




Research datasets executable workflows for experimental execution

Authors

Simon Lambert (Science and Technology Facilities Council)

April 2012

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 



Executive Summary

For the research datasets testbed in SCAPE, the STFC team has chosen data relating to the ISIS facility, a large neutron and muon source located at the Rutherford Appleton Laboratory. A number of challenges arise from long-term preservation of the scientific data obtained from this instrument, and these have been captured in a number of scenarios. An experimental workflow has been developed using the Taverna workflow management system, which represents some operations connected with the ingest of data. Some sample datasets have been made available for purposes of testing, evaluation and benchmarking.

A two-phase approach to development of the testbed is proposed, in which the first phase (the simple ingest-related scenario) is followed by a more advanced development of workflows and preservation actions concerned with preserving the research data in context—capturing and enhancing representation information to allow continued understanding and reuse of the datasets.

Table of Contents

Deliverable	i
Executive Summary	iii
1 Introduction	1
2 The experimental dataset	3
3 Research datasets testbed scenarios	8
3.1 Overview of the scenarios	8
3.2 Scenario RDST1: Scientific-data ingest-related scenario	9
3.2.1 Issues	9
3.2.2 Solutions	11
3.3 Scenario RDST2: Format migration of (raw) scientific datasets	12
3.3.1 Issues	12
3.3.2 Solutions	13
3.4 Scenario RDST3: Maintaining understandability and usability of raw data through external resources	13
3.4.1 Issues	13
3.4.2 Solutions	14
3.5 Scenario RDST4: Preserving the value of raw data and verifiability of processed datasets forming part of a scientific workflow	15
3.5.1 Issues	15
3.5.2 Solutions	16
4 Experimental workflow development	17
5 Conclusion and next steps	20

1 Introduction

The three testbeds play an important role in the SCAPE project, providing real-life opportunities to show the applicability and validation of the project's results in three diverse domains. The benefit is not only for the project, of course, but for the testbed domains themselves, producing "a reliable, robust integrated preservation system prototype" in each case, which can form the basis for further development and wider deployment beyond the project's lifetime.

The three parallel testbed work packages are concerned with the definition and set-up of the environment to which the SCAPE preservation system will be applied to achieve the preservation goals of the environment. The three work packages adopt a common structure for their plans, divided into twelve tasks. Some of these tasks overlap, but they follow a clear evolution starting with experimental and progressing to large-scale services. The aim of the experimental developments is to demonstrate the applicability of the SCAPE approach, while the large-scale focusses of course on scalability of solutions.

There are a number of important differences between the research datasets testbed¹ and the other testbeds in web content and large-scale digital repositories. The other testbeds have a strong focus on formats, particularly characterisation, migration and quality assurance, whereas the research datasets testbed has a focus on semantics. The aim of the preservation activity is to permit the continued understanding and use of datasets gathered in the past for purposes that might not necessarily have been expected at the time. That requires augmentation of the data with such information as is necessary to allow future users (or users from other scientific disciplines) to correctly reprocess the data for particular purposes. A simple example is the units of measurement of a set of numerical values.

Science data is part of a complex process and lifecycle that is of direct relevance to its preservation. The scientific dataset is associated with the software used to analyse it and the papers that were written from it, and these all provide valuable supplementary information to interpret the data itself.

Importantly for the phasing of the testbed development, special preservation action services for science data are being developed within SCAPE, rather than employing or extending existing tools, particularly for more advanced preservation actions. These build on the work of previous projects such as CASPAR, but the dependency on such developments within SCAPE—as opposed to the ready availability of say image format migration tools—has implications for the timeline of development of the research datasets testbed.

In short, as the Description of Work says, the research datasets testbed "will provide a validation of the SCAPE platform's ability to handle a diversity of preservation services." The main aspect of scalability being handled is the inherent complexity of the scientific research lifecycle—though the amounts of data are substantial.

The work package on the research datasets testbed (TB.WP.3) links to a number of other work packages, but especially:

¹ The term "science data" is also used in this document interchangeably with "research data" to reflect the specific testbed provided by STFC, without intending to exclude other varieties of research data that might not be considered scientific.

- **PC.WP.2 Action services components.** This work package will produce the preservation action services that will be deployed within the testbed, will be invoked by the preservation workflows and will perform actions on datasets as determined by the planning and watch components. The work package in fact has two tasks dedicated to the needs of the research datasets testbed: one on migration of research datasets, and one on the context and linking of datasets.
- **TB.WP.4 Evaluation of results.** This work package will provide the project with a mechanism for monitoring progress towards achieving project-level objectives by using the experience of the testbed implementation teams. The testbeds will provide input to defining evaluation procedures, and will apply those procedures to their own systems.

The Description of Work, describes the present deliverable thus:

Experimental scientific research data preservation workflows with documentation. These workflows are executable on the representative dataset and implement the first iteration of preservation scenarios, as well as providing input to the testbeds evaluation work package.

Several tasks have been accomplished to reach the state that the deliverable represents:

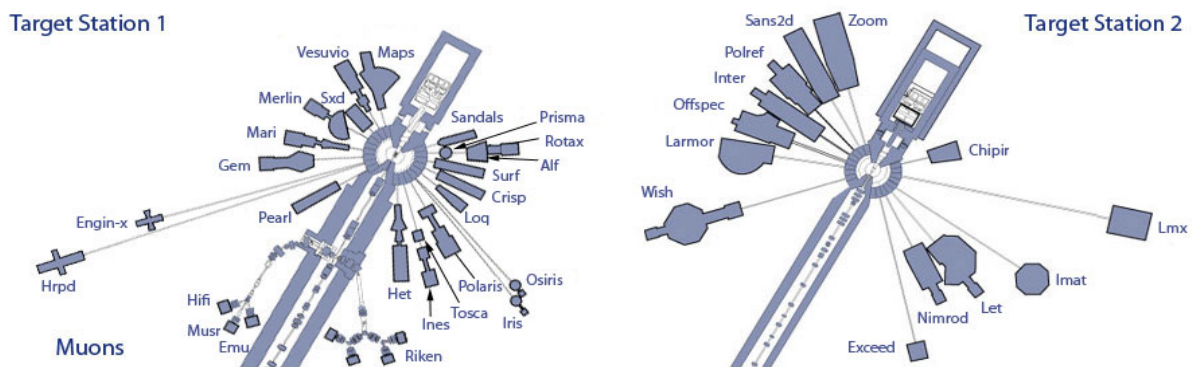
- **Task 1:** Define, capture and provide representative data set (M1–M9)
- **Task 2:** Develop first iteration of Preservation Scenarios including definition of requirements (M6–M9)
- **Task 3:** Design executable experimental preservation workflows (M10–M15)

The remainder of this document will lead up to the description of the workflows by way of introducing the experimental dataset that has been selected by STFC and describing the scenarios that are built around it. These scenarios collect together preservation issues and solutions related to the datasets, and have a key role to play in SCAPE because they provide a central focus for organising, prioritising and assessing the technical work of the project.

The single experimental workflow for STFC's research datasets testbed is then described. This workflow has been developed in discussion with the British Library (BL), the other partner in the corresponding Task 3. However, it does not include any datasets from the BL, as the intention is to focus (at least initially) on data held only by STFC, for simplicity. Thus the experimental workflow only represents the preservation actions associated with the STFC's research datasets.. In general, the aim of this task is to show a workflow implemented in Taverna invoking components of relevance to the testbed and applied to a sample dataset, including the preservation action services that it invokes. As hinted at above and explained in the discussion of scenarios below, the decision to develop just one experimental workflow reflects the different starting point of this testbed compared with the others. The initial aim has been to demonstrate applicability of the SCAPE approach, while in parallel working on necessary developments for additional scenarios and workflows.

2 The experimental dataset

The primary focus of STFC’s testbed will be data from the ISIS facility. ISIS is a world leading centre for research in the physical and life sciences. It is one of the world’s leading sources of neutrons and muons, used for exploring the structure of matter. The heart of ISIS is an accelerator that produces intense pulses of protons 50 times per second. Muons are produced when the proton beam passes through a carbon target. The protons then go on to collide with a tungsten target and produce neutron pulses. The neutron and muon beams produced at ISIS are used in research areas ranging from clean energy and the environment to pharmaceuticals, nanotechnology and IT.



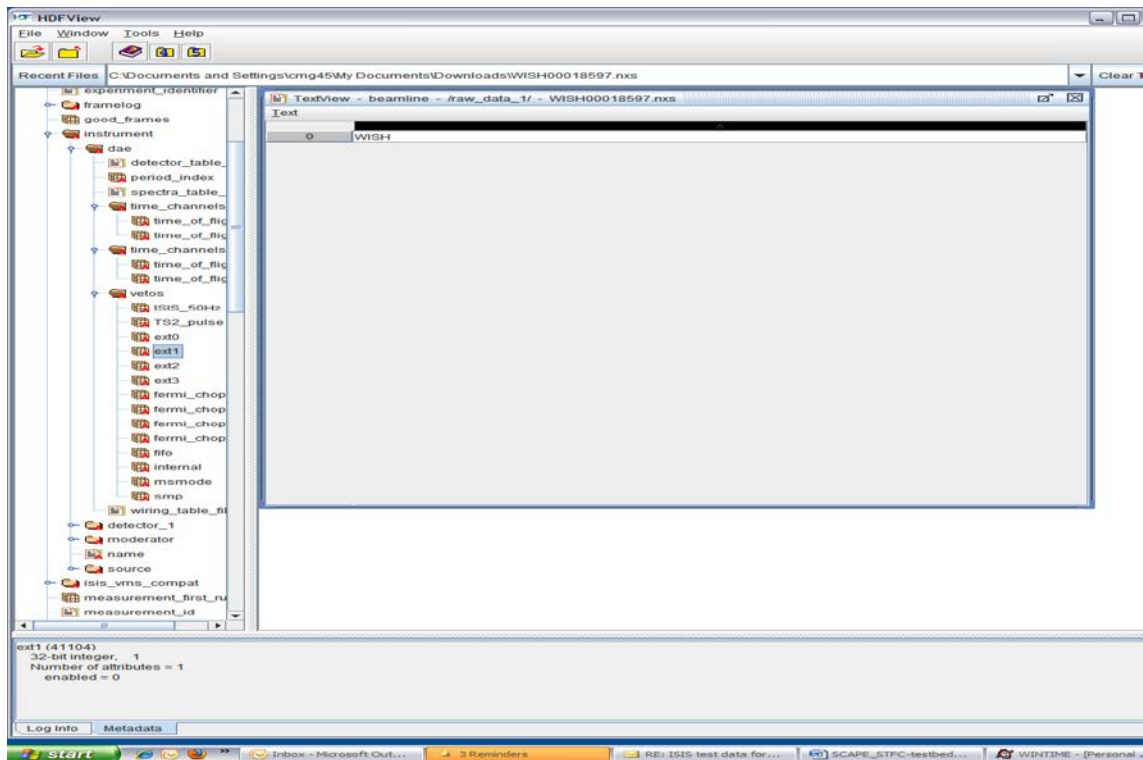
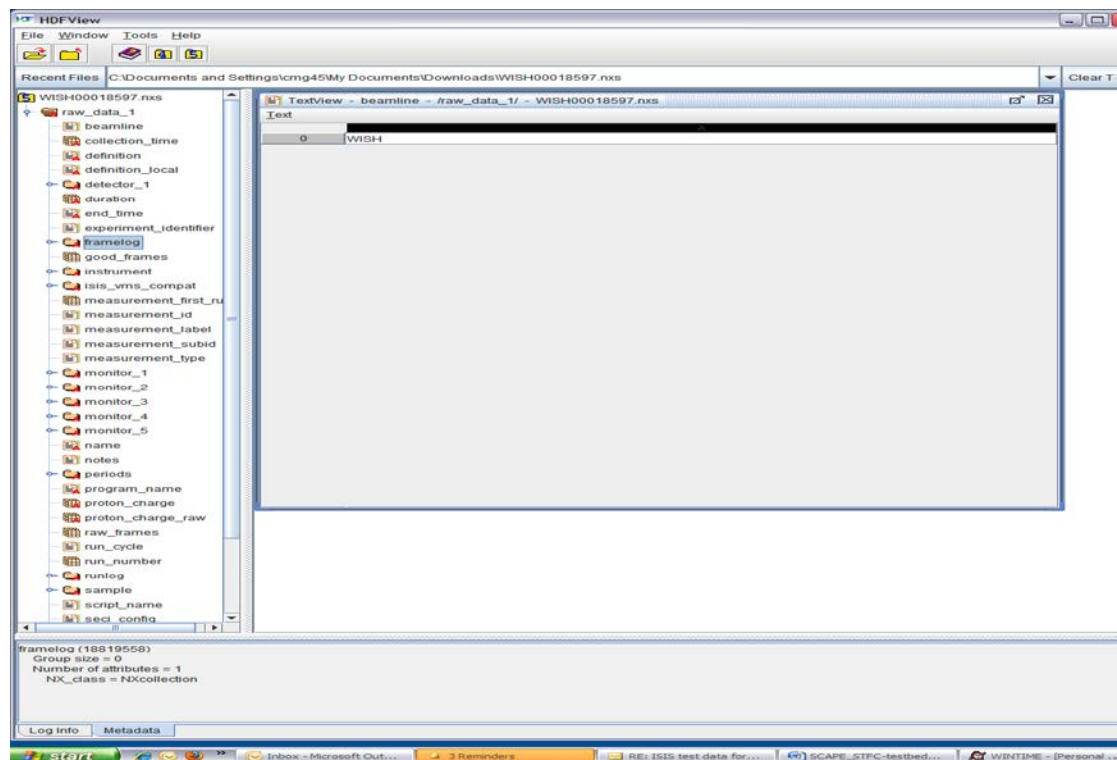
ISIS has two target stations, each with a number of instruments for different experimental techniques. Techniques include muon spectroscopy, neutron diffraction, neutron spectroscopy, reflectometry and small angle scattering.

When a proposal for an experiment is accepted the Principal Investigator and their team will be supplied with some beamtime. All experiments will involve a sample to be analysed which is put into the instrument before neutrons or muons are fired at it. An experiment may comprise many ‘runs’ and each run may produce one or more files. Two types of data output are produced: RAW files which include one or more log files, and/or NeXus files which do not have a log file, but have the logging information within the file itself. Metadata about the files is generated automatically and is kept in ICAT. NeXus is a common data format for neutron, x-ray, and muon science, and is being developed as an international standard². NeXus uses an underlying HDF5 representation³ and the file contents can be viewed using generic HDF5 utilities such as HDFView or analysis programs such as Open GENIE or Mantid.

The screenshots below show a typical NeXus file unpacked in an HDF5 viewer. The second screenshot illustrates the hierarchical structure.

² www.nexusformat.org

³ www.hdfgroup.org



In this process both the instruments and the associated computer systems are generated and maintained in-house, so although the RAW files are not an international standard, they have been consistently produced over the twenty five years of ISIS production running.

There is a standard process around the use of the ISIS facility by researchers. This is typical of all such facilities, whether neutron or photon-based. It can be summed up in the following steps:

- Researcher uses ISIS proposal system to ask for beam time for a specific experiment and sample. This includes linking to publications to prove the scientific worth of the experiment.
- Proposal considered at panels and either approved or rejected.
- Approved proposals will be scheduled.
- Experiment day: calibrations, configurations and experiment. Data collected in RAW format for old instrument, some newer ones collect NeXus files
- NeXus/RAW file together with configurations are catalogued into ICAT, together with information on proposal, dates, principal investigators, file size and parameters.
- Real experimental data sits in ISIS on a cluster of Windows servers, and local users can mount the file system to get to the data. External people either take their data away on physical storage or can retrieve from the web.
- The Mantid software enables analysis and at present loads the file from the shared disk. Future plans are to move away from this to use an integrated iCAT client which will query iCAT for the file's URL which will be used to retrieve data using a web service in front of the current file system.

A central part of the information system support for ISIS is the ICAT metadata catalogue. ICAT is an open source metadata management system designed for large facilities. It enables access to data, search, annotation and sharing.

The sample data set provided by ISIS for SCAPE has the data from four different instruments together with the XML ICAT metadata records for most of them. Data from the first three instruments were generated in 2007 and for the final instrument were generated in 2011 but are calibration data. The sample data is intended to be a representative sample of file types and sizes generated by the ISIS facility and its associated instruments.

The instruments selected are the following:

ALF: Excitations Alignment Facility for Single Crystals. ALF is typical of many RAW file instruments at ISIS as it produces many associated small log files. Depending on the actual instrument, these files can be generated at a rate between one run every 30 seconds to one run every 30 minutes and the file size varies accordingly. Typically the RAW file is around 5 Mb in size, and is associated with around a dozen supplementary log files (textual).

MAPS: high energy magnetic excitations in single crystals with varying energy resolution. MAPS is typical of some of the more recent instruments which produce fewer, but larger RAW files + log files.

EMU: Muon spectrometer. EMU is typical of muon instruments which produce all files in the NeXus format.

LET: Cold neutron multi-chopper spectrometer. LET is an ISIS Target Station 2 instrument which produces large NeXus files, roughly of the order of 1–2 Gb in size.

While the sample data is representative of the STFC's research dataset, in terms of size, however, it is only 6.3 Gb in total – this is only about 200+ runs. This is, in fact, just a fraction of the entire

dataset that is in need of long-term preservation. As illustrated in the case of LET instrument above, a single run of an experiment could produce data files (in NeXus format) that could be up to 1.2 Gb each in size, and considering, each experiment comprises tens or hundreds of runs, this could yield substantially large volumes of data from that experiment. For some instruments, the data could be in the form of one (big) file; but, for others, the data could be in a large number of small data files. From this perspective, it is, therefore, the sheer volume, not the size of the individual files of the STFC's research dataset that presents a considerable scalability challenge, and fittingly, data volume is one of the key areas of concern in SCAPE.

The other three areas in which SCAPE considers scalability are: size, complexity and diversity. As mentioned before, for the STFC research datasets, the size of individual files is not significantly large, hence it is not expected to pose any considerable scalability challenge. On the contrary, the data files can be highly complex, and considerably diverse, particularly in terms of file types; for some experiments, there are about ten different types of files generated for one run. Though most of them are in text or ASCII format, their contents may differ both in terms of structure and semantics, which effectively means that the dealing of these files can vary from instrument to instrument, and from experiment to experiment, making it potentially a complex procedure. These attributes of the STFC datasets would likely affect the operations needed for their effective preservation, and also the associated scalability requirements.

In addition, some examples of matching RAW and NeXus files have been made available—these are generated simultaneously by some instruments and may be useful for investigating format migration in due course.

The datasets are subject to restrictions on access, outlined at <http://www.isis.stfc.ac.uk/user-office/data-policy11204.html>. Users of the data in the SCAPE project are required to accept conditions of use before they may gain access to the data. Specifically, the following conditions must be accepted.

**Science and Technology Facilities Council (STFC)
ISIS Data Access Agreement for SCAPE Partners**

BEFORE YOU CLICK ON THE ACCEPT BUTTON AT THE END OF THIS DOCUMENT, CAREFULLY READ ALL THE TERMS AND CONDITIONS OF THIS AGREEMENT. BY CLICKING ON THE ACCEPT BUTTON, YOU ARE CONSENTING TO BE BOUND BY AND ARE BECOMING A PARTY TO THIS AGREEMENT. IF YOU DO NOT AGREE TO ALL OF THE TERMS OF THIS AGREEMENT, CLICK THE "DO NOT ACCEPT" BUTTON AND DO NOT DOWNLOAD THIS INTELLECTUAL PROPERTY.

1 Definitions

1.1 'raw data' are data collected from experiments performed on ISIS instruments. This definition includes data that are created automatically or manually by Facility-specific software and/or ISIS staff expertise in order to facilitate subsequent analysis of the experimental data, unless otherwise agreed.

1.2 'metadata' is information pertaining to data collected from experiments performed on ISIS instruments, including (but not limited to) the context of the experiment, the experimental team (in accordance with the Data Protection Act), experimental conditions and other logistical information.

1.3 the term 'on-line catalogue' pertains to a computer database of metadata containing links to raw data files, that can be accessed by a variety of methods, including (but not limited to) www-based browsers.

1.4. the term 'public domain' means belonging to the community at large, unprotected by copyright or patent and subject to appropriation by anyone.

2. Access to the Raw data and associated metadata

2.1 All raw data and the associated metadata obtained as a result of free (non-commercial) access to ISIS, reside in the public domain, with ISIS acting as the custodian.

2.2 Access to raw data and the associated metadata obtained from an experiment is restricted to the experimental team for a period of three years after the end of the experiment. Thereafter, it will become publicly accessible.

2.3 Data provided to the SCAPE project is publicly accessible.

2.4 To gain access to the data, you must register with the STFC team and agree that:

2.4.1 The data will be used only for the purposes of the SCAPE project

2.4.2 The data will not be passed on to other parties, within or without the SCAPE project

2.4.3 At the end of the project you undertake to destroy any personal copies of the data you may hold.

A final point to make concerns the representativeness of the datasets for science data in general. First, there is a high degree of commonality across facilities-based science—that is, experimental science based on the exposure of samples of materials at central facilities such as ISIS. There are many such facilities, utilising not only neutrons and muons like ISIS, but also photons, whether synchrotron radiation or lasers. The PaN-data consortium⁴, of which STFC is a member, brings together thirteen major world class European research infrastructures to create a fully integrated, pan-European, information infrastructure supporting the scientific process. Its interests include preservation of data; a recent FP7-funded project, PaN-data ODI, has preservation as one of its focusses.

More generally, the same problems of preserving semantics and the context of research arise in all scientific disciplines that gather data. Even though the data formats might differ, the challenges remain the same. Thus we are confident that the results of SCAPE applied to the research datasets testbed will be of wide applicability beyond ISIS and indeed beyond facilities science.

⁴ <http://pan-data.eu>

3 Research datasets testbed scenarios

3.1 Overview of the scenarios

The scenarios are intended to capture clusters of preservation concerns for each testbed that will be addressed in the SCAPE project. In the Description of Work, the assumption was made that the initial focus would be on migration of formats, followed later by more complex scenarios addressing the inherent complexity of the scientific lifecycle and its implications for preservation (for example, linking of outputs at different stages of the lifecycle to provide valuable supplementary information associated with a dataset).

There has however been a shift of emphasis concerning migration from RAW format to NeXus. This can no longer be regarded as the first priority for the testbed. The reasons are that this is not perceived as a preservation need by the ISIS facility at the moment, since the analysis tools are currently capable of coping with both formats—in other words, there is backward compatibility. Instead the initial focus of the testbed will be on other relatively simple and general preservation actions such as ingest, fixity checking and format validation.

Therefore the two phases can now be summed up as follows.

Phase 1. Although migration from RAW to NeXus is not of such immediate concern, in the same spirit, it is desirable to start with a relatively straightforward preservation scenario. This will be a simple workflow performing ingest of files and some checks on the files.

Phase 2. The second phase will have the more ambitious goal of “preserving datasets in context” and will employ preservation actions based on preservation network models⁵. Here there are four broad types of preservation strategy: risk acceptance and monitoring; capture of software and extension through the stack; description; and migration (transformation).

The following scenarios have been uploaded, in common with the other testbeds, to the OPF wiki⁶ which is being used as a working space by the SCAPE project. The scenarios all have a common structure, comprising a triple of dataset, preservation issue and possible solution. Since the datasets have already been described above, they will not be repeated in the individual scenario summaries. The preservation issues represent particular problems arising in long-term preservation of the datasets, which may be addressed by the proposed solutions. It should be noted that individual preservation issues and solutions are in some cases repeated from one scenario to another.

⁵ For an introduction to preservation network models, see for example E. Conway et al., “Managing risks in the preservation of research data with preservation networks”, Proc. 7th International Digital Curation Conference (IDCC2011).

⁶ STFC Scientific Datasets - <http://wiki.opf-labs.org/display/SP/STFC+Scientific+Datasets>

3.2 Scenario RDST1: Scientific-data ingest-related scenario

3.2.1 Issues

Title	IS29 Syntactic checking and validation of NeXus data from instrument
Detailed description	In order to ensure that the data to be preserved is of adequate quality, there is a need for structural/syntactical verification and characterisation upon ingesting data into repository
Scalability Challenge	The file size of these varies significantly: from 10s of MBs to 2GBs.
Possible Solution approaches	Use of existing characterisation tools, such as JHOVE with appropriate extension or plug-in for NeXus data format.
Context	<i>Details of the institutional context to the Issue. (May be expanded at a later date)</i>
Datasets	Nexus data files
Solutions	SO20

Title	IS30 Fixity capturing
Detailed description	Need for capturing the fixity information (e.g. checksum) to ensure continuous integrity of datasets to be preserved
Scalability Challenge	<i>What requirements are placed on the solution in terms of the SCAPE scales of scalability: content size, volume of content, complexity of content</i>
Possible Solution approaches	Use of commonly used fixity calculation algorithms (e.g. MD5, SHA-1)
Datasets	Nexus files
Solutions	

Title	IS31 Semantic checking of NeXus data from instrument
Detailed description	In order to ensure that the data to be preserved is of adequate quality , the contents of NeXus data files would need to be validated for their correctness against a given data model. Each data model is specified in a NeXus Definition Language (NXDL) file and contains assertions that define the expected content of a NeXus file. For example, a data model could define a metadata element (key-value pair) called “Integral” to represent the total integral monitor counts for grazing incidence small angle diffractometer GISAS for either x-ray or neutrons. In this scenario, the data type of the metadata element “Integral” would be an integer. For a NeXus data file conforming to this data model, it would be necessary to validate the value(s) assigned to “Integral” to ensure it is of appropriate data type.
Possible Solution approaches	Use of the NeXus validation toolkit - developed and used by the NeXus community - as part of the preservation ingest workflow.
Datasets	nexus data files
Solutions	SO21

Title	IS32 Basic migration of RAW to NeXus data
	See below under RDST2

Title	IS33 Enhanced migration of RAW to NeXus data
	See below under RDST2

3.2.2 Solutions

Title	SO20 Use JHOVE with appropriate extension or plug-in for NeXus data format
Detailed description	Use of existing characterisation tools, such as JHOVE with appropriate extension or plug-in for NeXus data format.
Corresponding Issue(s)	IS29
Evaluation	

Title	SO21 Use of the NeXus validation toolkit - developed and used by the NeXus community - as part of the preservation ingest workflow.
Detailed description	Use of the NeXus validation toolkit - developed and used by the NeXus community - as part of the preservation ingest workflow.
Corresponding Issue(s)	S31
Evaluation	

Title	SO 22 Developing a Raw-to-NeXus migration tool
	See below under RDST2

3.3 Scenario RDST2: Format migration of (raw) scientific datasets

3.3.1 Issues

Title	IS32 Basic Migration of RAW to NeXus data
Detailed description	Obsolescence of RAW files and preference for NeXus as standard
Possible Solution approaches	No suitable solutions exist at present. Hence, a suitable RAW-to-NeXus migration mechanism would need to be developed.
Datasets	nexus data files
Solutions	SO22

Title	IS33 Enhanced migration of RAW to NeXus data
Detailed description	Desire to enhance the value of the dataset with additional information about an experiment that is not present in the basic data file, so as to enrich the dataset with representation information.
Scalability Challenge	There are two scalability challenges: (1) enhanced migration needs to be applied to substantially large volumes of data and metadata files (2) with differing structures and semantics. Hence, automation is required while ensuring both semantic and structural validity of the data.
Possible Solution approaches	push the additional information into metadata fields of a nexus file; this requires (semi)-automated mechanism with appropriate semantic and structural validation methods.
Datasets	nexus data files , ICAT catalogue data
Solutions	SO22

3.3.2 Solutions

Title	SO 22 Developing a Raw-to-NeXus migration tool
Detailed description	No suitable solutions exist at present. Hence, a suitable RAW-to-NeXus migration mechanism would need to be developed.
Corresponding Issue(s)	IS32, IS33
Evaluation	

3.4 Scenario RDST3: Maintaining understandability and usability of raw data through external resources

3.4.1 Issues

Title	IS34 ISIS instrument website no longer applicable or available
Detailed description	Relevant description of the functioning of the instrument no longer available to help interpret data files
Scalability Challenge	
Possible Solution approaches	Watching the website and keeping local copies of key information content will help to avoid this situation.
Datasets	Nexus data files , ICAT catalogue data
Solutions	SO23

Title	IS35 Mantid website and/or software no longer applicable or available
Detailed description	<p>1. Relevant description of the analysis software and the algorithms it implement no longer available. Impact: loss of the ability to use the Mantid software to properly process the raw data because, for example, the up-to-date manual is loss.</p> <p>2. Data can no longer be analysed - i.e. becomes unusable, loss the capability of verifying the processed data</p>
Scalability challenge	
Possible Solution approaches	<p>1) This is a case where reasoning with a Preservation Network Model can help to find solutions. The dependencies indicate that either a local archive of the website or access to the UK Web Archiving Consortium archive of the STFC website would be able to substitute for the live site. 2) This is a case where reasoning with a Preservation Network Model can help to find solutions. The dependencies indicate that solutions could be provided by software on other platforms; specifications to reconstruct software.</p>
Datasets	Nexus data files
Solutions	SO23

3.4.2 Solutions

Title	SO23 Pushing additional metadata into NeXus metadata fields
Detailed description	<p>NeXus is an international standard format for neutron and synchrotron data files. It provides a whole range of pre-defined metadata fields (e.g. facility name, sample name). It also offers the flexibility of defining custom metadata fields. Some STFC ISIS instruments generate NeXus files from data acquisition systems, There is a well established automated process of doing so. In other projects, we define non-standard metadata fields for NeXus and push additional experimental information into these fields in NeXus files. Therefore, it is possible to push additional preservation-related metadata into NeXus metadata fields in SCAPE.</p>
Corresponding Issue(s)	IS34, IS35
Evaluation	SP.

3.5 Scenario RDST4: Preserving the value of raw data and verifiability of processed datasets forming part of a scientific workflow

3.5.1 Issues

Title	IS36 Examine the long term value of the preserved datasets
Detailed description	A large collection of raw data files are being collected into STFC archive every year capturing the experimental data captured straight from a large number of scientific instruments. We are trialling a basic bit-level preservation system on the newly created files. There is limited understanding of the preserved value of these collections. For example, how useful are they (e.g. are they containing enough information for researchers other than the original investigators to interpret them?) We need an efficient approach to measure and examine the value of these collections so that the preservation cost can be justified and the benefits can be quantified.
Scalability Challenge	Every year, some facilities generate millions of raw data files per instrument and there are often 10s of instruments per facilities. In addition, the file sizes vary significantly from instrument to instrument. Some generate files in the order of GBs, some in the order of KBs (but with a large number of small files). So, any approach for solving this problem has to be scalable (e.g. in terms of file size and volume) as well as fully automated.
Possible Solution approaches	
Datasets	Nexus data files , ICAT catalogue data
Solutions	SO24

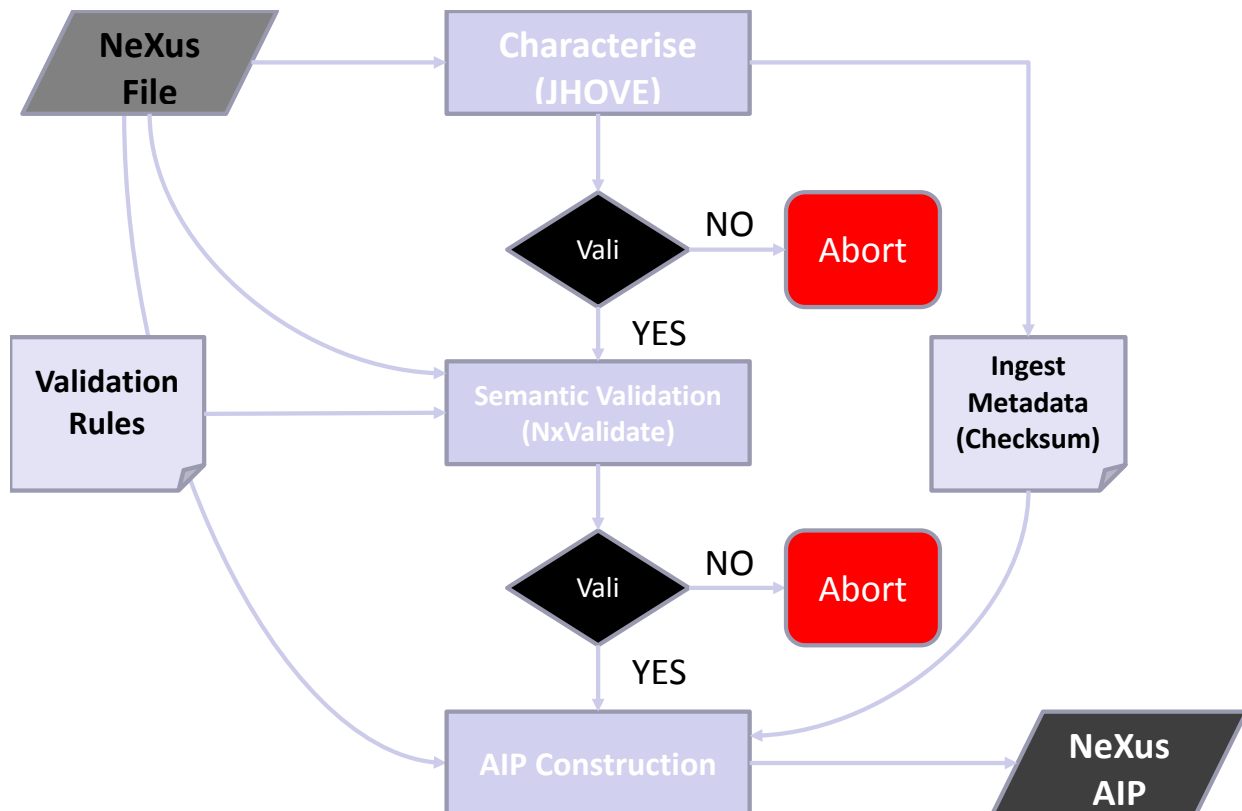
Title	IS37 Preserving the verifiability and provenance of processed datasets
Detailed description	Preserving the relationships between components of research objects is the challenge to tackle because it not only involves the components themselves but also the intrinsic relationship between them. We need to preserve the components but also the relationships between the components and allow the continuous evolution of such relationships to incorporate new components over time.
Scalability Challenge	Same as IS36
Possible Solution approaches	Use of Preservation Network Models to record “deep” dependencies and to allow for tracking over time.
Datasets	Nexus data files , ICAT data catalogue , workflow, software, and processed data
Solutions	SO24

3.5.2 Solutions

Title	SO24 Use Preservation Network Model to record "deep" dependencies and to allow tracking over time
Detailed description	Preservation Network Model (PNM) is a methodology to record "deep" dependencies between digital artefacts and to allow tracking over time (this is a solution yet to develop)
Corresponding Issue(s)	IS36, IS27
Evaluation	

4 Experimental workflow development

In the current reporting period, STFC has developed an experimental Taverna workflow that is representative of a data ingest scenario for STFC’s NeXus-based scientific datasets, corresponding to scenario RDST1. As illustrated in the diagram below, the workflow consists of the following preservation-related operations associated with a data ingest process.



Characterisation of NeXus files. The first operation in the workflow involves file format identification, structural validation and checksum calculation using an existing digital object characterisation tool, namely JHOVE2 (version 2.0.0⁷). It was also necessary to implement a specific plug-in/module for JHOVE to enable accurate identification of NeXus file format. In effect, this JHOVE plug-in/module for NeXus file format utilises existing Java-based utilities (developed by the NeXus user and developer communities) for checking structural validity of NeXus files.

Semantic Validation of NeXus files. In order to ensure that the data to be preserved is of adequate quality, an existing NeXus validation toolkit (developed and used by the NeXus community) is incorporated in the preservation ingest workflow as the second operation. This toolkit performs validation of the contents of NeXus data files for their correctness against a given data model. Each data model is specified in a NeXus Definition Language (NXDL) file and contains assertions that define the expected content of a NeXus file. For example, a data model could define a metadata element (key-value pair) called “Integral” to represent the total integral monitor counts for grazing incidence

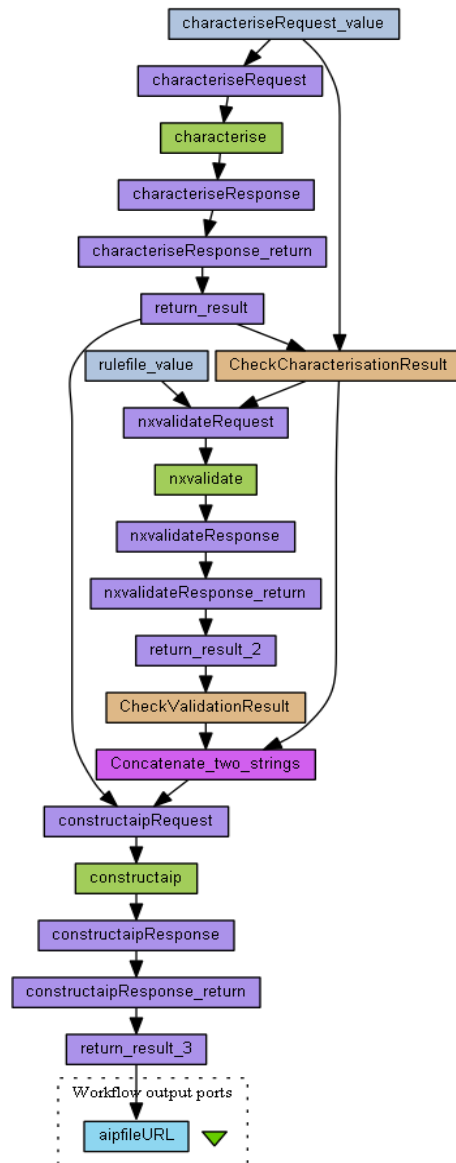
⁷ https://bitbucket.org/jhove2/main/wiki/JHOVE2-2.0.0_Download

small angle diffractometer GISAS for either x-ray or neutrons. In this scenario, the data type of the metadata element “Integral” would be an integer. For a NeXus data file conforming to this data model, it would be necessary to validate the value(s) assigned to “Integral” to ensure it is of appropriate data type.

Archival Information Package (AIP) Construction. The final operation of the experimental ingest workflow uses a bespoke Java-based tool to create an AIP that is essentially a zipped folder containing a syntactically and semantically valid NeXus file and the corresponding ingest-related metadata, such as checksum information.

Notably, all the three aforementioned tools/applications are command-line tools deployed as SOAP-based web services using the Toolwrapper developed by the SCAPE Testbed Subproject based on a similar tool developed by the IMPACT project.

The following diagram shows the workflow as developed in Taverna. The Taverna workbench and workflow management system are being utilised and enhanced in SCAPE for developing and executing preservation workflows.



In general, the Taverna experimental workflow worked as expected in demonstrating the underlying operations. For example, the JHOVE2 profile for NeXus file was able to correctly identify valid NeXus files while failing to recognise corrupt NeXus files or files masquerading as NeXus files. We also experimented with the "parallel service invocation" feature of Taverna in two ways: 1) processing multiple NeXus files through a single instance of the experimental workflow and 2) running multiple instances of the workflow in parallel to process multiple NeXus files. In both cases, the performance was much slower than that of a single workflow instance processing a single NeXus file as all the services were running on the same infrastructure. This level of performance is insufficient for the preservation of the STFC's research dataset, hence would need to be improved significantly for this Taverna-based approach to be effective. We anticipate that the Hadoop-based platform that is being developed by SCAPE would be able to address this issue by facilitating accelerated processing of Taverna workflows for large number of NeXus files.

5 Conclusion and next steps

This deliverable has reported on the experimental workflows developed for the research datasets testbed in SCAPE. Several steps were involved in achieving the experimental workflow in the Taverna workflow management system. A sample dataset was made available of data from the ISIS facility at STFC, representing a range of instruments at the facility and a variety of data formats. A number of scenarios were developed to characterise the goals for preservation in the context of SCAPE; the scenarios are formulated in terms of the datasets from which they arise, the preservation issues to be tackled, and approaches to solutions. A particular scenario was selected for the development of the experimental workflow.

In describing the workflow as experimental, a number of senses are intended. The first is that the workflow demonstrates that the SCAPE approach to the research datasets testbed is practicable, by producing a model, restricted in size and scope, that embodies key aspects of SCAPE and is relevant in useful in the testbed context. Second, the workflow is intended as a basis for experimentation in the context of the goals of SCAPE, principally that of scalability. Scalability has four facets: number of objects, size of objects, complexity of objects, and heterogeneity of collections. The research datasets testbed is primarily concerned with number of objects and especially complexity. In later stages of the project the complexity will be addressed directly; for now, detailed benchmarking of performance (some preliminary benchmarking was already conducted in the reporting period as outlined in Section 4) can take place on the experimental datasets and workflow.

Third, the experimental workflow is the basis for the larger-scale developments that follow it, as an experimental or prototype vehicle is for the production model that is based on it and takes advantage of the lessons learnt. The STFC team now has familiarity with Taverna and is in a good position to continue implementing more advanced workflows, building on the knowledge acquired in this experimental phase.

As well as the development of additional workflows, further work in the testbeds work package will involve a number of developments (sometimes in other WPs):

- Setting up a local instance of the SCAPE platform
- Development of preservation components to implement more advanced services that can be invoked from workflows
- Integration with policy-driven planning and watch
- Integration into the existing systems at STFC such as ICAT