subject-based, can range from simple community-driven databases to well-supported infrastructural services. These may be focussed on providing a central point of access rather than the long-term preservation of the content. Well-known examples of the latter include databases like Protein Data Bank or GenBank, organisations like EMBL or NCBI, the data centres funded by the UK research councils (e.g. the UK Data Archive, the UK NERC centres, UK ADS) and organisations like DANS (Netherlands). Generic, Web-based repository services are also beginning to emerge to support open data initiatives, e.g. figshare.



- **Institutions that employ researchers:** Changes with the research environment, with an increased focus of open data has encouraged research focussed institutions to provide research data management infrastructure providing services (e.g. repositories, unique identifiers), advocacy and training.

Although there are many types and scales of research data, the SCAPE approach is institutionally focussed; it assumes that the content to be preserved will be held in a repository with a cohesive management approach. It is not designed for an individual Project Investigator who needs to preserve the output of a project for the length of time dictated by their funder and so these guidelines are addressed to those who manage such repositories or intend to deposit their data into repositories for long-term access and storage.

7.2 Repository management

As the costs of physically storing digital data drop in relative terms there is the temptation to keep all the data; however not all data is worth curating and keeping. There is a need for repository to have collection management goals and tailored approaches to different parts of the collections. Repositories managed by funding organisations are providing more guidance on what might be collected, see [34], [35] & [36] for some examples.

7.2.1 Trustworthy repositories of data

One of the elements of an e-infrastructure for data is a network of repositories that can be trusted to keep their holdings safe, accessible and usable into the future. These repositories may be associated with particular subject areas or institutions, or might have a broader scope. In any case there is a need for some sort of assurance that the repository will indeed do a good job.

A fundamental standard in this area is the reference model for Open Archival Information Systems, ISO 14721 and CCSDS 650.0-M-2 [37], usually known as just OAIS. This provides a framework for the understanding and increased awareness of archival concepts needed for long term digital information preservation and access, and sets out several models for the functioning of a digital repository. It introduces the key concept of Representation Information, succinctly defined as “*The information that maps a Data Object into more meaningful concepts*”. The importance of Representation Information is in the recognition that some knowledge is required to reliably use or reuse the data, and that this cannot be taken for granted as time passes and so must be represented clearly.

Building on OAIS is another CCSDS and ISO standard, ISO 16363 “Audit and Certification of Trustworthy Digital Repositories” [38] which defines a process for assessing the trustworthiness of digital repositories with a long-term goal that a process of independent third-party certification of repositories will become possible. The standard has three headings (a) organisational infrastructure; (b) Digital Object Management and (c) Infrastructure and Security Risk management.

7.3 Data creators

Funding bodies are increasingly placing requirements on the management and curation of data generated by projects that they support. Research Councils UK (RCUK) has published a set of “Common Principles on Data Policy [19] that provide a framework for the individual Research Councils’ own policies. These principles are concerned with availability of data, and contain an explicit statement about long-term preservation: *“Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.”*

At a more practical level, there is a wealth of advice on data management planning, or of policies that place requirements on the storage of data. The Digital Curation Centre [3] is a world-leading centre of expertise in digital information curation with a focus on building capacity, capability and skills for research data management across the UK’s higher education research community provides a suite of resources to help institutions comply with UK funding organisations’ requirements.

One of the main motivations for preserving research data is to be able to reuse it in future. There are of course some difficult problems in this area, not least the balance between the sense of data as a public good and the rights of the researchers who gathered it; and how to support future reuse of data that necessarily cannot be anticipated. Adequate preservation underpins the potential for future reuse, and indeed this scenario makes strong demands on for example the supplementary information associated with the data, to enable researchers in different domains to feel comfortable in interpreting it, a similar level of context as required for effective preservation.

7.3.1 Research Data Lifecycle/process

Research data has a lifecycle in which it is created, analysed, used, preserved and reused. Most creators of data are concerned with the analysis and use of the data and the preservation is not necessarily their main concern. With the increase in emphasis on data management planning, there is more focus on preservation as part of the research data lifecycle. This means that some organisations/collaborations that have been focused on data use within their community are now more focused on what preservation might mean in their environment.

Two reports which address the importance of preservation for data as part of the research process are *Riding the Wave* of the High-Level Expert Group on Scientific data [7], published in October 2010 and the UK’s Royal Society’s “Science as an open enterprise”[8]. The first report recognises the importance of availability of increasing amounts of data and *“identifies the benefits and costs of accelerating the development of a fully functional e-infrastructure for scientific data – a system already emerging piecemeal and spontaneously across the globe, but now in need of a far-seeing, global framework. The outcome will be a vital scientific asset: flexible, reliable, efficient, cross-disciplinary and cross-border.”* It identifies challenges in being to be able to ensure that the information collected will be useable and understandable in the future and knowing what to preserve. The second also tackles the issues raised by the data deluge of modern science, and has a focus on openness in data. The report raises issues of provenance, a key factor in preservation: *“Tracking the provenance of data from its source is vital for its assessment and for attribution to its*

originators.” The general message is that an effective e-infrastructure must concern itself with long-term preservation of digital material for access and re-use, and that certain factors such as provenance are going to be essential in ensuring the success of the e-infrastructure

An example of this is the UK JISC’s funded MaRDI-Gross [33] project which considered Data Management Planning for Big Science Projects (Astronomy, Gravitational Waves and Particle Physics in particular). The starting point was an assumption that there is a need to preserve research data, and that—being a “big science” environment—there are the means to do that and it would build on the existing infrastructure which is there to support the collection and access of the data.

7.3.2 Perceived threats to data access and reusability

The threats to long-term availability of data are widely recognised. The PARSE.Insight project [10] conducted a number of surveys targeted at four key groups of stakeholders: researchers, data managers, publishers, and funders. Two of the key findings from researchers were:

- Researchers consider the possibility of re-analysis of existing data as the most important driver for the preservation of research data; 91% of the respondents thought this to be either important or very important.
- Researchers regard the lack of sustainable hardware, software or support of computer environment may make the information inaccessible as the most important threat to digital preservation. 80% believe this to be either important or very important.

The following chart from the study shows the perception of the importance of threats to preservation of digital data among the researchers surveyed. As can be seen from the table, two significant threats relate to the capture of context, dependencies and provenance of data. This is particularly relevant to research data where the context of the data is very important to the use and understanding of it.

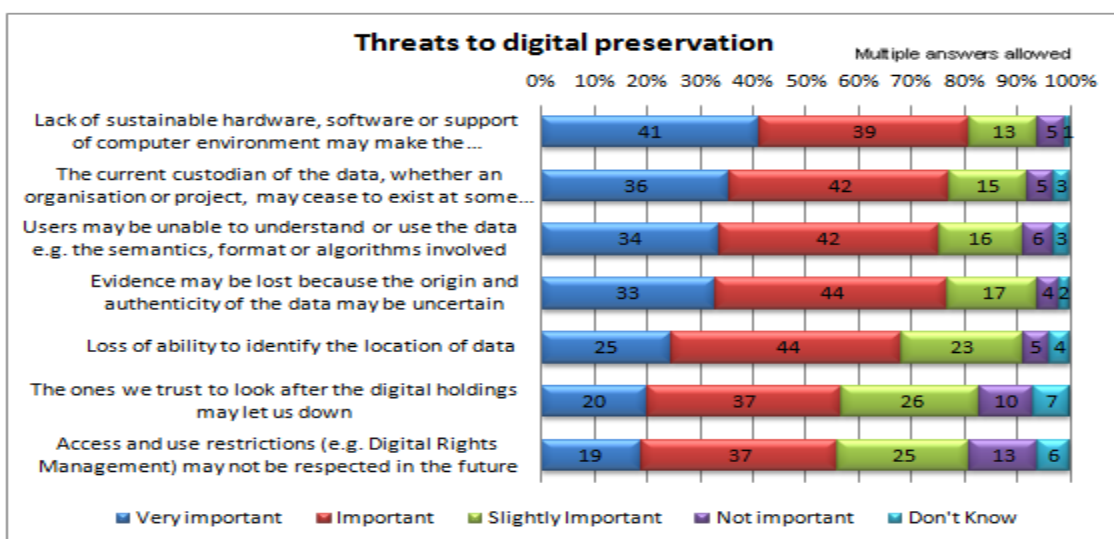


Figure 4 Threats to preservation from the PARSE.Insight project



7.3.3 Guidelines and Recommendations

The guidance outlined in this section is based on the work done in the wider community and experience gained through the SCAPE project. It is designed for those who are responsible for starting and managing repositories and preparing data for repositories, it is not the intention to produce comprehensive general best practice documentation on data preservation for research data but to build on knowledge gained through the SCAPE project.

Whilst the scale of research data and the management of it vary widely in practice, the focus of these guidelines and recommendations is on collections of datasets rather than individual ownership of data from a single project.

The approach taken in this document is one of organisational level repositories and as such there are three main sections to the advice: firstly around the organisation itself; secondly around the repository management and finally around the data itself.

Each of the items discussed is mapped to the three widely adopted standards for repository certification: Data Seal of Approval⁵, the ISO 16363 Audit and certification of trustworthy repositories⁶ (also available through CCSDS) and the German DIN 31644 through the Nestor Seal⁷. To ensure a comparable level of detail ISO16363 is mapped at the second level headings, rather than at the detailed level of individual metrics and submetrics. The following abbreviations are used DSA for Data Seal of Approval; NS for Nestor Seal and RAC for ISO 16363.

Preservation of any type of material is not a single act, the basic step is to ensure bit level preservation: that the files that are deposited are kept safely over the long term; to ensure continued use and reuse of the information additional steps described as functional preservation are required. For successful preservation, the preservation remit and policy must be in place.

⁵ <http://datasealofapproval.org/en/information/guidelines/>

⁶ Link is to the publically available text provided to ISO
<http://public.ccsds.org/publications/archive/652x0m1.pdf>

⁷ Nestor Seal of Approval using DIN 31644
http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html

7.4 Organisation

These organisational issues recommendations are intended to ensure that the right policy and resourcing framework is in place in support of preservation activities. These apply to any organisation, scalability issues comes with the management of increasing volumes of disparate data which may have conflicting requirements. Being clear about policy and resourcing will enable an organisation to make the most effective decisions for ongoing collection management.

ID	3.1.1
Activity	Set Preservation Goals
Description	<p>To be able to invest resources in an activity, an organisation needs to understand the purpose and benefit of doing so and who will be using the outputs of the preservation activity.</p> <p>Preservation goals are one way of identifying what is important to be preserved. They should define what is to be collected/preserved and what facet of the object is of importance.</p> <p>This can also be aligned to making business cases for research data management infrastructures.</p> <p>In a complex environment the needs of competing preservations goals needs to be considered and resolved.</p>
Guidance	<p>For any given collection, the organisation preserving the content should have clear and explicit preservation goals.</p> <p>These goals, or supporting documentation, should define the significant properties of the objects to be preserved so that the appropriate preservation strategies can be identified.</p>
Risks specific to research data	<p>Some large scale research data infrastructures are built to provide access and storage of the data and may not consider preservation as a specific objective. It is likely that bit level preservation activities will be addressed by the requirements for storage and data management. It is important that preservation is considered alongside current uses to ensure that access can be maintained over a long term where it is appropriate to keep the data over the long-term, considering the functional preservation aspects.</p>
Questions	<ul style="list-style-type: none"> • What is the preservation remit of the organisation? • Are the users of the preserved resources identifiable?

<p>Mapping repository certification Standards to</p>	<p>NS: C1 Selection of Information Objects and their representation NS C2: Responsibility for Preservation NS C3 Designation Communities DSA 4: The data repository has an explicit mission in the area of digital archiving and promulgates it RAC 3.1; Governance and organizational viability</p>
<p>Resources and Examples</p>	<p>SCAPE Policy Representation output D13.2 Contents of section 5.2 and [16] in particular for an example of high level principles</p> <p><i>“The Archaeology Data Service supports research, learning and teaching with freely available, high quality and dependable digital resources. It does this by preserving digital data in the long term, and by promoting and disseminating a broad range of data in archaeology.”</i> From http://archaeologydataservice.ac.uk/about (accessed Jan 2014)</p> <p><i>“DANS promotes sustained access to digital research data. For this purpose, DANS encourages researchers to archive and reuse data in a sustained manner, e.g. through the online archiving system EASY.”</i>From http://www.dans.knaw.nl/en/content/about-dans (accessed Jan 2014)</p>

ID	3.1.2
Activity	Define High Level Preservation Policy
Description	<p>To be able to effectively manage for the long-term the organisation responsible needs to be able to articulate the policy which underpins the activities. For an organisation this is likely to be at a high level.</p> <p>There may be an overlap here with the data management or collection management policies in place.</p>
Guidance	<p>Clear preservation policies should be in place in order for effective management of resources. These preservation policies should consider which preservation strategies would be most appropriate for the content being preserved. These strategies can include migration of content to another file format or emulation of the current environment.</p> <p>It should be clear what type of material is in scope and what type of material is better preserved by someone else.</p>
Risks specific to research data	<p>For those involved with creating and maintaining a research data infrastructure related to specific projects or facilities, there may not be an emphasis on long-term preservation.</p> <p>For research data which uses domain specific or local file formats, then specific notice should be taken of the methods in which this content is rendered/accessed. The preservation of the software, or even hardware, needed may need to be considered when considering functional preservation.</p>
Questions	<ul style="list-style-type: none"> • What high-level policies already exist concerning data management, storage or preservation? • What is the subject of these policies: the functioning of the data repository itself; the way that data is handled by researchers; the respective responsibilities of the parties involved; ...? • Can these policies be implemented, traced through to lower-level policies that eventually give rise to definite actions?
Mapping to repository certification Standards	<p>NS C2: Responsibility for Preservation</p> <p>DSA 6: The data repository applies documented processes and procedures for managing data storage</p> <p>RAC 3.3: Procedural accountability and preservation policy framework</p>

Resources and Examples	<p>SCAPE Policy Representation output D13.2</p> <p>Contents of section 5.2 and [16] in particular for an example of high level principles</p> <p>SCAPE has collected examples of policies: http://wiki.opf-labs.org/display/SP/Published+Preservation+Policies</p> <p><i>“The CMS collaboration is committed to preserve its data, at different levels of complexity, and to allow their re-use by a wide community”</i> The CMS data preservation, re-use and open access policy from the CMS experiment at the Large Hadron Collider https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/RetrieveFile?docid=6032&version=1&filename=CMSDataPolicy.pdf (accessed Jan 2014)</p>
------------------------	---

ID	3.1.3
Activity	Clarify Legal issues
Description	<p>There are three stakeholders in the issue of rights & legal considerations:</p> <ol style="list-style-type: none"> 1. The organisation as the provider of a service should be clearly identified from the start. 2. Those who deposit data content into the repository should be clear. 3. Those who wish to access and use the content <p>The rights, responsibilities and roles of all three should be clearly defined.</p>
Guidance	A clear rights management framework should be put in place.
Risks specific to research data	This is a new and developing area for research data. There can be some tension between the desire to make data open and available with as few barriers as possible and requirements for confidentiality of some data and monitoring the use.
Questions	<ul style="list-style-type: none"> • What are the licensing restriction/rights on the content? • How is information about rights kept and displayed to all the stakeholders? • How will secure content be kept secure? <p>These may vary down to the data level.</p>
Mapping repository certification Standards to	<p>NS C4: Access</p> <p>NS C6: Legal and contractual basis</p> <p>NS C7: Legal conformity</p> <p>NS C20: Technical authority</p>

	<p>DSA 5: The data repository uses due diligence to ensure compliance with legal regulations and contracts including when applicable, regulations governing the protection of human subjects.</p> <p>DSA 9: The data repository assumes responsibility from the data producers for access and availability of the digital objects</p> <p>DSA 14: The data consumer complies with access regulations set by the data repository</p> <p>DSA 15: The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.</p> <p>DSA 16: The data consumer respects applicable licences of the data repository regarding the use of data.</p> <p>RAC 3.5: Contracts, licenses and Liabilities</p>
Resources and Examples	<p>SCAPE Policy Representation output D13.2</p> <p>See section 5.4</p> <p>DANS license agreement for depositors: http://www.dans.knaw.nl/en/content/dans-licence-agreement-deposited-data (accessed Jan 2014)</p> <p>ADS terms and conditions for users of the service http://archaeologydataservice.ac.uk/advice/termsOfUseAndAccess (accessed Jan 2014)</p>

ID	3.1.4
Activity	Identify resources and plan for sustainability
Description	If the organisation is intending to provide long-term digital preservation, then the resources both in staff and recurrent budgets need to be estimated and provided for.
Guidance	The organisation should identify the resources for supporting this activity.
Risks specific to research data	<p>Different research domains may use specialised or proprietary formats and so it may be difficult to estimate costs of supporting these formats over time. There may be sustainability issues with specialised tools used to access or migrate the data from a specialised format.</p> <p>The scale of research data in some disciplines mean that the costs of curating the</p>

	<p>data must be balanced against the costs over the long term.</p> <p>Some data may be sensitive or confidential and the costs of keeping this over the long-term needs to be factored in.</p>
Questions	<p>If the organisation is intending to provide long-term digital preservation, then the following questions need to be addressed:</p> <ul style="list-style-type: none"> • Are there sustainability plans in place for the organisation or repository? • Who is responsible for ensuring the resourcing is appropriate? This includes staffing, storage and physical buildings. • Will the repository use a certification scheme for assurance?
Mapping repository certification Standards to	<p>NS C8: Funding</p> <p>NS C9: Personnel</p> <p>NS C10: Organisation and processes</p> <p>NS C12: Crisis/successorship management</p> <p>DSA 5: The data repository has a plan for long-term preservation of its digital assets.</p> <p>DSA 6: The data repository applies documented processes and procedures for managing the data storage.</p> <p>RAC 3.1 Governance and organisation viability</p> <p>RAC 3.2: Organisation structure and staffing</p> <p>RAC 3.3 Procedural accountability and preservation policy framework</p> <p>RAC 3.4 Financial sustainability</p> <p>RAC 5.1: Technical infrastructure risk management</p>
Resources	<p>See contents of section 5.3 and also ongoing work from the 4C project</p>

7.5 Repository Management

SCAPE is concentrating on the management of large scale preservation where the information is held in repositories, this section addresses the recommendations for those who are responsible for the management of repositories and the data held within them.

Over the long-term repositories will need to be migrated to new technology. The process and advice regarding this is covered in separate guidelines.

Content within repositories which are intended to provide long-term access and storage require that the content is actively managed through a cycle of activities designed to look for potential changes, plan as a result of changes and potentially perform preservation activities on the content. See figure below.

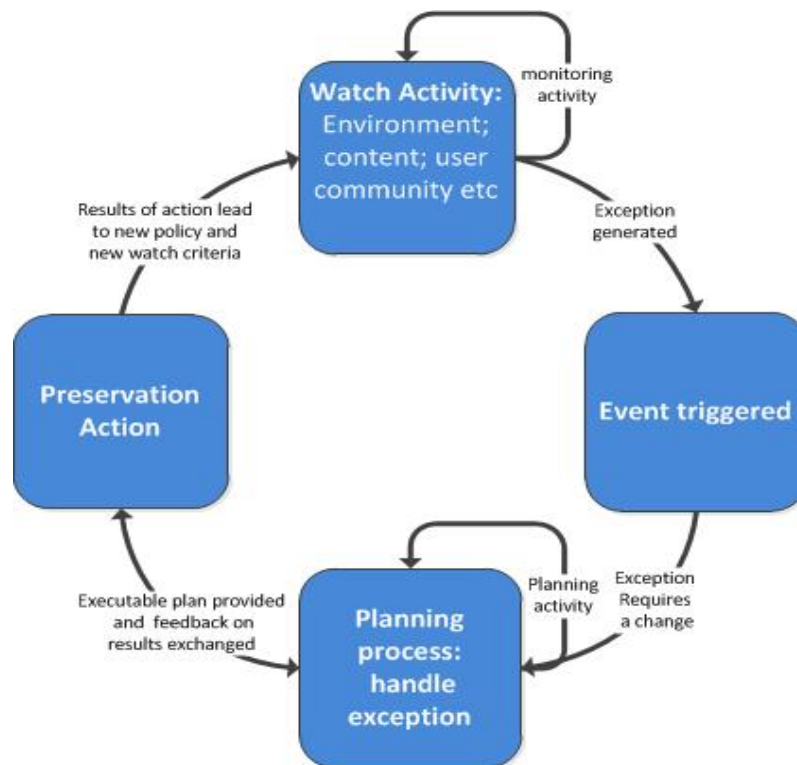


Figure 5: Watch, planning and action cycle

7.5.1 Repository & content set-up

These guidelines refer to some general pointer which would apply to all repositories, but are particular importance to those intending to provide long-term access and storage.

ID	3.2.1
Activity	Repository or Collection Level Preservation Procedure Policy
Description	<p>Although as an organisation there will be preservation policies, these will perform be at a high level; for actual management of the repository and the data held within it there needs to be lower level, more practical policy to address the procedures and resources required to run the service effectively.</p> <p>The policy should cover all stages in the lifecycle:</p> <ul style="list-style-type: none"> • Acquisition • Ingest • Description • Data management/preservation activities • Retrieval • Re-use • Disposal
Guidance	<p>Detailed policy addressing the practicalities of preservation should be in place.</p> <p>If there are significant differences in treatment of different material within the repository then it may be a good idea to have collection specific policies in place.</p> <p>Policy at the collection level may need to be machine understandable in order to use preservation specific tools to enable the automation of activities or to automatically control the access to collections.</p>
Risks specific to research data	<p>Different research domains may use specialised or proprietary file formats and so there may be additional specialised requirements.</p>
Questions	<ul style="list-style-type: none"> • Where will the content of the repository come from? • Will it be homogenous content? • Does there need to be deposit agreements in place? • Does any of the content need specific descriptive information limited to that file type or specifics of the domain? • Are there any restrictions on management or access of the material? • How long will the content be retained for? • Will the original digital object be kept even after preservation actions such as migration? • Does the content need any specialized software or hardware to be able to access or use it?

	<ul style="list-style-type: none"> • Who is the intended user group of the content? • What tools & techniques are there for machine understandable policies and what process is there for deriving them from the high level policy?
Mapping repository certification Standards	<p>to</p> <p>NS C1: Selection of the information objects and their representations</p> <p>NS C13: Significant properties</p> <p>NS C21: Submission Information packages</p> <p>NS C28: Descriptive metadata</p> <p>NS C29: Structural metadata</p> <p>NS C30: Technical metadata</p> <p>NS C32: Administrative metadata</p> <p>DSA 8: Archiving takes place according to specific work flows across the data life cycle.</p> <p>RAC 3.1 Governance and organisation viability (see collection management)</p> <p>RAC 3.3 Procedural accountability and preservation policy framework</p>
Resources and Examples	<p>SCAPE Policy Representation output D13.2</p> <p>Examples of policies collected in SCAPE wiki:</p> <p>http://wiki.opf-labs.org/display/SP/Published+Preservation+Policies</p> <p>See also section 5.2 and 5.5</p>

ID	3.2.2
Activity	Purpose and audience for the repository related to collections
Description	<p>The intended audience/user group will affect the preservation decisions as it will guide the identification of the important aspects of the content which is to be preserved.</p> <p>For example: if for the users of a specific textual item the content of the file is more important than the layout, then a different preservation strategy could be adopted.</p>
Guidance	The context and requirements for functional preservation should be clear and decisions on how much of the preservation of the context should be the responsibility of the institution preserving the dataset itself.
Risks specific to research data	Some research data is generated using very specialized formats and it is important to ensure that additional information required to be able to access and use the

	<p>content is collected with the object. What level of detail that information is will depend on the level of the intended audience – for example experts in the field may need different information to undergraduate students.</p> <p>Depending on the context of the repository – whether it is very specific to a domain or whether it is capturing a variety of domains – there may be a very varied user community which may bring additional complexity.</p>
Questions	<ul style="list-style-type: none"> • Who is the intended audience/user community? • Is there a process of “data release” in research data - what special demands does this make? For example does a new release supercede a previous version? Is there enough resource to keep all versions of a data release? • What additional contextual information is required for effective preservation and reuse of research data? Who is responsible for collecting and preserving this additional context? Are they a trusted source? • What additional constraints/requirements are imposed by the preservation of research data? There is often a dependency on the analysis software – what is the preservation position of the repository on this? • How is the content selected?
Mapping repository certification Standards to	<p>NS C3: Designated communities</p> <p>NS C5: Interpretability</p> <p>DSA 4: The data repository has an explicit mission in the area of digital archiving and promulgates it.</p> <p>RAC 3.3: Procedural accountability and preservation policy framework</p> <p>RAC 4.5 Information Management</p> <p>RAC 4.6 Access Management</p>
Resources and Examples	<p>See SCAPE Policy Representation output (due M36)</p> <p>See section 5.5</p> <p>UK Data Service defines their audience as “<i>researchers, teachers and policymakers who depend on high-quality social and economic data.</i>” From http://data-archive.ac.uk/about/services/uk-data-service (accessed Jan 2014)</p> <p>UK Archaeology data Service “<i>supports research, learning and teaching</i>” From http://archaeologydataservice.ac.uk/about (accessed Jan 2014)</p> <p>The Austrian National Library (ONB) “<i>The Austrian National Library regards itself as a centre of information and research oriented toward serving the public, as an outstanding national memory institution and as a many-sided centre of education and culture.</i>” http://www.onb.ac.at/ev/about/mission.htm (accessed Jan 2014)</p>

ID	3.2.3
Activity	Good Management Practices and Trusted Digital Repository standards
Description	<p>It is important that repositories holding content for the long term are successfully managed and perform good data management practices.</p> <p>This applies to all stages of the data life-cycle within the repository, and involves implementing policy as described in 3.2.1. Life cycle stages:</p> <ul style="list-style-type: none"> • Acquisition • Ingest • Description • Data management/preservation activities • Access/Retrieval • Re-use • Disposal <p>There are currently two main standards for digital repositories which enable those who provide repositories to validate their practices against these standards to ensure that the repositories are run to high and sustainable standard. These are DataSeal of Approval and ISO 16363</p>
Guidance	<p>Achieving a formal certification is a rigorous process and may use much resource within the repository staffing and must therefore be part of the organisations strategy and policy. . However it will demonstrate that the repository is well managed and will be sustainable over the long-term. For some preservation specific repositories the act of certification provides additional reassurance to their community of depositors and users that it is sustainable over the long term.</p>
Risks specific to research data	<p>If the repository covers many different domains and/or types of data then the data management procedures and practices will need to become more complex to address specific issues.</p>
Questions	<ul style="list-style-type: none"> • Who is responsible for ensuring the repository is fit for purpose? • What processes & procedures are recommended for good data management? • What processes are in place for ensuring bit preservation? • What risks linked to the preservation of content can be identified? • What standards will be used for content description and preservation metadata? • Who is responsible for any retention decisions? • Does the repository intend to become certified?
Mapping repository certification to	<p>NS C4: Access</p> <p>NS C14: Integrity: ingest interface</p>

Standards	<p>NS C15: Integrity: Functions of the archival storage</p> <p>NS C16: Integrity: user interface</p> <p>NS C17: Authenticity: ingest</p> <p>NS C19: Authenticity: Use</p> <p>NS C22: Transformation of the submission information packages into archival information packages</p> <p>NS C23 Archival information packages</p> <p>NS C25:Transformation of archival information packages into dissemination information packages</p> <p>NS C26: Dissemination information packages</p> <p>NS C33: IT infrastructure</p> <p>NS C34: Security</p> <p>DSA 6: The data repository applies documented processes and procedures for managing data storage</p> <p>DSA 11; The data repository ensures the integrity of the digital objects and the metadata</p> <p>DSA 12: The data repository ensures the authenticity of the digital objects and the metadata</p> <p>DSA 13: The technical infrastructure explicitly supports the tasks and functions described in international accepted archival standards like OAIS</p> <p>RAC 3.2: Organizational Structure and Staffing</p> <p>RAC 4.1: Ingest: acquisition of content</p> <p>RAC 4.2: Ingest: creation of the AIP</p> <p>RAC 4.4 AIP Preservation</p> <p>RAC 4.5 Information management</p> <p>RAC 4.6: Access management</p> <p>RAC 5.1 Technical infrastructure risk management</p> <p>RAC 5.3 Security</p>
Resources	<p>See SCAPE Policy Representation output D13.2</p> <p>See section 5.5</p>

ID	3.2.4
Activity	Unique Identification of data sets
Description	Within a repository data sets need to be uniquely identified, in the past this might

	<p>have been as straightforward as some type of accession number. In the changing world where the citing of data is becoming more accepted then the identification of data should be persistent over the longer-term.</p> <p>An important issue regarding persistent identifiers is the level at which this is assigned. It can be assigned to the entire dataset but this can present problems if the dataset is not completed and is growing; it can be at an intermediate logical level, or even at the file level. Each of these levels has trade-offs between management and the precise definition of what is being identified.</p> <p>There is also the issue of multiple identifiers, for a dataset where the creator has assigned a persistent identifier, such as a DOI or a PURL handle, there is a need for the preservation infrastructure to preserve that, even though the preservation copy may not be the copy of record and so will need a local unique identifier. The original persistent identifier will resolve to the original copy of the data, not necessarily the copy in the preservation infrastructure.</p>
Guidance	<p>Providing persistent identifiers for the dataset is good practice.</p> <p>The level at which these are assigned will depend on the domain and practice within the domain.</p>
Risks specific to research data	<p>Issues such as datasets continuing to grow and how that is resolved for persistent identification is still a developing field.</p>
Questions	<ul style="list-style-type: none"> • What type of persistent identifier scheme will be used? • What approach will be taken if the data set is ingested with an existing persistent identifier attached to it • Are there any versioning issues associated with datasets being preserved in this repository? • What happens to the persistent identifier if the content is migrated to another format?
Mapping repository certification Standards to	<p>NS C4: Access</p> <p>NS C27: Identification</p> <p>DSA 10: The data repository enables users to discover and use the data and refer to them in a persistent way.</p> <p>RAC 4.2 Ingest: creation of the AIP</p>
Resources	<p>See section 5.6</p>

7.5.2 Preservation activities

This next section addresses activities which are designed to ensure effective preservation of the content of the repository.

ID	3.2.5
Activity	Preservation Watch activities
Description	Preservation Watch is the process of routinely looking for changes in the environment, be it policy or, technical which will impact on the way that the repository and its content are managed. By looking for changes the repository can proactively react to changes and plan to minimise the impact.
Guidance	<p>It is important to ensure that changes in the local and wider environments are monitored to be able to adopt a pro-active approach.</p> <p>The changes being monitored are likely to be identified as a result of risk management activities and policy decisions and will depend in part on the collection and remit of the repository and organisation.</p> <p>For large scale collections then automated watch activities may be appropriate, as it may be too complex a landscape, or too time consuming to perform.</p>
Risks specific to research data	Changes to specialised file formats may be more difficult to watch for than file format standards which are widely adopted across multiple subjects.
Questions	<ul style="list-style-type: none"> • What are the important risks which need to be monitored? • What machine understandable policy is needed to be able to automate a watch function? • Where will the information about changes be found? • Who is responsible for the monitoring process? • What happens if some change is identified through the monitoring process? • How can watch be automated effectively? • What kinds of obsolescence might affect the long-term preservation of the data? • What other possible changes might also be relevant, for example in the knowledge of the community for whom the data is being preserved?
Mapping repository certification Standards to	<p>NS C11: Preservation measures</p> <p>DSA 7: The data repository has a plan for long-term preservation of its digital assets</p> <p>RAC 4.3 Preservation Planning</p> <p>RAC 5.1: technical infrastructure risk management</p>
Resources	SCAPE SCOUT watch tool has been developed to enable automated watch activities.

	See section 5.7
--	-----------------

ID	3.2.6
Activity	Preservation Planning
Description	<p>Preservation planning is an activity which should be undertaken when change is detected, either in the technical or policy landscape or as a result of new collections. It is designed to ensure that the appropriate activities to maximise the preservability of the digital objects is undertaken.</p> <p>Taking a preservation action without planning and considering alternatives may result in wasted resources or a poor choice of action.</p>
Guidance	<p>The ingest of new content, or changes to the environment should always be analysed and the most appropriate action within available resources identified.</p> <p>As part of the preservation planning process, risks should identified and mitigated.</p> <p>As collections become more complex and larger, then a repository may benefit from consistent and automated planning tools.</p>
Risks specific to research data	<p>If the collection is complex and heterogeneous in nature then planning at the appropriate level is more time –consuming.</p> <p>For data in proprietary formats, there may not be many suitable alternatives.</p>
Questions	<ul style="list-style-type: none"> • How can planning be automated effectively? • What kinds of obsolescence might affect the long-term preservation of the data? • Are there specific scale aspects to the potential preservation actions to be identified? • Are there any specific restrictions on actions that can be undertaken? • Have the decision criteria which will distinguish between different options been identified?
Mapping repository certification Standards to	<p>NS C11: Preservation measures</p> <p>DSA 7: The data repository has a plan for long term preservation of its digital assets.</p> <p>RAC 4.3: Preservation Planning</p>
Resources	<p>SCAPE PLATO tool</p> <p>See section 5.7 in particular</p>

ID	3.2.7
Activity	Preservation Actions
Description	<p>Following planning, an appropriate preservation action will be identified. The precise action will depend on the file format and the risk identified. They may include:</p> <ul style="list-style-type: none"> • Transformation of the object • Replacement of the repository • Quality assurance of the content • Additional enhancement of the metadata for the object
Guidance	<p>Preservation actions should always be documented, both in a human readable form and also in the preservation metadata for the item (s).</p> <p>PREMIS⁸ is the standard for preservation metadata.</p> <p>All preservation actions should be tested on a small scale before implementing over the entire dataset.</p> <p>All preservation actions should include a component for testing the successful outcome to provide quality assurance.</p>
Risks specific to research data	<p>For some preservation actions there may be specialist, community tools as the file formats may be specialised to a small focused domain.</p> <p>Large scale research data may need the use of novel architectures to ensure that preservation actions on the whole collection are able to be achieved within a reasonable time frame.</p>
Questions	<ul style="list-style-type: none"> • What types of actions might be needed to ensure that data continues to be accessible, usable and understandable in future? • Is transformation of file formats envisaged in future? What might be lost through such transformations? • Would it be possible to add supplementary information to data as the world changes? • Are all file formats known and familiar? • Is there a dependence on software to analyse or reuse the data? What if the software is no longer available? • Is there a need to record provenance of the data—a record of the processing that has been done to the object?

⁸ <http://www.loc.gov/standards/premis/>

<p>Mapping repository certification Standards to</p>	<p>NS C15: Integrity: functions of the archival storage</p> <p>NS C18: Authenticity: Preservation measures</p> <p>NS C24: interpretability of the archival information packages</p> <p>NS C31: Logging the preservation measures</p> <p>DSA 7: The data repository has a plan for long-term preservation of its digital assets</p> <p>DSA 8: Archiving takes place according to the explicit workflows across the data life cycle.</p> <p>RAC 4.3 Preservation Planning</p> <p>RAC 4.4: AIP Preservation</p> <p>RAC 4.5: Information Management</p>
<p>Resources</p>	<p>SCAPE tools, and workflows, have been developed for file characterisation, some transformations and quality assurance for a selection of file formats. See the SCAPE tool catalogue and My experiment for Taverna workflows.</p> <p>See section 5.7</p>

7.6 Data concerns

This advice is aimed at those who produce data and then provide that data to a third party (which may be part of the same organisation) to manage, preserve and provide access to. This puts additional responsibilities upon the data creator to ensure that the data is well documented and is of an appropriate quality to be preserved.

<p>ID</p>	<p>3.3.1</p>
<p>Activity</p>	<p>Data Management Planning</p>
<p>Description</p>	<p>There is a trend toward funders and research institutions requiring those who create data through project funding to provide a data management plan which details what data is to be produced, in which formats, how it is to be managed over the project and whether there are long-term requirements.</p>
<p>Guidance</p>	<p>Effective planning for data management and preservation should be done at the start of the project whenever possible.</p> <p>Domain standards for file formats, experimental methods and analysis should be adopted where-ever possible and exceptions should be documented.</p>
<p>Risks specific to</p>	<p>Data management planning is designed to be used for research data to assist the</p>

research data	activities of good data management and preservation.
Questions	<ul style="list-style-type: none"> • Does the funder of the research expect a formal data management plan? • Has the type of data to be collected/created been identified? • What requirements are there for long term storage and preservation of content? • Are there any restrictions or constraints on the data which will have an impact on the long term preservation.
Mapping to repository certification Standards	Not directly applicable, although good data management planning will enable the data to be prepared for deposition in a repository as part of the management process
Resources and Examples	<p>See section 5.4</p> <p>Advice from the UK Digital Curation Centre: http://www.dcc.ac.uk/resources/data-management-plans (accessed Jan 2014)</p> <p>Advice from the UK Centre for Environmental Data Archival http://www.ceda.ac.uk/help/archiving-with-ceda/outline-data-management-plans/ (accessed Jan 2014)</p> <p>Advice from the UK's Medical Research Council http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/datasharing/DMPs/index.htm (accessed Jan 2014)</p> <p>Advice from MIT http://libraries.mit.edu/guides/subjects/data-management/checklist.html (accessed Jan 2014)</p>

While the contents of every data management plan will be different, the Data Curation Centre has a checklist and template which is a good starting point. The table below summarises this. For full details see DCC. (2013). *Checklist for a Data Management Plan*. v.4.0. Edinburgh: Digital Curation

The Data Curation Centre's Checklist
Administrative information:
Details about the project, purpose, personnel, funding body, links to applicable institutional policies
Data Collection
Information about the data being collected or created such as type, format and amount. This should include information on your intended use of standards & formats
Documentation and Metadata
Information on what additional materials will be provided to enable the data to be understood and

<p>reused. This should include use of standards and some details of how this material is going to be produced.</p>
<p>Ethics and Legal Compliance</p>
<p>Are there any ethical issues to be considered and if so what procedures and protocols will be put in place? What is the IPR and copyright position? If there are collaborators from different organisations how will this be agreed and organised?</p>
<p>Storage and Backup</p>
<p>What are the arrangements for data storage and IT back-up whilst the project is underway?</p>
<p>Selection and Preservation</p>
<p>It is important to be clear about what data from the project will be selected for long-term preservation and sharing with others. This should be a conscious decision. This should include where the long-term home for the data is</p>
<p>Data Sharing</p>
<p>What arrangements will be put in place for sharing your data. Are there any ethical, legal or commercial reasons which may make data sharing arrangements more complex?</p> <p>Many funders' expectations are that data should be shared, when it has been publically funded.</p>
<p>Responsibilities and Resources</p>
<p>Who is involved in the data management process and are all the roles clearly defined? This is of especial importance when more than one institution is involved.</p>

ID	3.3.2
Activity	Produce Data documentation
Description	For successful use and reuse of research data, then it needs to have proper documentation. This may as straightforward as ensuring the columns in a spreadsheet are unambiguously labelled and have the units of measurements through to documentation of the experimental intent.
Guidance	<p>There needs to be sufficient information about the dataset so that it can be preserved and reused in the future.</p> <p>Consideration should be made for the preservation of any supporting documentation/information required to make the data understandable.</p> <p>Consideration of grouping like data into a collection and documenting the collection should be made.</p>
Risks specific to research data	Any information needed for re-use that is only available at the creation point may need to be collected then. This may pose difficulties for large volume data or that which is automatically created.
Questions	<ul style="list-style-type: none"> • What is important about the data/the way it was collected/the collection purpose which needs to be explicitly documented? • What tools/software packages are required to be able to view and use the data? • How will the documentation be accessed by those who wish to see/use the data • What standards will be used for content description and preservation metadata? • Is there a formal process of “data release” in research – what special demands does this make? • What additional contextual information is required for effective preservation and reuse of research data? Who is responsible for collecting and preserving this additional context? Are they a trusted source? • For large volume data, what methods can be adopted to minimise the requirements of creating/associating the additional information to ensure that the documentation process does not become an insurmountable task.?

<p>Mapping repository certification Standards to</p>	<p>NS C5: Interpretability</p> <p>DSA1: The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with the disciplinary norms.</p> <p>DSA 2: The data producer provides the data in formats recommended by the data repository</p> <p>DSA 3: The data producer provides the data together with the metadata requested by the data repository.</p> <p>RAC 4.1: Ingest: Acquisition of Content</p>
<p>Resources and Examples</p>	<p>See section 5.4</p> <p>UK Data Service (social science data) : http://data-archive.ac.uk/create-manage/document (accessed Jan 2014)</p> <p>Australian National Data Service Guide to Metadata http://ands.org.au/guides/metadata-working.html (accessed Jan 2014)</p> <p>University of Minnesota: https://www.lib.umn.edu/datamanagement/metadata (accessed Jan 2014)</p> <p>Advice on describing images from JISC Digital Media http://www.jiscdigitalmedia.ac.uk/guide/approaches-to-describing-images/ (accessed Jan 2014)</p>

ID	3.3.3
Activity	Consider use of standard controlled vocabularies and ontologies
Description	<p>Part of the description of research data can include the use of controlled vocabularies or ontologies to enhance the information to enable better description or location.</p> <p>A common use of controlled vocabularies is to add additional subject descriptors to enable the item to be placed in the wider or narrower context. The level of detail used will reflect the expertise of the intended audience. However controlled vocabularies can be used for other purposes such as describing relationships between objects or associating additional characteristics.</p>
Guidance	<p>Using a standard vocabulary to add additional information to the research data can ensure a consistent approach to description and can aid information location.</p> <p>Using ontologies which link additional information to objects can be used to automatically add further details.</p> <p>For example an ontology which links experimental techniques and instrument could be used to add information on experimental techniques to datasets generated from specific instruments.</p>
Risks specific to research data	The ontologies and controlled vocabularies may be limited to a small specific domain and there may be sustainability risks associated with it.
Questions	<ul style="list-style-type: none"> • What is important about the data/the way it was collected/the collection purpose which needs to be explicitly documented? • What tools/software packages are required to be able to view and use the data? • How will the documentation be accessed by those who wish to see/use the data • What standards will be used for content description and preservation metadata? • Is there a formal process of “data release” in research – what special demands does this make? • What additional contextual information is required for effective preservation and reuse of research data? Who is responsible for collecting and preserving this additional context? Are they a trusted source? • For large volume data, what methods can be adopted to minimise the requirements of creating/associating the additional information to ensure that the documentation process does not become an insurmountable task.?
Mapping repository to	DSA1: The data produce deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with the

certification Standards	disciplinary norms. DSA 3: The data producer provides the data together with the metadata requested by the data repository.
Resources	See section 5.4

ID	3.3.4
Activity	Clarify any rights/consent issues
Description	The data should have clear right/licensing information associated with it. If the data is about human subjects, then information about what purposes the data could be used for that the people consented to is of particular concern.
Guidance	There should be clear rights information, both for preservation purposes and for accessing the content.
Risks specific to research data	Certain types of research may have specific rights or consent issues and need to be discussed and agreed before data is collected. In particular research using human subjects needs to take ethical considerations into account. Some research done with commercial partners may have additional commercial in confidence issues.
Questions	<ul style="list-style-type: none"> • Have preservation processes been considered when discussing/agreeing any restrictions due to ethical or commercial constraints? • How will be the rights information be associated with the data sets concerned? • If the data set is gathered over a long time period (decades) will the same rights apply to all data collection events? • Is any national legislation likely to impact how this data might be preserved over the long term?
Mapping repository certification Standards	<p>to</p> <p>DSA1: The data produce deposits the data in a data repository with sufficient information for others to assess the quality of the data and compliance with the disciplinary norms.</p> <p>NS C6: Legal and contractual basis</p> <p>DSA 15: The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information</p> <p>DSA 16: The data consumer respects the applicable licences of the data repository regarding the use of the data.</p>



	RAC 3.5: Contracts, Licenses and Liabilities
Resources	Resources from the bibliography , see [30]



8 Conclusions

The guidance outlined in section 3 is based on the work done in the wider community and findings from the SCAPE project.

Managing and curating research data is an active topic and will continue to develop, but it is important to understand the special characteristics of this type of material to ensure that appropriate preservation decisions and actions are undertaken.

9 Bibliography and useful websites

9.1 Useful bodies and conferences

There are a variety of useful bodies and communities in this area who provide advice and guidance

- [1] Keys conferences in the Digital Curation area are iPres and IDCC which are annual events.
- [2] Open Planets Foundation
- [3] Alliance for the Permanent Access to the Records of Science
- [4] National Bodies such as
 - o UK Digital Curation Centre
 - o Digital Preservation Coalition in the UK
 - o NESTOR in Germany
 - o DANS in The Netherlands
 - o Australian National Data Service
 - o US National Digital Stewardship Alliance US bodies

9.2 Policy and strategic: High level view

Context setting reports and articles

- [5] Riding the Wave, High Level Expert Group, 2010
<http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>
Vision of the High Level Expert Group on Scientific Data, includes information on long-term access and usability of data.
- [6] Science as an open enterprise (2012), Royal Society
<http://royalsociety.org/policy/projects/science-public-enterprise/report/>
- [7] ODE
<http://www.alliancepermanentaccess.org/index.php/community/current-projects/ode/>
A project which examined the perceptions and needs of stakeholders in data sharing
- [8] Parse.Insight
<http://www.parse-insight.eu/>
Perceptions and needs of stakeholders in data sharing
- [9] A surfboard for Riding the wave: 4 countries view of implementing the Riding the Wave report
Available from the Knowledge Exchange website <http://www.knowledge-exchange.info>
- [10] OECD (2007). OECD Principles and Guidelines for Access to Research Data from Public Funding. Paris: OECD.
<http://www.oecd.org/dataoecd/9/61/38500813.pdf>
- [11] RIN (2008) Stewardship of digital research data: a framework of principles and guidelines
London: RIN.
<http://rinarchive.jisc-collections.ac.uk/our-work/data-management-and-curation/stewardship-digital-research-data-principles-and-guidelines>
- [12] Addressing Digital Presentation - Proposals for New Perspective
<http://cs.harding.edu/indp/papers/barateiro7.pdf>



Paper oriented towards risk management approach to digital preservation, high level view for general public.

- [13] Preservation Modelling Goals to Guide Digital Preservation, Angela Dappert, Adam Farquhar, The British Library, Boston Spa, Wetherby, West Yorkshire, UK, In: The international journal of digital curation, Issue 2, volume 4, 2009 p 119 doi:10.2218/ijdc.v4i2.102
- [14] KEY PERSPECTIVES. (2010), "Data dimensions: disciplinary differences in research data sharing, reuse and long term viability: A comparative review based on sixteen case studies". DCC SCARP Synthesis Report commissioned by the Digital Curation Centre.
<http://www.dcc.ac.uk/sites/default/files/documents/publications/SCARP-Synthesis.pdf>
- [15] Neil Beagrie, Robert Beagrie, Ian Rowlands. "Research Data Preservation and Access: The Views of Researchers". July 2009, *Ariadne* Issue 60 <http://www.ariadne.ac.uk/issue60/beagrie-et-al/>
- [16] RCUK Common Principles on data policy
<http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>
- [17] Open Data Dialogue
<http://www.rcuk.ac.uk/documents/documents/TNSBMRBRCUKOpendatareport.pdf>
- [18] SCAPE – Published Preservation Policies wiki
<http://wiki.opf-labs.org/display/SP/Published+Preservation+Policies>

A wiki collecting published preservation policies from a variety of organisations.

9.3 Costs and benefits

Information on how to assess costs and benefits of digital preservation.

- [19] Beagrie N, KEEPING RESEARCH DATA SAFE - A COST MODEL AND GUIDANCE FOR UK UNIVERSITIES,
<http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf>
- [20] Blue Riband task force final Report, 2010
http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf
- [21] Estimating Digitization Costs in Digital Libraries Using DiCoMo
<http://dl.acm.org/citation.cfm?id=1887780>

Paper focused on time and cost estimates for digital preservation in general

- [22] FRY, J., LOCKYER, S., OPPENHEIM, C., HOUGHTON, J., & RASMUSSEN, B. (2008). Identifying benefits arising from the curation and open sharing of research data produced by UK Higher Education and research institutes". Loughborough University, Centre for Strategic Economic Studies <http://hdl.handle.net/2134/4600>
- [23] 4C Project : Collaboration to Clarify the Costs of Curation <http://4cproject.eu/>

This is an EU project which is examining issues concerning the costs of curation.

9.4 Practical Guidance

Links to specific guidance of a more practical nature.

- [24] Study IISH Guidelines for preserving research data
<http://www.surffoundation.nl/en/publicaties/Pages/StudyIISHGuidelinesforpreservingresearchdata.aspx>

[25]DCC Digital Curation Reference Manual

<http://www.dcc.ac.uk/resources/curation-reference-manual>

[26]DCC Curation Lifecycle Model

<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

[27]DCC How to guide: Appraise & Select Research Data for Curation

<http://www.dcc.ac.uk/resources/how-guides/appraise-select-data>

[28]DCC How to guide: Develop a Data Management and Sharing Plan

<http://www.dcc.ac.uk/resources/how-guides/develop-data-plan>

[29]Archaeology Data Service Guides to Good Practice

<http://archaeologydataservice.ac.uk/advice/preservation>

[30]Ball, A. (2012). 'How to License Research Data'. *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available online: <http://www.dcc.ac.uk/resources/how-guides>

[31]Bicarregui J, Gray N, Henderson R, Jones R, Lambert S, Matthews B (2012) 'DMP Planning for Big Science Projects: MaRDI-Gross project' <http://purl.org/nxg/projects/mardi-gross/report>

[32]Managing and Sharing Data: A Guide for Researchers, 3rd ed, 2011, UK Data Archive

<http://www.data-archive.ac.uk/media/2894/managingsharing.pdf>

[33]NERC Data checklist <http://www.nerc.ac.uk/research/sites/data/documents/data-value-checklist.pdf>

[34]DANS checklist for selecting data for preservation

http://www.dans.knaw.nl/sites/default/files/file/archief/Factsheet_Checklist_storing_and_selecting_DEF.pdf

[35]DANS Data Managing Planning

[http://www.dans.knaw.nl/sites/default/files/file/Datamanagementplan%20UK\(1\).pdf](http://www.dans.knaw.nl/sites/default/files/file/Datamanagementplan%20UK(1).pdf)

9.5 Repository Management & Infrastructure

Advice relating to the management of the repository and the underpinning infrastructure rather than the data within it.

[36]Digital Preservation, Architecture and Technology for Trusted Digital Repositories

<http://www.dlib.org/dlib/june05/jantz/06jantz.html>

[37]Open Archival Information Systems ISO 14721 and CCSDS 650.0-M-2

<http://public.ccsds.org/publications/AllPubs.aspx>

[38]ISO 16363 "Audit and Certification of Trustworthy Digital Repositories

<http://public.ccsds.org/publications/AllPubs.aspx>

[39]The Significance of Storage in the 'Cost of Risk' of Digital Preservation

<http://www.bl.uk/ipres2008/programme.htm>

[40]Dataseal organisation & DCC case study

<http://www.datasealofapproval.org/>

<http://www.dcc.ac.uk/resources/case-studies/ads-dsa>

[41]SESINK, L., VAN HORIK, R., & HARMSEN, H. (2008). *Data Seal of Approval: Quality Guidelines for Digital Research Data in the Netherland*. The Hague: Data Archiving and Networked Services. The Hague, Data Archiving and Networked Services - 2nd ed. DANS, 2010. ISBN 978 9490 531 027.

[42]Trusted audit and certification work



<http://wiki.digitalrepositoryauditandcertification.org/bin/view>

- [43] GREEN, A., MACDONALD, S., & RICE, R. (2009). Policy-making for Research Data in Repositories: A Guide. London: JISC funded DISC-UK Share Project.

<http://www.disc-uk.org/docs/guide.pdf>

- [44] TRELOAR, A., GROENEWEGEN, D., & HARBOE-REE, C. (2007). The Data Curation Continuum: Managing Data Objects in Institutional Repositories. D-Lib Magazine [online], 13 (9/10)

<http://www.dlib.org/dlib/september07/treloar/09treloar.html>

9.6 Data citation

- [45] DCC How to guide: Cite Datasets and Link to Publications

<http://www.dcc.ac.uk/resources/how-guides/cite-datasets>

9.7 SCAPE related papers

For a complete list see the SCAPE web site, this subset relates to issues discussed in this document.

- [46] ROMAN G, HUBER-MORK R, SCHINDLER A, SCHLARB S: Duplicate Detection Approaches for Quality Assurance of Document Image Collections. The International ACM Conference on Management of Emergent Digital EcoSystems (MEDES 2013)
- [47] BECKER C, FARIA L, DURETEC K: Scalable Preservation Intelligence for Information Longevity, OCLC Systems & Services, 2013
- [48] MATTHEWS BM, JONES C, BUNAKOV V, CROMPTON S: Investigations as research objects within facilities science, Linking and Contextualizing Publications and Datasets Workshop at TPDL, 2013
- [49] FARIA L, DURETEC K, KULMUKHAMETOV A, RAUBER A: Tools for uncovering preservation risks in your large repositories, iPres2013 2013
- [50] SCHLARB S: An Open Source Infrastructure for Preserving Large collections of Digital Objects, ELAG2013
- [51] SIERMAN B, JONES C, BECHHOFFER S, ELSTROM G: Preservation Policy Levels in SCAPE, iPres, 2013
- [52] NEUDECKER C, SCHLARB S: The Elephant in the Library, Hadoop Summit Europe, 2013
- [53] SCHLARB S: An open source infrastructure for quality assurance and preservation of a large digital book collection (Proceedings), IS&T Archiving 2013
- [54] SCHMIDT, R: An Architectural Overview of the SCAPE Preservation Platform, iPres 2012 (short paper), 2012
- [55] ASSEG F, RAZUM M, HAHN, M: Apache Hadoop as a Storage Backend for Fedora Commons, Open Repositories 2012
- [56] BECKER C: A Capability Model for Digital Preservation: Analysing Concerns, Drivers, Constraints, Capabilities and Maturities, iPRES 2011 - Proceedings of the 8th International Conference on Preservation of Digital Objects, 2011
- [57] CONWAY E, LAMBERT S: Managing Preservation Networks: Issues of Scale for Scientific Research Assets, iPRES 2011 - Proceedings of the 8th International Conference on Preservation of Digital Objects, 2011
- [58] BECKER C: Control Objectives for DP: Digital Preservation as an Integrated Part of IT Governance, ASIST 2011 - Proceedings 74th Annual Meeting of the American Society for Information Science and Technology, 2011
- [59] SCHMIDT R, An Approach for Processing Large and Non-Uniform Media Objects on MapReduce-based Clusters, Lecture Notes in Computer Science, 2011
- [60] KING R, Evolving Domains, Problems and Solutions for Long Term Digital Preservation iPres 2011 - Proceedings of the 8th International Conference on Preservation of Digital Object, 2011



- [61]BECKER C, RAUBER A: Decision criteria in digital preservation: What to measure and how. Journal of the American Society for Information Science and Technology, 2011
- [62]BECKER C, RAUBER A: Preservation Decisions: Terms and Conditions apply, Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries, 2011



Bit Preservation

Best practice guidelines and
recommendations



Table of Contents

1	Introduction	1
1.1	Scope	3
1.2	Audience	3
2	Policy, Management and Risks	4
2.1	Policy	4
2.2	Management: IT Governance frameworks	5
2.3	Risk assessment and management	5
2.4	Physical Environment Risks	7
2.5	Hardware and media risks	8
2.6	Data Security risks	11
3	Technical Approaches	12
3.1	Commercial Preservation Systems	12
3.2	Bit preservation and the cloud	13
4	Case Studies	17
4.1	Bit preservation at the British Library	17
4.2	STFC Large Hadron Collider Tier 1 bit storage	19
5	Conclusions and recommendations	23
5.1	Management of the bit level preservation infrastructure	23
5.2	Technical and Operational concerns	28
6	Bibliography	34
6.1	General	34
6.2	Policy, Management and Risks	34
6.3	Technical Approaches:	36
6.4	Case Studies	37

10 Introduction

Bit preservation is a necessary part of digital preservation activities; it is essential but not sufficient for successful preservation. In this document the following working definition is used:

Bit preservation activities are concerned with ensuring the *persistence* of the file or digital object over time, while digital preservation activities, also known as functional preservation, are concerned with the *accessibility and usability* of the file or digital object. That is to say, bit preservation is a precondition of digital preservation.

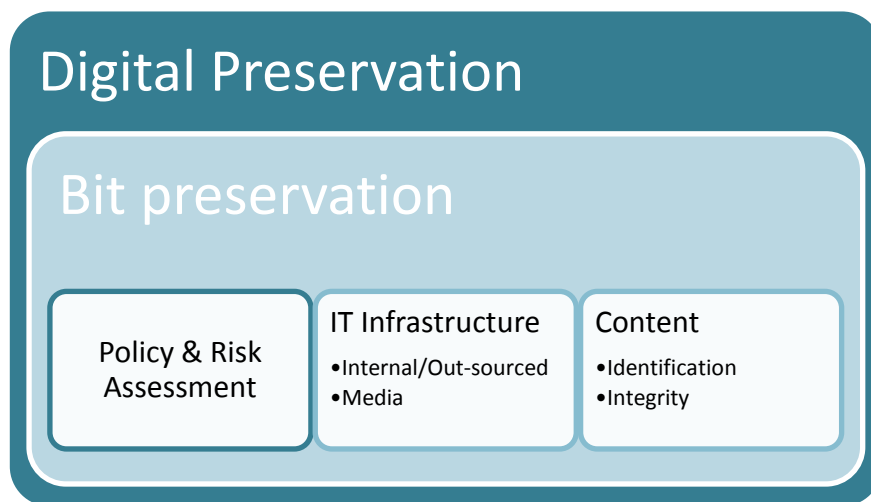


Figure 6: Diagram showing relationship between bit and digital preservation

The persistence activities which should be undertaken by a content holding organisation can be split into three main areas: policy and risk assessment; IT infrastructure management and content related activities.

Policy and risk assessment form the bedrock of an organisations approach to bit preservation. Policy should define what the preservation remit of the organisation is, what is being preserved and who for. It should consider resourcing and sustainability issues. Bit preservation is, as most activities are, a balancing act between the best possible outcome and the resources available to support that. Risk assessments enable judgements to be made about the risks to content and what can be done to mitigate those risks.

The USA National Digital Stewardship Alliance have proposed four levels of Digital Preservation [2] which are defined as:

- Level 1: Protect your data
- Level 2: Know your data
- Level 3: Monitor your data
- Level 4 Repair your data



There are then recommendations for five areas: storage & geographic location; file fixity and data integrity; information security; metadata and file formats. Whilst not all of these areas can be considered to be bit preservation, the first three areas are important to ensuring the persistence of bits. So at level 4, which is the most comprehensive level, some of the recommendations include having at least three physical copies in different locations and on IT infrastructure which is subject to different risks (i.e. different manufacturers) and that fixity checks should be done after every event and no-one should have write access to all copies. These decisions for a particular organisation should be made in response to risk assessments and policy.

It could be argued that basic bit preservation activities are applied everywhere that IT systems are in use, however the issue is the ability to keep the bits intact over a long period of time. In the digital preservation literature, bit preservation has been a controversial and confusing subject, David Rosenthal's paper from 2010, "Bit Preservation: A Solved problem?" [22] p.144 attests to that:

"If bit preservation were a solved problem then it would be reasonable to expect that no bits would be lost. This is not the case; just as in paper archives, preserved content in digital archives will be lost or damaged. Setting unreasonable expectations for the performance of our preservation systems, for example by continually making unsupported claims to have solved the bit preservation problem, is simply setting ourselves up to be perceived as failures."

From a hardware and media perspective there are advances in technology which provide storage solutions which aim to detect errors as part of the standard working of the storage and there is consensus in the digital preservation community that the improvement of bit preservation in storage systems depends 1) on advances in the state-of-the-art storage technologies developed by the storage industry in partnership with academic research and 2) on the competition in the market. In practice, however, digital archives may not be able to afford state-of-the-art bit preservation.

The content itself impacts on the bit preservation decisions as issues such file formats affect the size of content to be stored, some file types take more space than others to store the same information and the error proneness of the format. Other factors such as confidentiality matters and access times also affect how the digital objects are stored. Once the content has been ingested then there is a need to ensure that it is kept unchanged, that the object integrity is not damaged, this process is called fixity checking. Fixity checks should be done on ingest, to check what has been received has been accepted correctly, when it is moved onto new media or into a new system and then there are regular, random, checks to see if existing content is uncorrupted. There is a delicate balance between looking for errors and causing errors by repeated reading of media as reading & writing to disks and tapes cause wear on the system.

Bits are either lost or safeguarded. Bit-preservation applies to everything an institution decides to store and keep. It makes no sense to store something without doing bit-preservation. However, there are bit-preservation methods that can minimize the rate of loss.



The aim of the SCAPE project was to investigate the issues that large scale collections bring to the subject of digital preservation. It can be straight forward to manage tools and activities to support small amounts of material, but are these scalable to real life collections? Some of the SCAPE partners are content holding institutions and their expertise has been used to write this report. Real-life experiences are shared through the use of two case studies.

10.1 Scope

This report concentrates on the issues surrounding ensuring that the bits are kept safe and are known. It is not designed to give technical information on setting up bit preservation infrastructures.

10.2 Audience

The audience for these guidelines are those who are responsible for digital preservation, both at a technical and managerial level and those who are about to become responsible for keeping bits safe.



11 Policy, Management and Risks

The management of the infrastructure to enable bit preservation is, on one hand, no different to any other IT infrastructure as corrupted data causes problems for all systems; however the major difference is in the initial intention to ensure the bits are unchanged and accessible over a long time frame, beyond the current hardware, software and storage media in use.

To achieve this longevity there are three interdependent strategic aspects which need to be considered & addressed:

- **Preservation policy:** what is the preservation remit of the organisation, what is to be preserved and what bit preservation specific policy is in place?
- **IT infrastructure:** how are the bits going to be stored and managed?
- **Risk management:** what are the risks to the content which need to be considered and mitigated?

These aspects do not stand alone and need to be put into an organizational context with considerations of budgets and staff resources and capabilities to be considered.

It should be recognised that some data collections might have a higher preservation value than others, so that different preservation regimes might apply within the same repository – such as for example in the case of the BnF. [8].

11.1 Policy

An organisation should define policy for all parts of the preservation activities; however this section is concerned with policy specifically for bit level preservation. The SCAPE Catalogue of Policy elements (D13.2) discusses the following topics:

- Define Bit preservation
- Define Bit preservation levels
- Decide on Ingest activities
- Develop Integrity Measures
- Assign Persistent Identifiers
- Decide on number of copies, geographical distribution and organisational distribution
- Define Policy for Disaster Recovery

The USA's National Digital Stewardship Alliance [2] proposed four levels of Digital Preservation which addressed the topics of storage and geographic location; file fixity and data integrity; information



security; metadata and file formats. Each level builds on the previous one and so to achieve level 4 (the highest level) the greatest investment needs to have been made.

11.2 Management: IT Governance frameworks

There are established frameworks and best practice models in IT Governance which can be used in Data Preservation infrastructures as described in earlier work done in the framework of Erpanet [11] and work from more recent DP literature [18] & [5] Governance frameworks and management best practices have evolved to help organisations ensure alignment between IT services with business needs and realising optimal value from IT assets.

The following set of 3 complementary frameworks and best practices are relevant to organisations that consider digital preservation is core to achieve their mission:

- i. The COBIT framework is an authoritative, international set of generally accepted IT-control objectives for day-to-day use by business managers and owners, IT professionals and assurance professionals.

COBIT version 5 has five key principles for the governance and management of organizational IT which ensure that the needs of the stakeholder are met, that the IT infrastructure covers all parts of the business and has an integrated holistic approach and ensures that the governance of the services provided is separated from the management. The principles are supported by seven categories of enablers which cover practical aspects such as the services, processes, staff skills and organizational structure and the cultural aspects such as the principles, policies, ethics and behavior as well as the underpinning aspect of information.

- ii. The Capability Maturity Model (CMM) is a development model in which the term "maturity" relates to the degree of formality and optimization of processes in an organisation. IT-related processes become more mature as they develop from ad-hoc practices to well-managed processes, which follow quality improvement cycles through performance measurement and optimization. CMM originally aimed to improve software development processes, but it is also applied to other IT-related business processes.
- iii. At a more operational level, the Information Technology Infrastructure Library (ITIL) is a set of practices (procedures, tasks and checklists) that enable good practice IT service Management (ITSM) and so allows an organization to establish a baseline from which it can plan, implement, and measure. It is used to demonstrate compliance and to measure improvement.

11.3 Risk assessment and management

Risk assessment and management follows a standard set of steps although the methodologies and



supporting tools may be different from organisation to organisation. The key stages are:

- Identify the assets that are being managed and the relative value of each
- Identify and assess threats in the context of the organisation
- Assess the vulnerability of the important assets to the threats identified
- Determine the risk and consequences of the threats actually occurring
- Identify ways to minimise those risks and consequences
- Prioritise risk reduction measures based on a strategy.

The threats to digital information are likely to fall into one of the following areas:

1. Physical environment
2. It infrastructure: Hardware and Media related issues
3. Data security issues/Malicious damage (internal or external)
4. Software related issues
5. Organisational failure
6. Curatorial errors

Although all these threats may damage the bits, in this document we will be focusing on those areas which are concerned with maintaining the bits directly and so will not be covering in detail items 4 to 6.

The two main risk assessment methodologies developed for digital preservation specifically are DRAMBORA [8] and the SPOT Model [19]

DRAMBORA was developed for use in digital repositories and enables the auditing of a repository so that appropriate risk assessments can be made. It provides a toolkit to provide a standard method of capturing: mandate and scope; activities and assets; risks and vulnerabilities associated with the repository. The risks can then be assessed and mitigated on.

The SPOT Model for risk assessment enumerates sets of threats associated with six properties of successful digital preservation (availability, identity, persistence, renderability, understandability, and authenticity). The SPOT Model defines the “persistence” property as follows:

“Persistence is the property that the bit sequences comprising a digital object continue to exist in a usable/processable state, and are retrievable/processable from the medium on which they are stored.”

Defined as such, “persistence” covers bit-preservation fully: it covers the threats of bit rot and data

loss. Bit-preservation or persistence is a basic requirement for digital preservation. It is the minimal level of preservation.

The SPOT Model identifies following policy areas that address persistence: *“the major threats to persistence reside in physical media management, media refreshment policy, hardware migration policy, and data security policy.”*

The implementation of these policy areas typically fall under the remit of the (external or internal) IT-organisation or computer centre with whom the institution has negotiated appropriately designed service level agreements. The SLA’s should reflect how the data redundancy policy, adherence to proper storage condition standards, storage medium refreshment policy, security policy, etc. are implemented.

Further reading in the bibliography 15.2

11.4 Physical Environment Risks

Risks to the physical environment should be assessed and form part of any Business Continuity Plan. Physical risks which have the greatest potential impact on the preservation of bits include:

Issue	Outcome	Mitigating actions	Notes
Damage or destruction of all or part of the building housing IT infrastructure.	Building destroyed, computing hardware and media damaged and consequently loss of bits	Building designed to minimise risks such as flooding and fire. Appropriate fire suppression systems put in place. Consideration of the number of locations these bits/files are stored in to ensure more than one copy Back-ups, where appropriate include a copy kept off site	The policy decisions about the number of copies kept will depend on many factors including the remit of the organisation, the importance of the availability of the content and rarity of the material.
Working environment in the machine room rendering the hardware & media unusable, such as air conditioning failure or incorrectly set	Data destroyed Loss of hardware	Active monitoring of the air flow within the computer room.	Large scale computer/machine rooms need to be kept at the appropriate temperature for the hardware within it as temperature spikes can cause hardware failure.
Power failure causing hardware to shut down mid-action	Data destroyed Loss of hardware	Ensure key equipment is on an Uninterruptible power supply system so that it continues to run	

		<p>even if mains power is switched off</p> <p>Where possible ensure that systems are resilient to sudden power spikes or loss of power</p>	
<p>Unauthorised access to the building and hence the hardware & media enables non-staff member to enter causing risk of malicious damage</p>	<p>Data destroyed</p> <p>Loss of hardware</p>	<p>Ensure that access to the machine/computer room is controlled to ensure only specific staff have access.</p> <p>Ensure visitors are accompanied at all times</p>	

11.5 Hardware and media risks

There are many ways in which bits can be lost – from accidental commands by those who are responsible for using & running systems, through unexpected consequences of upgrades to silent bit rot. This section discusses some of the issues & risks that may need to be addressed when dealing with data on the large scale, which requires comprehensive infrastructure to support bit level data management. These risks are grouped into three headings related to (1) human actions; (2) system management and (3) technical failure.

It assumes that the infrastructure is managed in-house and the bits may be stored on either disks or tapes depending on the requirements of the system utilising the infrastructure. These risks and mitigating actions are based on the bit level infrastructure run by STFC, but are not necessarily domain specific.

For preservation at the large scale, then there is the requirement for much storage – both spinning disk and tapes. The management of this type of infrastructure and how the data flows through the infrastructure depends on the requirements for the service being provided. The value of specific data will mean that there is a different approach to the number of copies and the media used.

11.5.1 Risks related to the way people work

Not all the risks to data come from technical failure and obsolescence, inadvertent errors by those who support and use the services can also cause problems for long term bit preservation. These risks are the hardest to mitigate against as they are the most unpredictable.

Issue	Outcome	Mitigating actions	Notes
<p>File deletion by user/service admin/system admin</p>	<p>Data destroyed</p>	<p>Checks and balances within the service to ensure no unauthorised deletions</p>	<p>If deletion allowed, then no process will stop mistaken deletion of the wrong files.</p>
<p>Disk partition containing data deleted by system</p>	<p>Data destroyed</p> <p>Loss of files on part or all</p>	<p>Ensure staff are competent</p>	

admin	of a disk server		
A data base admin accidentally drops or modifies a production database table	Loss of data in the table. The database may hold the catalogue for the service	Limit access to suitably trained staff. Ensure back-ups and data audit trails	
A set of tapes IDs for tapes containing data are accidentally placed in a free tape pool, thus indicating that they are available for use	That set of tapes may be overwritten	Ensure that that system will not overwrite tapes that are recorded in the system as containing data.	This is a system specific mitigation
Where tapes are removed from a tape robot existing protections within file storage management system will be bypassed.	Many (10s-100s of) tapes	This operation is exceptionally rare; tapes are not routinely removed from the robot.	Eject followed by re-label is a good way to lose custodial data
Where tapes are found to be faulty a system admin may choose to directly access the media - for example overwrite its contents to test it. A typographical error would lead to the wrong tape being overwritten	Contents of the tape destroyed.	Manual tape interventions are rather rare. Staff take care not to make mistakes.	

11.5.2 Risks related to the management of the system

These next set of risks and mitigation factors are about the way that the system is managed. These examples come from a specific service, but can be generalised for any computing infrastructure.

Issue	Outcome	Mitigating actions	Notes
Disk server incorrectly marked for "recycling" or non-recovery after an incident.	Disk server incorrectly "wiped"	Hardware database/inventory to track server state. Instructions to clear file system have to be requested through helpdesk ticket Detailed written process Out of hours recovery discouraged	Amount of data loss depend on how much disk is supported by the server. This could be a large amount in the TB.
An database upgrade has unintended	An upgrade destroys or modifies meta data so	Change control process. Upgrades tested on	

consequences	data is no longer identified	snapshot of production database. Final database snapshot before patches applied. Routine backups and journals	
Logic error introduced by upgrade of dark data cleanup causes mass file deletions.	Data managed by the catalogue	Difficult to track	This is where files on the hardware which don't have a record in the metadata system are automatically deleted from the file system.
Upgrade to the catalogue system destroys the data through unexpected changes to the schema	All data destroyed/unavailable	Change control process. All schema upgrades are tested on snapshot of production database. Database is backed up prior to schema upgrades. Check files written into storage system before upgrade and validated before production is restarted. Routine database backups and journal files.	
A RAID controller firmware update on a batch of disk servers causes loss of device contents. For example by initiating an array rebuild	Loss of content of a generation of disk servers (1PB approx.)	Change control process. Ensures firmware is properly tested. Rollout is phased in.	
A Quattor ⁹ software configuration change accidentally overwrites data partitions	Loss of contents of a generation of disk servers	Protective measures in Quattor to limit activities to primary device. Quattor does not delete files.	

11.5.3 Technical failure

There will always be some technical failure and the potential impact of the failure determines how the resilient the infrastructure is built to be.

Issue	Outcome	Mitigating actions	Notes
Disk driver on	No loss of data if within the	Ensure RAID is appropriate	

⁹ Quattor is a **system administration toolkit** providing a powerful, portable, and modular set of tools for the automated installation, configuration, and management of clusters and farms.

RAID array fails	recoverable number of drivers for the array	Always replace disk drivers when they fail	
Snapped tapes	Data on the tape will be lost	Ensure tape wear is monitored	CERN use rule of thumb that a tape life is 5000 accesses
Tape drive fails	Possible loss of data during write action	Always replace tape head when they fail	

11.6 Data Security risks

Ensuring that there is no unauthorized access to the bit preservation infrastructure, either from internal or external people is an important part of ensuring effective management. see also section 2.4 and 2.5.1

Issue	Outcome	Mitigating actions	Notes
Files/bit level infrastructure accessible over the network through incorrect security measures	Data destroyed	Ensure that the service has taken appropriate security measures Checks and balances within the service to ensure no unauthorised deletions	



12 Technical Approaches

There are many different approaches to ensuring the technical infrastructure supports the organisation's requirements for bit preservation. This infrastructure may be supported by a commercial preservation system, by in-house technical development or may be out-sourced to the cloud. This section discusses approaches using commercial systems and cloud preservation systems. For examples of use of contrasting approaches, see also section 4: Case Studies.

12.1 Commercial Preservation Systems

While digital preservation is generally conceived as chiefly concerned with format obsolescence, hardware integrity is a precondition to any digital preservation activity. As such, a preservation system should address the numerous and various risks concerning bit preservation by providing tools to ensure bit-health of the repository filestreams over time. These tools should either be an integral part of the preservation system or integrate with it by appropriate APIs. Proper procedures and security policy are fundamental to any information system and are typically set by the institution's information security officer

Bit preservation needs to mitigate for two types of risk – human and infrastructural. The human factors should be dealt with by the institution's information security officer through policy and procedures. Those relating to infrastructure are in part determined by the digital preservation system as that one that manages storage will typically move files from location to location, thus assuming responsibility for file integrity. Additionally the system relies heavily on data integrity in order to extract technical metadata and significant properties that are critical for digital preservation risk analysis. Finally, users' need for a single solution that ensures accessibility presents a requirement that bit and digital preservation be handled as one.

The following describes aspects of bit preservation within the context of a digital preservation system, exemplified by ex Libris's Rosetta. Rosetta's fixity capabilities are based on combination of users' requirements with community standards, and are already part of institutions' digital curation and preservation workflows.

Internal Fixity checks. These checks are run within the repository and cross check the stored checksum against the actual checksum of the file in storage. A match will be recorded as statistical event, while a mismatch be recorded in a more notable manner, (see below) and allow for additional actions.

External (Storage layer). Depending on the hardware, best results may be best achievable by allowing the storage layer to manage and run fixity checks. A typical use case is tape storage, where the storage layer is best suited to determine the optimal method of retrieving this information. In such cases, APIs are called by an external application to retrieve the stored checksum value and provide to the storage layer's tool, and to update the preservation system with fixity results, creating appropriate preservation events.

The frequency of checks to ensure content is unchanged is best determined using criteria such as the type of storage hardware (reliability, durability, performance etc.) and other considerations such as expense. These considerations vary from repository to repository and often from collection to collection, and repository managers are at the best position to set appropriate policies in place.



Selecting a checksum algorithm (or several algorithms) should be a possibility – either as built-in functionality or as part of a plug-in framework.

Redundancy is achieved by storing these details in both the file system and the database, both of which should be protected by appropriate backup strategies.

Other fixity types are targeted at preventing malicious attacks, e.g. tampering with checksum results. These include cryptographic hash functions and digital signatures, used to authenticate the signer of an object and/or the information contained in the object, allowing for verifying the identity of the depositor and that the file was unchanged in transmission. PREMIS has issued recommendations for using these methods, and the system should be able to accommodate storing the relevant information per the PREMIS data dictionary [29]. As this type of validation is more concerned with content integrity, a fuller evaluation of these tool and recommendation exceeds the scope of bit preservation and should be addressed in the context of digital preservation itself.

Storing information regarding date and outcome of fixity checks is crucial for monitoring bit health. Ongoing monitoring of fixity checks should be done via a PREMIS Event-driven engine. We recommend differentiating between several fixity-related events:

- Initial fixity (checked and/or generated during loading)
- Ongoing fixity checks with no status change
- Ongoing fixity checks with status change

While each of the above three checks generate an event, we regard the first and third check as more substantial than the second, which should serve mostly as an indicator for analysing the outcome of the third type. In other words, a fixity check failure would suggest looking at the date of last successful fixity check as a point of reference for running fixity checks on other files on the file system in question, as well as identifying an appropriate backup, should one exist.

Accordingly, the first and third generate a provenance event, which is stored on the object level, while the second generate a statistical event, retrievable for reporting purposes, but one that will not become part of the object metadata, and if the object is exported from the repository the information will not be retained with it.

The content of the provenance event information is to be stored as PREMIS events on the file level. The event should include the date, fixity algorithm used, and outcome.

Retrieving information on the outcome of the fixity checks is necessary in order to evaluate the health of the repository. Reports indicating the result of the checks will determine additional steps to consider such as restoring and hardware replacement. Reports on fixity failure should include full timestamps and paths to the filestreams, along with detail of the last successful check, providing system administrators with the all the required information to take more thorough action. Reports should be delivered automatically and independently of the preservation system (xls, pdf) to all stakeholders.

12.2 Bit preservation and the cloud

One of the recent developments in technology is the ability to outsource data to external storage providers of remote on-line storage. As Zachary P et al [29] noted the exponential growth of electronic data has led private organizations and governmental agencies, with limited storage and IT resources, to outsource data storage to cloud-based service providers. This business model can be a



cost-effective one, but for those who are responsible for the long-term preservation of material, there are some issues that need to be considered before choosing this type of arrangements.

The providers of the storage service preserve and make data available for retrieval under the conditions of a formal service level agreement (SLA). In addition to availability, SLAs may also guarantee that data will be stored only at data centres within a specific geographical region for performance, regulatory and continuity reasons.

When it comes to preserving the bits in the cloud there are two important questions:

- Are all the bits accessible?
- Where are the bits?

12.2.1 Are all the bits accessible?

Cloud storage offers clients a logical view of their files and collections, without detailing how they are actually stored in the infrastructure. This abstraction according to Kevin B et al.[31] is appealingly simple. In reality Cloud Service Providers (CSP) generally store files/objects with redundancy or error correction to protect against data loss. Amazon and Microsoft, for example, claim that their S3 services store three replicas of each object. Additionally, cloud providers often spread files across multiple storage devices. Such distribution provides resilience against hardware failures, but these are not visible to the clients and verification of the storage policy is difficult but essential. Remote testing of fault tolerance is a vital complement to contractual assurances and service-level specifications.

Currently there are multiple techniques to tackle this challenge of ensuring the data is complete and accessible without downloading everything, the following list only represents core techniques:

Proof of Data Possession (PDP)

In the Carnegie Mellon University Research showcase “Provable Data possession at the Untrusted Stores” [32] a new model for provable data possession is defined. Archival storage requires guarantees about the authenticity of the data within the storage, namely that storage servers possess the data. It is insufficient to detect that data have been modified or deleted when accessing the data because it may be too late to recover lost or damaged data. Archival network storage presents unique performance demands. Given that file data are large and are stored at remote sites, accessing an entire file is expensive due to input/output costs of the storage server and in transmitting the file across a network. Reading an entire archive, even periodically, greatly limits the scalability of the network stores. Clients need to be able to verify that a server has retained file data without retrieving the data from the server and without having the server access the entire file.

The model for provable data possession (PDP) which provides probabilistic proof that a third party stores the file is unique in that it allows the server to access small portions of the file in generating the proof; all other techniques must access the entire file. This model enables provable and secure scheme for remote data checking

Proof of Retrievability (POR)

According to Qingji Z. and Shouhuai X. [33] POR allows a cloud storage provider to convince the data owner that its outsourced data are kept intact. Existing POR schemes can deal with static data and are not secure when used to deal with dynamic data. Intuitively, the difficulty can be



attributed to the fact that the retrievability property is more demanding than the possession property. Another problem inherent to dynamic POR is *fairness* which ensures that a data owner cannot falsely accuse a cloud storage service provider of manipulating its data. Note that *fairness* in the setting of static POR is easily solved, for example, by requiring the client to digitally sign its data before the data are outsourced to the server. In the setting of dynamic POR, however, the problem is challenging because the updated data is held in the Cloud Storage. One solution is to download and sign the whole data after each update operation, which is also clearly not acceptable in practice because of the communication costs. **Fair and dynamic proof of retrievability (FDPOR)** is a useful extension of static POR in practice. Efficiently designed FDPOR scheme simultaneously offers both retrievability and fairness in the setting of dynamic data.

Remote Data Checking (RDC)

According to Bo C. and Reza C. [34] RDC is a technique that enables the checking of the integrity of data stored at a third party, such as a CSP. RDC can be used for data auditing, allowing data owners to assess the risk of outsourcing data in the cloud. In an RDC protocol, the data owner (client) initially stores data and metadata with the cloud storage provider (server); at a later time, an auditor can challenge the server to prove that it can produce the data that was originally stored by the client; the server then generates a proof of data possession based on the data and the metadata. Several RDC schemes have been proposed for static data, including Provable Data Possession (PDP) and Proofs of Retrievability (POR), mentioned above. RDC schemes have also been proposed for the dynamic setting PDP, which supports updates on the outsourced data. A scheme for auditing remote data should be both *lightweight* and *robust*. Lightweight in that there are no significant processing and bandwidth burdens on the infrastructure which can be achieved by spot checking random small samples. Robust in that the auditing scheme has mechanisms to mitigate arbitrary amounts of data corruption which is usually achieved by integrating forward error-correcting codes (FECs) with remote data checking. Although there may be tension between FECs and dynamic data as securely updating even a small portion of the file may require retrieving the entire file.

Remote Assessment of Fault Tolerance (RAFT)

Kevin D et al.[31] also develop and describe a protocol for remote assessment of fault tolerance for stored files (RAFT). It enables a client to obtain proof that a given file F is distributed across physical storage devices to achieve a certain desired level of fault tolerance. Storage is referred as units of drives. For protocol parameter t , these techniques enable a cloud provider to prove to a client that the file F can be reconstructed from surviving data given a failure of any set of t drives. For example, if Cloud Service provider were to prove that it stores a file F fully in triplicate, i.e., one copy on three distinct drives, this would imply that F is resilient to $t = 2$ drive crashes.

12.2.2 Where are the bits?

Moving to the cloud requires organizations to interact with their data at a new level of abstraction. This comes with significant benefits but also has some limitations which are the motivation for position paper on Data Sovereignty [29]. According to this paper verifying that cloud storage service providers are meeting their contractual geographic obligations is a challenging problem, and one that has emerged as a critical issue. For example, careless or naive storage service providers may move



data, in violation of an SLA, to an overseas data centre to leverage cheaper IT costs. Such actions, however, may make data available to foreign governments through search warrants or other legal mechanisms.

Data sovereignty protocols may also be a complementary technology providing solutions to other data security problems. For digital provenance, when determining the origin and history of a digital document, one of the most fundamental questions is: where is this data right now? With no reliable answer to this question at any point in the data's lifetime, one may never establish reliable provenance data.

Within the problem of data sovereignty, key concerns include developing techniques that minimize storage and network (thus, economic) costs. Tools which break the abstractions of the cloud to geolocate data, may be essential in the future to gather evidence and establish compliance (or show non-compliance) with contracts and laws. The problem of verifying that data exists only at allowed locations—and copies have not moved to some location that violates a policy— is a difficult problem in general; data sovereignty provides a much weaker guarantee but it is a step toward actively monitoring compliance with some SLA policies.

Further reading in the bibliography 15.3



13 Case Studies

The two case studies discussed here demonstrate different approaches taken by two SCAPE partners and reflect their remits and key concerns.

The British Library is responsible for legal deposit for the UK and has in place a Digital Library System to manage the content it is responsible for preserving. The CERN Tier1 centre at the Science and Technology Facility Council is not responsible for long-term preservation, however it does have a responsibility to ensure that the bits have been received correctly and are managed according to good practice.

The technical infrastructures run by these two organisations are similar in that they are both part of a wider collaborative infrastructure which ensures geographical and technical separation of the content but the BL is responsible for the bits, the content and access, whereas the STFC Tier1 centre is responsible for the bits and ensuring access to them, but content management, what the files are and where they should be kept, is the responsibility of the CERN experiment that generates it. This collaborative approach to minimising the loss of collections can also be seen in approaches such as LOCKSS (Lots of Copies Keeps Stuff Safe) in the electronic journal domain.

13.1 Bit preservation at the British Library

The British Library is the main legal deposit library for printed and electronic material, which has been published in the UK. The other UK Legal Deposit Libraries are: the National Library of Wales, the National Library of Scotland, the Bodleian Library, Oxford, the Cambridge University Library and Trinity College Library, Dublin. The British Library uses a Digital Library System (DLS) to store the content which is accessed by all UK Legal Deposit Libraries. The vision behind the Digital Library System is to have shared technical infrastructure for non-print legal deposit in the UK.

The DLS is built from the bottom up with many of the requirements of a Trusted Digital Repository (TDR) in mind. The whole basis of the system is in-perpetuity and trust (authenticity). The DLS assumes that digital media will fail and that failure may be silent as well as hard failure. The system currently has about a petabyte of managed storage which is continually growing, the rate of which is projected to increase to about half a petabyte per storage node per annum over the next few years.

13.1.1 Principles guiding the design of the Digital Library System

From a preservation perspective it is assumed that bit loss is inevitable and is mitigated by having the digital object replicated across all four storage nodes within the system. It is important to ensure that there are multiple uncorrupted copies so that if there is an issue at a specific node, the content will be reingested from another node. Corruption is identified by periodic checking of the integrity of the objects, through inspection of the check-sums and signature files. All digital objects include preservation metadata to ensure the preservation of meaning and context of the object. Due to the distributed nature of the overall design, multiple copies of digital objects are stored at different geographic locations on different devices which also enables a technical separation so that administrators at one node can only access that node. Additional security measures include a separation between the backend and user facing systems.

The DLR has been designed to be able to expand both the content and new requirements and services to minimise re-engineering. It is also designed to handle failures gracefully and enable processes to be restarted from the point of failure rather than the start to minimise disruption. The storage used for the BL system is intended to meet the design goals of vendor independence, standard commodity hardware and the ability to be extended easily.

13.1.2 Architectural overview of the storage solution

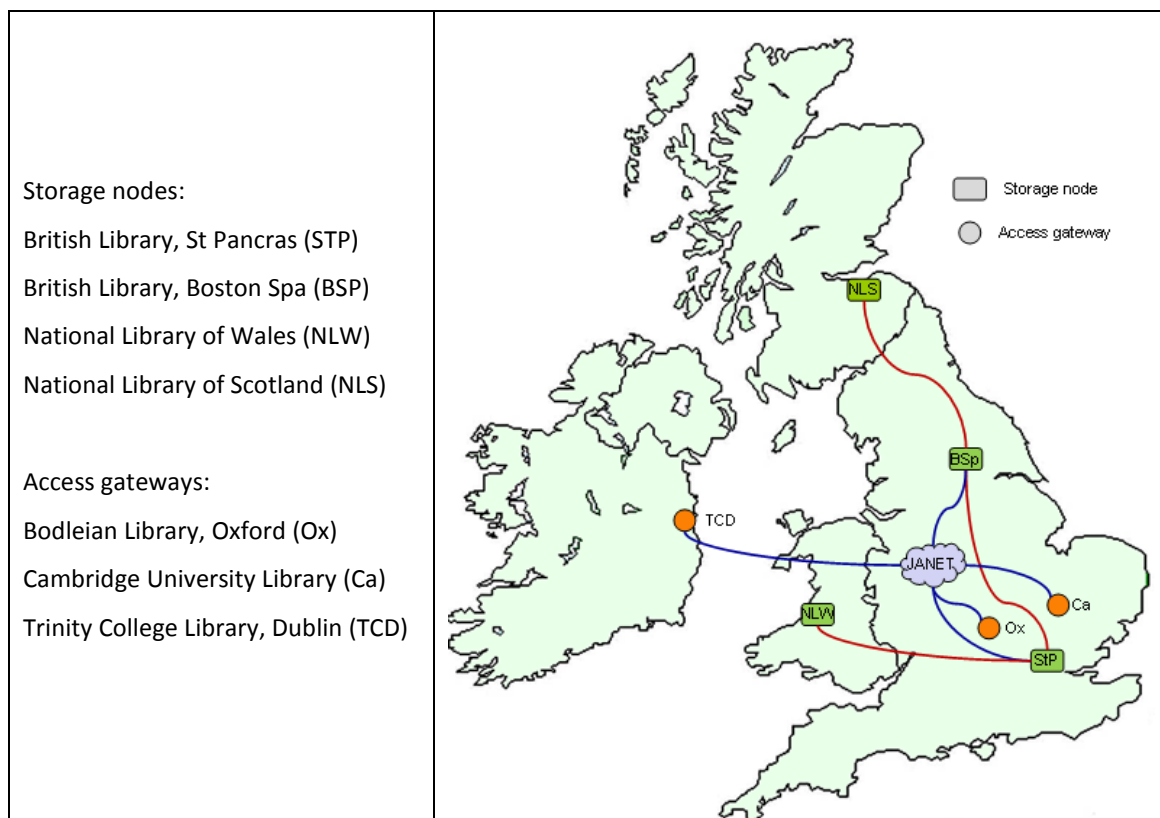


Figure 7: Storage nodes and access gateways

Four national centres hold full copies of the system. These are the storage nodes depicted above. The British Library has two nodes, one at its St Pancras site in London and the other at its Boston Spa site in Yorkshire. The university legal deposit libraries are entitled to access legal deposit content but they do not hold full copies of the system. They access the content across Janet4, the UK's research and education network.

Ingest of digital objects to the DLS takes place using ingest systems at either the BL's St Pancras or Boston Spa sites. There is a variety of ingest streams at each site; for example the BSP systems deal with e-journals, voluntary deposit items, digitised newspapers, web archive content. During the ingest process, each digital object is assigned a storage identifier called a DOM id. This is written into the METS metadata file that accompanies each object. Metadata about objects is held in the Metadata database. The METS file contains all relevant digital preservation metadata in PREMIS form. This includes any events relating to format validation, format migration and format characterisation etc. Each digital object has a signature file, created during ingest by a Digital Signing



Engine, or Signer. The signature file contains a hash value for the object, a secure timestamp and uses certificated cryptography.

The hardware used on each storage node is different; different storage media is used, including base operating system on at least one of the nodes to mitigate common mode failure. There is no trust model between the different stores; the digital signature of each object is used for trust. The security profile of the system is very high, using a layered approach and a very limited interface to the stores. Independent penetration tests are performed on a regular basis. No staff have access to all the nodes and each node operates in its own domain. Each node verifies the signature of each object on receipt (either for ingest or replication). The signature file can be used to detect corruption or tampering in objects and every object is checked on a regular basis. If any defect is found, automatic recovery from another node takes place.

Every object ingested is replicated to each of the four nodes in the system. A non-functional requirement of the system is to check the authenticity and integrity of each object at least every thirty days on each storage node hence it will take the coincidental corruption of all four copies within a thirty day period to lose the object. If corruption is detected it is automatically recovered and re verified from one of the other storage nodes.

Access to the digital objects is via an Access Gateway at the firewall of an individual node. The legal deposit libraries will usually access their nearest node, but resilience is built in to the system so that in the event of a disaster at one node, another node will be able to provide services whilst restoration is taking place.

13.2 STFC Large Hadron Collider Tier 1 bit storage

The observations in this section have been collated with the assistance of staff at the UK's Large Hadron Collider Tier 1 Data Centre.

The computer infrastructure for the Large Hadron Collider at CERN [47] is designed to achieve geographic and technical separation. There is a single Tier 0 centre at CERN which is responsible for the safe keeping of the raw data and the initial analysis. The raw and analysed data are distributed to the eleven world-wide Tier 1 centres, of which STFC is one. Tier 1s are responsible for holding a share of the whole data and for distributing parts to the Tier 2 centres which are mostly based in Universities and other scientific institutes, there are 140 of these centres world-wide.

The Tier 1 Data Centre is a bespoke and sophisticated data management system which is responsible for storing and providing access to data in real-time and on a very large scale. The issues around ensuring that the bits are there and are the bits that one would expect to be there are similar to those systems which have a bit preservation remit

13.2.1 Principles guiding the design of the Bit level infrastructure

The infrastructure to support the Large Hadron Collider at CERN depends on the Worldwide LHC Computing Grid (WLCG) which is a global collaboration of computing centres. It is designed to store, distribute and analyse the 15 petabytes of data generated each year by the LHC.

This infrastructure is a distributed one, so that multiple copies of data are kept at different geographical locations, managed by different teams using different technology. This means that there is no single point of failures, the advantages of having different data centres in different time



zones means that the system, overall, can be monitored 24 hours a day for the service provided to the researchers.

13.2.2 System description

The approach taken by the STFC Tier 1 Centre is to use routine hardware, but to invest in the staff to support this service. Both discs and tapes are used to support this service, as although disks give rapid access they are more expensive to run than tapes due to the costs of keeping the disks spinning. It is rare that the STFC Tier 1 holds the only copy of the data due to the data management arrangements. The Tier 1 service provides computing resource with the logical equivalent of 10 thousand CPUs, 11PB of disk and 11 PB of tape storage.

The service performs a selection of routine maintenance tasks to ensure that the bits and files are still available and these are described in the next sections.

Disk Infrastructure

The STFC Tier1 data infrastructure has approximately 200 disc servers providing 7PB of storage. The data files written to those disks come over the network from CERN, other sites around the UK and worldwide within the LHC collaboration. The checksum algorithm Adler32 is used as it is very fast although it is not the most robust algorithm as it can produce clashes. The checksum calculated as the file comes in over the network is stored both in a database and when the file is written to disk from the cache the same checksum is stored as an extended attribute within the file. Once every 24 hours, all the preceding days new writes are checked, a new checksum is generated and compared between the checksum held in the database and that which is kept in the extended attributes.

In addition once every 10 minutes a random file from each of the disk servers is chosen and the same checks are performed. The choice of 10 minutes between checks is because tests revealed that the biggest file took 5 minutes to perform this action and so 10 minutes gives enough leeway to ensure jobs have finished before new ones start.

If a mismatch is identified then the user of the file is informed and:

- 1) If the file is only on disk,
 - a) if the database or file attribute checksum is wrong it is replaced by the new file generated checksum as it is assumed to be a checksum error;
 - b) if the file checksum doesn't match then the production team will inform the owner, via the Helpdesk as there may be a problem with the file itself. The owner can choose to delete the file and retransfer from another source
- 2) If the file is on the disk cache in front of the tape; then the file is removed from the disk and retrieved from tape and checksums re-run.

The disk and disk servers are on a formal replacement plan and are replaced every 3-5 years. This mitigates the risks of old hardware although it should be noted that transferring files from disk to disk carries risk of bit loss as well.

Checksum failure rates – STFC Tier 1

The data infrastructure runs a random checksum comparison job on a file on each of the 200 disk servers once every 10 minutes looking for corrupted files.

In 4 years of checking (2009 – 2013) approximately 70 files of the 53,000,000 files on the 7PB of disk have been found to be corrupted.

Tape Infrastructure

STFC, based at RAL, runs two StorageTek tape robots. There are 10,000 tapes slots and 64 slots for Tape drives in each.

At present these are using T10K series of tapes and tape read/write heads. This technology uses both changes in Tape drives and media to increase the amount of data stored on the tapes. Each improved Tape drive means that more data can be written on the same tape. There is a 2/3 year gap between each new development.

Tapes driver	Activity	Media change from previous version	Notes
T10KA	Can be read & written as A	T10KA tape	
T10KB	Can read A tapes but writes B tape	Same media as T10Ka	
T10KC	Can read A & B tapes but writes C tapes	Change in media required for writing C tapes	C drives has the ability to checksum each block on the tape and so a media scan will identify issues.
T10KD	Can read A, B & C tapes, can rewrite C tapes and writes D tapes	Same media as T10KC	About to be released

Tapes have a finite life, which depends on their use but is around 5 years for active tapes. Tape technology has moved to BaFe for the magnetic layer. The archive life of the tapes has moved from 10 to 30 years.

The tape robots are monitored by software which looks at both the tape drives and the tapes and can see the traffic on the fibre channel and can look for errors. A database holds the confidence level and if this is breached then it will drain a tape before errors cause a major impact. In particular it monitors for write errors. If a drive has a problem writing to tape, it will retry over the same piece of tape, it also uses the two heads on the tape drive to write with one and read back from another so that errors can be identified, either with a particular tape or with the tape drive heads. If it is the tape then this is switched out, approximately 2/3 a year are switched out; if it is the Tape drive head then this is replaced.

The monitoring program also checks the tapes for the bits in the right place. The control software for storage on Tier 1 also has checked tapes against the catalogue checksums by looking at the first files, end files and random files in the middle. It has checked every full tape and every tape which has not been read in the last 3 years.



Due to the changes in tape technology and the benefits that increasing capacity of the new tapes brings, STFC have done 3 tape to tape migrations in the last 3 years.

Year	Number of tapes	Number of problems	Notes
2010	From 3000 T10KA tapes to T10KB tapes (1.5PB data)	6 tapes with software problems reconciling catalogue to data on tape 2 tapes with hardware issues	This transition halved the number of tapes required
2011	From 3000 T10KA tapes to T10KC tapes (1.5PB data)	5 tapes	No data lost as it was all recovered from other sources
2012	From 2334 T10KA tapes to 204 T10KC tapes (1.5PB data)	1 tape	4 files lost, all others recovered

All these migrations show a very small level of tape failure, under 0.5% in all three cases. The resilience of the infrastructure is demonstrated by the few number of data files which were actually lost (as recorded in 2011 and 2012)

14 Conclusions and recommendations

These recommendations are based on the expertise within the SCAPE project. They are split into two sections. The first deals with management and policy issues of bit preservation and the second section addresses technical issues and activities.

14.1 Management of the bit level preservation infrastructure

This section addresses the issues of policy, management and resourcing.

ID	8.1
Activity	Policy required for bit preservation
Description	Policy provides a framework in which technical and operational decisions can be made effectively. Bit level preservation policy will support and interact with other IT management policies, but it is important to ensure that it exists.
Guidance	Policy explaining the rationale behind the bit preservation infrastructure will enable a common understanding of the priorities and assist in resourcing decisions. It is likely that this exercise will be done with consideration to risk assessments.
Risks specific to large scale collections	It is rare that all collections have equal value and need to be treated in the same way. By having written policy, the norms and exceptions can be explicit and this may inform resourcing and infrastructure decisions.
Questions	<ul style="list-style-type: none"> • What type of material is being preserved? • Are there any collections of special importance? • Are there any additional resourcing issues with particular file formats? • Is there any central policy about IT infrastructure? For example must it be done in-house or outsourced. • Is there any policy about use of commercial systems vs. open source systems? • Has a risk assessment of the bit preservation activities been undertaken?
Resources and examples	<p>SCAPE Policy Representation deliverable D13.2</p> <p><i>“Parliament may use the services of external contractors or partners to provide preservation and access services for some or all of its digital collections. Decisions about this will be based on the requirements of the collections, and Parliament’s existing or planned capabilities with regard to the required services. Where external services are used, proper arrangements must be in place”</i> UK Parliamentary Archive, Digital Preservation Policy for Parliament. http://www.parliament.uk/documents/upload/digitalpreservationpolicy1.0.pdf (accessed Feb 2014)</p> <p><i>“In order to reduce technology dependencies and to manage the risks of</i></p>



hardware/software obsolescence or storage failure, the technological strategy for digital preservation for Gloucestershire Archives is to:

- *avoid reliance upon single software or hardware products or suppliers*
- *prefer standards based, open source and cross platform (not hardware specific) software solutions to proprietary or patent encumbered solutions"*

Gloucester Archives Digital Preservation Policy.
<http://www.gloucestershire.gov.uk/extra/CHttpHandler.ashx?id=25143&p=0>

ID	8.2
Activity	Management of a local bit level preservation infrastructure
Description	For a successful IT related service there needs to be both policy and good management processes in place to ensure that appropriate standards are followed.
Guidance	<p>Successful data infrastructure management will have aspects of the following:</p> <ul style="list-style-type: none"> • Written processes • Processes to manage the hardware infrastructure: Such as a hardware database/inventory to track server state and a rolling replacement plan so that hardware is current and supported. • Formal testing process for changes to the hardware, firmware and software for the infrastructure • Change control process • Process to track instructions for operational activities which carry risk such as instructions to clear file system have to be requested through helpdesk ticket <p>Other, more ethos related points, are that for good management that:</p> <ol style="list-style-type: none"> i. Out of hours recovery should be discouraged as this may be at an unusual time and there is likely to be less team discussion about the best way of resolving the issue. ii. An awareness of rarity of the operation may encourage staff take pause before undertaking it as mistakes are more likely to occur for unusual operations.
Risks specific to large scale collections	The scale of data means that poor management of the system may lead to greater data loss.
Questions	<ul style="list-style-type: none"> • What external service management standards does your organisation use? • Is the management and governance of the bit preservation infrastructure clear? • What resources are available for the bit preservation infrastructure? • If there is an cross-organisational bit preservation infrastructure are the roles and responsibilities of all the partners clear and agreed?
Resources and examples	<p>See STFC case study</p> <p>See section 6.2 relating to governance</p> <p><i>“Application upgrades and migrations between applications are planned and documented unless these constitute a minor operation.”</i> Archaeology Data Service Preservation Policy</p> <p>http://archaeologydataservice.ac.uk/attach/preservation/PreservationPolicyV1.3.1.pdf (accessed Feb 2014)</p>

ID	8.3
Activity	Using a cloud provider for bit level preservation
Description	An organisation may not wish to be responsible for providing the technical infrastructure and one choice for out-sourcing this activity is using a Cloud Storage Provider.
Guidance & Questions to consider	<p>As part of any decision making process about out-sourcing archival storage to one (or more) cloud providers you may wish to consider the following questions:</p> <ul style="list-style-type: none"> • What level of assurance/process is in place for the fact that the bits/files are satisfactorily stored – how many copies & where? • What process is in place to ensure that the bits/files are checked, and what happens if there is an issue • How can the bits be retrieved? • How long are the bits guaranteed for? • What are the long-term costs for bit level preservation? • Are there any geographic restrictions for the data to be stored in the cloud? • How would any geographical restrictions be adhered to? • Is there any particularly sensitive data?
Risks specific to large scale collections	Costs and time involved in retrieving the whole collection if one wishes to change provider
Resources	<p>See section 6.3</p> <p>Cloud Computing Toolkit: Guidance for outsourcing information storage to the cloud from the Department of Information Studies, Aberystwyth University and the Archives and Records Association of UK and Ireland</p> <p>http://www.archives.org.uk/images/documents/Cloud_Computing_Toolkit-2.pdf</p>

ID	84
Activity	Minimising Human errors when managing bits
Description	<p>It is not possible to design computer systems in such a way that human error is completely removed. The appropriate balance between risk reduction and useable system should be maintained.</p> <p>One approach to mitigating the risk that human error may compromise the bits/files is to ensure that there are multiple copies in existence and that each copy is managed by a different system administrator/team of systems administrators.</p>
Guidance	<p>Only trained & competent staff should be performing system/bit level operations which have the potential to damage bits.</p> <p>Consider whether the data is of sufficient value to require multiple copies managed by different people. Multiple different versions should not be managed by the same person.</p>
Risks specific to large scale collections	The bigger the collection, the greater the risk that an accidental command could delete or damage large parts of the collection.
Resources	See STFC case study

14.2 Technical and Operational concerns

ID	8.5
Activity	Number of copies
Description	This is the decision to have duplicates of the system and associated data to ensure that the bits are less vulnerable
Guidance	<p>For successful preservation there needs to be more than one copy of the objects available. The media on which the copies are stored should also be considered, some storage media such as CDs may be suitable for initial collection, but may not be suitable for long term preservation.</p> <p>There are discussions as to whether back-ups provide a method of ensuring additional copies. Although a back-up will copy the data, usually there are no object level integrity checks to ensure that all objects have been copied successfully and so if integrity issues are identified in the future it may not be possible to identify which copy is uncorrupted. So it is recommended that additional preservation copies or replicas are produced as part of an intended process rather than through back-ups.</p> <p>The number required is set by the policy of the organisation holding the data. This depends on the importance of the data and the resources available to manage it.</p>
Risks specific to large scale collections	The size of big collections means that the decision on the number of copies to keep concurrently will have large resourcing implications both at the point of replication and during sustainability discussions.
Resources and examples	<p><i>“4.2 Long-term storage of electronic records covers a variety of methods and media, including online*, near line* and off line* for both magnetic and optical media. The ideal digital preservation programme should ensure that three copies of a born-digital item, and two copies of a digital surrogate are made available on different storage media in different locations.”</i> Hampshire Records Office (UK) Digital Preservation Policy http://www3.hants.gov.uk/archives/hro-policies/hro-digital-preservation-policy.htm (accessed Feb 2014)</p>

ID	8.6
Activity	Spread of locations
Description	<p>There are potential risks in holding bits in a single geographic location as disasters such as fire, floods, earthquakes or other damage to the buildings or power supply issues can make the bits vulnerable.</p> <p>A number of different locations also gives the opportunity for the use of different IT infrastructure to store the material, thus reducing the risk of loss through a specific hardware or media issue.</p>
Guidance	<p>The number of different geographic locations and whether full systems or just data are held there depend on the policy of the organisation holding the data. This depends on the importance of the data and the resources available to manage it.</p> <p>It is recommended that if more than one copy is held that it is not held in the same building. Ideally there should be some geographic distance to enable additional copies to be held somewhere which would not be subject to the same natural disasters.</p>
Risks specific to large scale collections	It may be more difficult to find suitable alternative locations.
Questions	<ul style="list-style-type: none"> • Do you have suitable off site locations to store additional copies of the collection? • How will the copies be transferred to this location? • How often will the locations be synchronised? • Will the remote site(s) be able to use different hardware and media to store the objects to reduce vendor specific risks?
Resources and examples	<p>See BL and STFC case studies for examples of collaborations enabling geographic spread.</p> <p><i>“In order to ensure resilience and provide an adequate level of redundancy, the preservation system consists of on -site, near-site and off site storage. For the same reasons, mirror versions of on -site systems are provided.. Furthermore, to reduce risk further different operating systems will be installed across the systems.”</i> UK Data Archive Preservation Policy</p> <p>http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf (accessed feb 2014)</p>

ID	8.7
Activity	Hardware refresh
Description	<p>The hardware used to support the storage of the bits will have a finite life; this is usually in the region 3 -5 years. As well as increased risks of failure there is the issue of technical obsolescence to consider. As with all IT infrastructure it is important to ensure that there is a plan for replacement and for the new kit to come into production.</p>
Guidance	<p>Ensure that there is a plan for replacing hardware and the associated media at an appropriate frequency.</p> <p>Ensure that the routine management of the infrastructure looks for hardware devices which are showing unexpected errors and that there is a plan for dealing with these errors which has some form of escalation if they become more frequent.</p> <p>It is, generally, more complicated and expensive if one has to do a big technical change, rather than ensuring the infrastructure keeps pace with changes.</p> <p>Depending on the size and load of the physical infrastructure, it may be beneficial to have a test/load testing period before the hardware is put into service to ensure that they are performing to the specification.</p>
Risks specific to large scale collections	<p>Buying a large amount of equipment at the same time has several risks associated with it:</p> <ul style="list-style-type: none"> • The effort required to test and put into place • Problems associated with the manufacture of a particular batch of equipment – if it is all the same and there is a manufacturing error then all of your infrastructure will be affected.
Resources	See bibliography

ID	8.8
Activity	Media refresh
Description	The media used for storage, both spinning disks and off-line copies on magnetic tape are subject to wear and tear as well as technical obsolescence
Guidance	<p>Ensure that there is a plan for replacing media at an appropriate frequency.</p> <p>Ensure that the routine management of the infrastructure looks for media which are showing unexpected errors</p>
Risks specific to large scale collections	The larger the amount of media in use will mean that there is a greater rate of routine, expected failures.
Resources and examples	<p>See bibliography</p> <p><i>“Every media refreshment action will be verified at the bit level, to ensure that the content has been copied without corruption or loss. Parliament will implement procedures to rectify any errors or losses identified as a result of media refreshment”</i> UK Parliamentary Archives Digital Preservation Policy</p> <p>http://www.parliament.uk/documents/upload/digitalpreservationpolicy1.0.pdf</p> <p><i>“The UK Data Archive operates a media monitoring procedure as part of its AMASS® preservation system This allows it to check for potential future problems of wear and tear on media and act before the problems become severe.”</i> UK Data Archive Preservation Policy</p> <p>http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf</p>

ID	8.9
Activity	Checksums/fixity checks generation
Description	<p>This is the process and series of activities undertaken to ensure that the digital objects in the collection are not corrupted or altered. This can also be described as fixity information.</p> <p>This will ensure that you have the digital objects you were expecting to have; that they are not corrupted or altered and that you are able to prove both of these facts.</p>
Guidance	<p>There are some standard methods for establishing fixity information:</p> <ul style="list-style-type: none"> • Checksums • Cryptographic hashes • Digital Signatures • File counts • File size information <p>There are some standard points in the preservation lifecycle where the fixity of a digital object may be checked:</p> <ol style="list-style-type: none"> 1. On ingest. If the digital object arrives with fixity information, after the ingest process the fixity can be recalculated and checked against the original information. 2. On transfer to another system or different media. Once the object is within your bit infrastructure then it should be possible to check the recalculated fixity information against the original fixity. 3. Routine fixity checking to detect silent bit loss. This is checking the fixity information on a routine schedule to see if there are any changes. An increased rate of errors may identify media (tape or disk) failure or hardware failure (read/write heads for example). <p>Increasingly there are developments in the hardware and media used to support digital preservation systems which support fixity checks within the standard working of the infrastructure.</p> <p>There are some considerations for when choosing the most appropriate method. Reading tapes for any purpose impacts on the life of the tape, and the same is true of the read/write heads on disk drives, so that the act of doing fixity checks may shorten the life of the media/hardware being used. The speed at which the fixity checking process runs may have effects on the general processes, along with the computing power needed to run them.</p> <p>It is important for preservation systems to record the results of fixity checks and any actions undertaken as a result.</p>
Risks specific to large scale collections	<p>The scale of the collection held may have an impact on effective mechanisms for fixity checking, especially for routine checks.</p>

Resources

"The frequency and method of integrity checking will be determined with regard to the susceptibility of the current storage media to corruption, and its performance limitations, and will be periodically reviewed. Parliament will implement procedures to rectify any integrity errors detected, through recovery from an alternative copy." UK Parliamentary Archives, Digital Preservation Policy for Parliament.

<http://www.parliament.uk/documents/upload/digitalpreservationpolicy1.0.pdf>
(accessed Feb 2014)

*"Data refreshment is an ongoing process. It is undertaken regularly (minimally on a weekly basis) during the already noted synchronisation of locally held data to an off site data repository within the UKDA. This one way synchronisation compares checksum values at source and destination to detect change and acts accordingly."*UK Archaeology Data Service preservation Policy

<http://archaeologydataservice.ac.uk/attach/preservation/PreservationPolicyV1.3.1.pdf>
(accessed Feb 2014)

See also the STFC case study where these issues are discussed.

Blog from the The Signal discussing fixity:
<http://blogs.loc.gov/digitalpreservation/2014/02/check-yourself-how-and-when-to-check-fixity/> (accessed Feb 2014)

15 Bibliography

15.1 General

- [1] Eld Maj-Britt Olmütz Zierau, (2012) "A holistic approach to bit preservation", Library Hi Tech, Vol. 30 Iss: 3, pp.472 – 489. <http://dx.doi.org/10.1108/07378831211266618>
- [2] National Digital Stewardship Alliance: Preservation Levels
<http://www.digitalpreservation.gov/ndsa/activities/levels.html>

15.2 Policy, Management and Risks

- [3] *L Bairavasundaram, G Goodson, B Schroeder, A Arpaci-Dusseau, R Arpaci-Dusseau*. An Analysis of Data Corruption in the Storage Stack. Proceedings of the 6th USENIX Conference on File and Storage Technologies (FAST '08) Feb 2008, San Jose California, USA.
- [4] *C. Becker, J. Barateiro, G. Antunes, J. Borbinha, R. Vieira*. On the relevance of Enterprise Architecture and IT Governance for Digital Preservation. In: Electronic Government, 332-344. (2011) http://publik.tuwien.ac.at/files/PubDat_203363.pdf
- [5] *C. Becker, G. Antunes, J. Barateiro, and R. Vieira*. Control objectives for dp: Digital preservation as an integrated part of it governance. In Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology (ASIST), New Orleans, Louisiana, US, October 2011.
- [6] COBIT website <http://www.isaca.org/COBIT/Pages/default.aspx>
- [7] Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics. DPHEP-2012-001 May 2012. [arXiv:1205.4667](https://arxiv.org/abs/1205.4667) (accessed 28/11/2013)
- [8] DRAMBORA website and toolkit <http://www.repositoryaudit.eu/objectives/> Accessed Feb 2014
- [9] DPC Digital Preservation Risks
http://wiki.dpconline.org/index.php?title=Digital_preservation_risks (accessed Jan 2014)
- [10] *J Elerath*. Hard-disk Drives: The Good, the Bad and the Ugly. Communications of the ACM Vol 52. No.6 Pages 38-45. 10.1145/1516046.1516059 (accessed 27/11/2013)
- [11] ERPANET project <http://www.erpanet.org/index.php>
- [12] *Louise FAUDUET, Sébastien PEYRARD*. A DATA-FIRST PRESERVATION STRATEGY: DATA MANAGEMENT IN SPAR. iPRES 2010.
- [13] The Federation of American Scientists (FAS) works to provide science-based analysis of and solutions to protect against catastrophic threats to national and international security.
<https://www.fas.org/sgp/crs/homsec/RL32561.pdf>
- [14] T Gollins, Parsimonious preservation: preventing pointless processes! The National Archives
<http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation-in-practice.pdf>



- [15] *Haris Hamidovic*; An Introduction to Digital Records Management. In: ISACA Journal, vol. 6; 2010 <http://www.isaca.org/Journal/Past-Issues/2010/Volume-6/Pages/An-Introduction-to-Digital-Records-Management.aspx>
- [16] ITIL Website <http://www.ital-officialsite.com/>
- [17] *P Kelemen* Silent Corruptions. Presentation given at Linux Clusters for SuperComputing workshop (LCSC) , 2007 Sweden www.nsc.liu.se/lcsc2007/presentations/LCSC_2007-kelemen.pdf (accessed 28/11/2013)
- [18] *Steve Knight*. Developing a Digital Preservation Programme at a National Library. In: Proceedings of the 1st International Digital Preservation Interoperability Framework Symposium. ACM New York, NY (2010) <http://dx.doi.org/10.1145/2039263.2039265>
- [19] *B. Lavoie, P. Caplan, S. Vermaaten*. Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment. D-Lib Magazine. Vol.18; nr. 9/10; 2012. <http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html>
- [20] *RW Moore, J Jala, R Chadducj*. Mitigating Risk of Data Loss in Preservation Environments. Proceedings of the 22ms IEEE/13th NASA Goddard Conference on Mass Storage Systems and Technologies (MSST 2005)
- [21] *Andreas Rauber*; Digital Preservation in Data-Driven Science: On the Importance of Process Capture, Preservation and Validation. Theory and Practice of Digital Libraries (TPDL) 2012 http://timbusproject.net/component/docman/doc_download/84-digital-preservation-in-data-driven-science
- [22] *David S.H. Rosenthal*. Bit Preservation: A Solved Problem? In: The International Journal of Digital Curation; Issue 1, Volume 5 (2010); p. 134-148 <http://www.ijdc.net/index.php/ijdc/article/view/151/224>
- [23] *DSH Rosenthal, T Robertson, T Lipkins, V Reich*. Requirement for Digital Preservation Systems: A Bottom-Up Approach. arXiv:cs/009018v2 6 Sep 2005
- [24] *DSH Rosenthal*; Keeping Bits Safe: How Hard Can It Be? Communications of the ACM, Vol. 53 No. 11, Pages 47-55 10.1145/1839676.1839692
- [25] *Saideep Raj, Jack Sepple and Leslie Willcocks*; IT governance: Spinning into control. In: Outlook, The journal of high-performance business. Issue nr. 1, February 2013. <http://www.accenture.com/us-en/outlook/Pages/outlook-journal-2013-information-technology-governance-spinning-into-control.aspx>
- [26] SCAPE wiki which collect together links to the published preservation policies. <http://wiki.opf-labs.org/display/SP/Published+Preservation+Policies>
- [27] *B Schroeder, G Gibson*. Disk Failures in the real world: What does an MTTF of 1,000,000 hours mean to you? Proceedings of 5th USENIX Conference on File and Storage Technologies (FAST '07)
- [28] *Greet Volders*. How to use CobiT to assess the security & reliability of Digital Preservation? Erpa WORKSHOP Antwerp, 14 - 16 April 2004 http://www.erpanet.org/events/2004/antwerpen/presentations/erpaWorkshop-Antwerpen_Volders.pdf http://www.powershow.com/view/92edd-M2M0N/How_to_use_CobiT_to_assess_the_security_powerpoint_ppt_presentation



15.3 Technical Approaches:

[29]PREMIS: <http://www.loc.gov/standards/premis/v2/premis-2-2.pdf>

Case study: [http://documents.el-una.org/923/2/ELUNA Digital Preservation at the Church Final 2012 3 7.pdf](http://documents.el-una.org/923/2/ELUNA_Digital_Preservation_at_the_Church_Final_2012_3_7.pdf)

[30]A *Position Paper on Data Sovereignty: The Importance of Geolocating Data in the Cloud*. Zachary N. J. Peterson, Mark Gondree, and Robert Beverly. 3rd USENIX Workshop on Hot Topics in Cloud Computing. HotCloud '11.

[31]*How to Tell if Your Cloud Files Are Vulnerable to Drive Crashes*. Kevin D. Bowers, Marten van Dijk, Ari Juels, Alina Oprea, Ronald L. Rivest. CCS'11, October 17–21, 2011, Chicago, Illinois, USA.

[32]*Provable Data Possession at Untrusted Stores*. Giuseppe Ateniese, Randal Burns, Reza Curtmola, Joseph Herring, and Lea Kissner. Carnegie Mellon University, Research Showcase. Department of Electrical and Computer Engineering. 1 January 2007.

[33]*Fair and Dynamic Proofs of Retrievability*. Qingji Zheng and Shouhuai Xu. CODASPY'11, February 21–23, 2011, San Antonio, Texas, USA.

[34]POSTER: *Robust Dynamic Remote Data Checking for Public Clouds*. Bo Chen and Reza Curtmola. CCS'12, October 16–18, 2012, Raleigh, North Carolina, USA.

[35]*Windows Azure Storage: A Highly Available Cloud Storage Service with Strong Consistency*. Brad Calder, Ju Wang, Aaron Ogus, Niranjana Nilakantan, Arild Skjolsvold, Sam McKelvie, Yikang Xu, Shashwat Srivastava, Jiesheng Wu, Huseyin Simitci, Jaidev Haridas, Chakravarthy Uddaraju, Hemal Khatrri, Andrew Edwards, Vaman Bedekar, Shane Mainali, Rafay Abbasi, Arpit Agarwal, Mian Fahim ul Haq, Muhammad Ikram ul Haq, Deepali Bhardwaj, Sowmya Dayanand, Anitha Adusumilli, Marvin McNett, Sriram Sankaran, Kavitha Manivannan, and Leonidas Rigas. SOSP '11, October 23-26, 2011, Cascais, Portugal.

[36]*Towards Self-Repairing Replication-Based Storage Systems Using Untrusted Clouds*. Bo Chen and Reza Curtmola. CODASPY'13, February 18–20, 2013, San Antonio, Texas, USA.

[37]HAIL: A High-Availability and Integrity Layer for Cloud Storage. Kevin D. Bowers, Ari Juels, Alina Oprea. CCS'09, November 9–13, 2009, Chicago, Illinois, USA.

[38]*High Availability in DHTs: Erasure Coding vs. Replication*. Rodrigo Rodrigues and Barbara Liskov

[39]In M. Castro and R. van Renesse (Eds.): IPTPS 2005, LNCS 3640, pp. 226–239, 2005. *Building Confidence in the Cloud: A Proposal for Industry and Government Action for Europe to Reap the Benefits of Cloud Computing*. Microsoft.

[40]Centre for the International Study of Contemporary Records and Archives - <http://www.ciscra.org/>

[41]Not trusting cloud storage - <http://blog.dshr.org/2013/06/not-trusting-cloud-storage.html>

[42]Is cloud storage the answer to preservation? - http://www.researchinformation.info/news/news_story.php?news_id=1120

[43]Preservation in the Cloud - http://www.lockss.org/locksswp/wp-content/uploads/2011/12/3-3_Rosenthal-provider_cloud_storage1.pdf

[44]Digital Preservation Cloud Services for Libraries and Archives - <http://www.slideshare.net/qnguyen/digital-preservation-cloud-services-for-libraries-and-archives>



[45] Report on Digital Preservation and Cloud Services:

http://www.mnhs.org/preserve/records/docs_pdfs/Instrumental_MHSReportFinal_Public_v2.pdf

15.4 Case Studies

[46] APARSEN storage solutions paper D23.1 http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D23_1-01-1_0.pdf

[47] <http://home.web.cern.ch/about/computing>