

Project no. 269977

**APARSEN**  
**Alliance for Permanent Access to the Records of Science  
Network**

**Instrument:** Network of Excellence

**Thematic Priority:** ICT 6-4.1 – Digital Libraries and Digital Preservation

**D14.1 REPORT ON TESTING  
ENVIRONMENTS**

---

Document identifier:	APARSEN-REP-D14_1-01-1_4
Due Date:	2013-02-28
Submission Date:	2013-06-30
Work package:	WP14
Partners:	STFC, APA, CSC, DNB, KNAW-DANS, KB, FORTH, UNITN, SBA, UESSEX, CINES
WP Lead Partner:	Tessella
Document status	Released
URN	urn:nbn:de:101-2014051670

---

**Abstract:** This document provides a report in support of the work that has been undertaken within WP14 Common Testing Environments. The objective of this work package is to “collect together a set of environments that can be used to test the efficacy of tools and techniques for digital preservation, against changes in hardware, software, environment and knowledgebase of the Designated Communities, and design new ones if necessary”. This deliverable is a report which summarises the test environments which have been examined and the proposed common testing approach, and also provides a framework within which to evaluate the efficacy and applicability of proposed preservation tools and techniques, for example what types of digital objects the tool/technique is/is not useful for and what types of changes the tool/technique can guard against (changes in hardware, software, environment and knowledgebase of designated communities).

**Delivery Type**

**Author(s)** Ashley Hunter, Pauline Sinclair, Robert Sharpe (Tessella), David Giaretta (APA/STFC). Additional supporting input from Rene van Horik (KNAW-DANS), Sabine Schrimpf (DNB), John Lindström (LTU), Mark Guttenbrunner (SBA), Heikki Helin (CSC), Hervé L'Hours (UK Data Archive), Yannis Tzitzikas (FORTH)

**Approval****Summary****Keyword List**

**Availability** ☒ PUBLIC

**Document Status Sheet**

Issue	Date	Comment	Author
0.1	11-May-2011	Wrote initial proposal for how the testbed work should be carried out.	Pauline Sinclair, Tessella
0_2	03-Jul-2011	Update to provide a view of the landscape we need to address	David Giaretta/APA/STFC
0_3	12 Jul-2011	Add further details of testing landscape	David Giaretta, APA/STFC
0_4	24 Jul-2011	Comments from Tessella plus responses from DG	Tessella and David Giaretta
0_5	09 Jan-2012	Reformatted to improve readability, and updated based on recent work	Ashley Hunter, Tessella
0_6	16 Jan 2012	Clarify a number of points including test environment vs preservation techniques	David Giaretta
0_7	18 Jan 2012	Collate changes from partners	Ashley Hunter, Tessella
0_8	23 Jan 2012	Final Modifications	Ashley Hunter, Tessella
1_0	24 Sep 2012 – 23 Jan 2013	Restructure based on comments from the 1 <sup>st</sup> Year review	Ashley Hunter, Tessella
1_1	31 Jan 2013	Minor corrections and additions	David Giaretta, APA
1_2	08 Feb 2013	Minor corrections and additions	René van Horik, DANS
1_3	27 Feb 2013	Final edits based on internal review	Ashley Hunter, Tessella
1_4	30 Jun 2013	Revisions following Year 2 Review	Ashley Hunter, Tessella and David Giaretta, APA

**Project information**

Project acronym:	<b>APARSEN</b>
Project full title:	<b>Alliance for Permanent Access to the Records of Science Network</b>
Proposal/Contract no.:	<b>269977</b>

---

**Project coordinator: Simon Lambert/David Giaretta**

Address:	STFC, Rutherford Appleton Laboratory Chilton, Didcot, Oxon OX11 0QX, UK
Phone:	+44 1235 446235
Fax:	+44 1235 446362
Mobile:	+44 (0) 7770326304
E-mail:	<a href="mailto:simon.lambert@stfc.ac.uk">simon.lambert@stfc.ac.uk</a> / <a href="mailto:david.giaretta@stfc.ac.uk">david.giaretta@stfc.ac.uk</a>

## CONTENT

<b>1</b>	<b>EXECUTIVE SUMMARY .....</b>	<b>6</b>
<b>2</b>	<b>INTRODUCTION .....</b>	<b>6</b>
2.1	DESCRIPTION OF WP14 OBJECTIVES .....	8
2.2	DESCRIPTION OF WP14 DELIVERABLE .....	8
<b>3</b>	<b>TEST ENVIRONMENTS AND TESTBEDS.....</b>	<b>9</b>
3.1	TEST ENVIRONMENTS DEVELOPED BY PREVIOUS PROJECTS .....	10
3.1.1	<i>Significant Properties based Test Environments (SPT).....</i>	<i>10</i>
3.1.1.1	Testbed Digitale Bewaring .....	11
3.1.1.2	DELOS Network of Excellence.....	11
3.1.1.3	Planets Testbed.....	12
3.1.1.4	DCC Test Methodology.....	12
3.1.2	<i>Designated Community based Test Environments (DCT) .....</i>	<i>13</i>
3.1.3	<i>Transformational Information Properties based Test Environments (TIPT) .....</i>	<i>14</i>
3.1.4	<i>Multi-Valent Test Environments (MVT).....</i>	<i>15</i>
3.1.5	<i>Commentary on previous test environment approaches.....</i>	<i>15</i>
3.2	APARSEN TEST ENVIRONMENT SYSTEMS .....	16
3.3	THREATS TO DIGITAL MATERIAL .....	17
<b>4</b>	<b>TEST ENVIRONMENTS EVALUATION FRAMEWORK .....</b>	<b>18</b>
4.1	DEVELOPING A COMMON APPROACH .....	18
4.2	IMPLEMENTATION STRATEGY .....	19
4.2.1	<i>Classification.....</i>	<i>19</i>
4.2.2	<i>Defining a Digital Preservation Landscape .....</i>	<i>20</i>
4.3	CAPTURING PRESERVATION SCENARIOS .....	21
4.3.1	<i>Preservation Scenario Template .....</i>	<i>21</i>
4.3.2	<i>Collection of Preservation Scenario data .....</i>	<i>22</i>
4.4	ANALYSIS OF PRESERVATION SCENARIOS .....	22
4.4.1	<i>A method for assessing the capabilities of Test Environments .....</i>	<i>23</i>
4.4.2	<i>Monitoring the evolution of a Test Environment over time .....</i>	<i>23</i>
4.4.3	<i>Assessing the capabilities of Tests Environments to preserve digital objects .....</i>	<i>26</i>
4.5	CLASSIFICATION SCHEMES.....	27
4.5.1	<i>Types of digitally encoded information object.....</i>	<i>28</i>
4.5.2	<i>Types of preservation techniques .....</i>	<i>29</i>
4.5.3	<i>Common Test Data Sets .....</i>	<i>30</i>
<b>5</b>	<b>TESTING COMMON TEST DATA SETS IN THE APARSEN TEST ENVIRONMENT SYSTEMS .....</b>	<b>31</b>
5.1	KOLIBRI (DNB) .....	31
5.2	PRESERVICA (TESSELLA).....	32
<b>6</b>	<b>CONCLUSIONS.....</b>	<b>36</b>

## 1 EXECUTIVE SUMMARY

This document details the work that has been undertaken within the WP14 Common Testing Environments work package of the APARSEN project. The objective of this work package is to “collect together a set of environments that can be used to test the efficacy of tools and techniques for digital preservation, against changes in hardware, software, environment and knowledgebase of the Designated Communities, and design new ones if necessary”.

We in APARSEN believe that people outside of the digital preservation community are looking for guidance on how to preserve their digital material against threats to its continued accessibility and comprehensibility. In particular, we believe that such people are looking to be given the answer to the question: “I have a/these specific type(s) of digital object and I’m worried about this/these threats to them; how do I preserve them?” For example, “I have some CAD files, some images, several data set files and lots of documents and I’m worried about being able to read and use them all in the future; how do I preserve them?”

APARSEN is aiming to build a Virtual Centre of Excellence (VCoE) and one of the services such a VCoE would provide is to answer such questions. To do so, the VCoE needs a body of knowledge which records which digital preservation techniques and tools preserve which types of digital objects against which threats, and conversely, what techniques and tools are not appropriate to use in other instances.

The aim of WP14 is to provide a generic testing methodology that allows digital preservation tools (and similarly digital preservation techniques) to be tested in a rigorous, impartial and replicable evidence based manner such that the results are comparable. Following on from this, the work package aims to develop a way of assessing the capability to preserve digital objects by populating a matrix of:

- digital object types
- threats to the continued accessibility and comprehensibility of digital material
- digital preservation tools (or techniques)

to provide a reliable body of knowledge on which preservation planning decisions can be based.

By being able to periodically review (snapshot) the capabilities of test systems at different points in time it is anticipated that we will be able to demonstrate both the maturing evolution of these test systems as they are enhanced to improve their preservation capabilities, and to use the combined results to highlight any areas for improvements where preservation techniques (and/ or tools) need to be added or further refined to provide better preservation strategies for those types of digital objects.

## 2 INTRODUCTION

Work Package 14 is made up of 2 tasks: **Task 1410 - Identification of testbed techniques and tools** that “collects together the various testbeds which are available”; and **Task 1420 – Testbed suite** that “produces a testbed suite with associated testbed procedures. To facilitate this, project partners will make their testbeds, procedures, test data and software available to other partners”.

Our initial work in Task 1410 therefore concentrated on the identification of digital preservation test environments and test frameworks that were either current or pre-dated the APARSEN project, and the types of techniques and tools that they employed. This has helped us to understand the process through which current test architectures have evolved. Through this work we have identified that document-centric National Archives and Libraries have made significant contributions to digital preservation research in recent years, resulting in a large number of digital preservation test environment designs and preservation strategies that focus primarily on preserving document centric types of digital information object (e.g. text and image based file formats). The CASPAR project (see: <http://www.casparpreserves.eu/>) and its

precursor projects sought to address the digital preservation needs of the data-centric community where document-centric preservation strategies such as simple transformations may not be directly applicable. Our analysis of use cases for scientific datasets has helped us to establish good examples of such instances (described later on in this report) that require us to consider implementing alternative preservation strategies and techniques if these types of digital information objects are to be adequately preserved and remain accessible in the longer term.

We recognise that all preservation techniques have some usefulness, but almost certainly none are usable in every situation. The key for us as a knowledge providing entity (the APARSEN VCoE) is to find out the boundaries of applicability of each of the preservation techniques so that we can provide useful and meaningful advice to people who have these real world digital preservation issues. We are also interested in finding out the capabilities of each system to implement these strategies.

To provide a level playing field on which to assess all the various preservation systems, techniques and tools being developed by the digital preservation community at large we concluded that APARSEN had to come up with a way of defining each digital preservation problem (test case) so that comparative tests could be implemented in any number of suitable test environments. We wanted these assessments to be constructive and preferably quantitative rather than just qualitatively. We also wanted to ensure that our approach would be widely applicable to any existing or future test environments, provided either inside or outside of the APARSEN community, so that this becomes a useful resource in the longer term. We decided to develop an approach that allowed us to group together digital information object types and assess what threats could limit the continued accessibility and comprehensibility of those objects, and then construct tests to assess the performance of preservation techniques and/or strategies to mitigate against those threats.

APARSEN partners were asked to capture “Preservation Scenarios” that describe the real world digital preservation challenges being faced by their user communities. We initially drew our test cases from our own experiences, and then sought to extend these by reviewing the publicly available reports of ongoing digital preservation research projects (e.g. ENSURE, SCAPE, TIMBUS, PrestoPRIME etc) for details of their own test scenarios and test environment activities.

Collating together these digital preservation scenarios provides us with the basis for generating realistic and objective tests of the efficacy of digital preservation tools and techniques to preserve specific types of digital object against specific threats. In our second task in the work package, Task 1420: Testbed Suite, we run these Test Cases with the digital preservation tools, techniques and systems available to the WP14 partners allowed us to capture a snapshot of the capability of each test system. Moreover, analysing the results should highlight those areas of the digital preservation landscape where there is poor coverage in terms of available preservation tools and techniques. In addition, we should be able to compare the ability of different tools, techniques and systems to protect a specific type of digital object against specific threats.

Partner test systems include Tessella’s Preservica cloud based instance of SDB, and a koLibRI based system being developed by the DNB, as well as using the results from the APA’s CASPAR theoretical approach. Preservation Scenarios that provide digital preservation ‘challenges’ for each test environment system were selected so as to provide an assessment of the system’s functional capabilities. By applying these shared tests across all test environment systems we can start to make direct like for like comparisons of each system’s ability to preserve the test objects under threat.

We were keen to highlight the types of digital objects that are currently at risk primarily from a lack of coverage in terms of available preservation strategies or techniques to mitigate the threats that face them. (e.g. National Archives and Libraries now routinely migrate Word DOC files to PDF/A in order to ensure future accessibility of the information objects that they contain, but what would they do with AutoCAD v4 files if they had to maintain this type of content in their collections?)

It is hoped that a body of knowledge can be developed from these Preservation Scenarios about the usage of these preservation techniques and strategies when applied to specific digital object types. The knowledge generated through this process will:

- 1) help identify to the APARSEN partners where our communities preservation capabilities are weaker, and so highlight areas that may benefit from further investigation and research;
- 2) be used by the Virtual Centre of Excellence facilities that are to be established by the project to inform and advise others on preservation strategies and techniques that will be beneficial for them to adopt in order to protect their own digital objects for obsolescence issues.

## **2.1 DESCRIPTION OF WP14 OBJECTIVES**

The DOW describes the objectives of WP14 as:

“Collect together a set of environments to test the efficacy of tools and techniques for digital preservation, against changes in hardware, software, environment, and the knowledge base of the designated communities, and design new ones if necessary.”

## **2.2 DESCRIPTION OF WP14 DELIVERABLE**

The DOW describes deliverable D14.1 (the only deliverable for WP14) as:

“Report on testing environments”

This deliverable consists of a report which summarises the test environments which have been examined and the proposed common testing approach, and also provides a framework within which to evaluate the efficacy and applicability of proposed preservation tools and techniques, for example what types of digital objects the tool/technique is/is not useful for and what types of changes the tool/technique can guard against (changes in hardware, software, environment and the knowledgebase of the designated communities).”



### 3 TEST ENVIRONMENTS AND TESTBEDS

Digital Preservation test systems have been developed over recent years by the digital preservation community to derive evidence for the quantitative assessment of specific preservation strategies and/or actions to safeguard digital information objects of various types from digital obsolescence. The test systems that have been produced to date typically appear either as:

1. Test Environments, where vendor specific instances of preservation systems have been installed in a development or QA server configuration (i.e. not a “production” system) to provide a test framework, allowing investigators to construct and manually run test preservation actions in controlled workflows so that the investigator can quantify and assess any resulting changes in significant properties and behaviours of the digital objects under test, both before and after preservation actions are applied
2. Testbeds, typically created as outputs from collaborative research projects, provided some additional level of automation over a test environment to assist investigators to run comparative analyses of varying preservation strategies on test digital objects. Sometimes also known as a ‘Sandbox’. The idea being that a testbed system can be reset easily to allow new investigations to be constructed and implemented without any impact on other users. More complicated Testbeds allowed the investigator to test preservation scenarios involving a large number of digital objects so that organisational infrastructure decisions could also be tested, e.g. “Could a large set of high resolution TIFF based images be reduced in size by implementing a file format transformation preservation action to JP2 to reduce the overall size of the collection without drastically reducing the ‘human readability’ of the images under test.” The results would inform the investigator of the typical cost saving on IT infrastructure that may also result from making such a preservation strategy decision.

Both Test Environments and Testbeds work towards providing the digital archivist or curator with the information they need to make recommendations on the relative worthiness of applying one specific preservation strategy to a digital object when compared to available alternative preservation strategies. Both types of system should record characteristics of the digital information object both before and after a preservation action has taken place to allow quality assessments to take place. The community has identified these characteristics as being either “significant properties” of the object or various types of understandability or usability of the object by the designated community; when migration tools are used to transform the physical file format type of the object, the “transformational information properties” are used by the investigator to assess whether any quality or usability changes introduced into the digital object through the preservation process are within tolerable limits or not.

We undertook a review of test environments and testbeds that have been developed previously by the community in order to baseline our understanding of the types of techniques and tools that they used to assess the preservation strategies for digital objects. The sub-sections of section 3.1 describe each of the test environments that were identified as part of this study.

During this work we noticed that digital preservation terminology is evolving over time, and that actually different digital preservation communities were coming up with different names for the same actions or objects, or that actually the same word was being used by different people but with different meaning or context, hampering communication between those individuals. In order to combat these issues, we initiated the collection of digital preservation terms for various sources in to a combined APARSEN glossary that was developed on our private wiki, and later published on to the APARSEN website as part of our broader spread of project results.

Some terms that we found relevant to define as part of our work included:

- A digital preservation strategy is a method for preserving digital objects. Examples include migration, emulation and gathering together appropriate representation information and

preservation description information for the object (i.e. CASPAR). A strategy may mandate the use of a particular tool or technique to implement the strategy. E.g. the use of the ImageMagick image file format conversion tool to transform TIFF file objects to JP2000 (JP2) file formats.

- A Testbed is a controlled test environment (including testing procedures & test data) that facilitates evaluation and comparison of systems and strategies through the generation of metrics <sup>[1]</sup>. A Testbed is a more specialised version of a Test Environment, in that it has additional services and functionality to support the automation of tests to facilitate the comparative assessment of alternative and competing processes, such as preservation actions as applied to digital objects.
- A digital preservation system is a computer system which conforms to the OAIS reference model and includes specific functionality to preserve over the long term the digital content entrusted to it.

Specifically, in relation to the APARSEN project:

- The APARSEN Test Environments are controlled environments that enable the evaluation and empirical comparison of digital preservation tools and techniques through the generation of metrics concerning how well these tools and techniques can maintain meaningful access to digital information against changes in hardware, software, environment and the knowledgebase of the designated communities.
- The APARSEN Test Environments will share a common testing framework, which will include a set of test data and a rigorous and systematic testing procedure, in order to ensure that the experiments run in the Test Environments are comparable and reproducible.

### 3.1 TEST ENVIRONMENTS DEVELOPED BY PREVIOUS PROJECTS

There have been a number of previous projects that have carried out systematic testing of digital preservation tools or techniques within their own test environments and have published their test protocols. It is worth examining the outputs of these activities in order to learn from their experience, and try to identify ways of classifying those techniques, tools and environments.

Note: We have adopted a classification scheme for the Test Environments themselves, based on their general approach to providing a preservation capability.

#### 3.1.1 Significant Properties based Test Environments (SPT)

There is a family of closely related test environments which are essentially based on transforming digital objects from obsolete or less accessible file formats to more common and/or accessible file formats and then analysing the changes in a number of pre-selected significant properties associated with those types of digital objects in order to provide a way of quantifiably measuring the performance of the transformation process.

The significant properties that are typically selected mainly relate to the quality, structure, appearance and behaviour of rendering the digital information object both before and after the transformation. A significant property will ideally remain invariant during a transformation event, indicating that no change in the objects quality and / or performance has occurred during the transforming process. The ability to define and use significant properties is also somewhat dependent on the tools that are available to extract (characterise) the properties of the digital object in a quantifiable and repeatable way for the file formats both before and after the transformation process is applied in order to determine the value of the significant property.

Although the use of significant properties has primarily been used to evaluate and assess changes in a digital information object under test following a file format migration preservation strategy, other strategies such as emulation have also used the assessment of significant properties in order to assess the quality of the overall preservation strategy for the object under test <sup>[2]</sup>.

### 3.1.1.1 Testbed Digitale Bewaring

The Dutch Digital Preservation Testbed project, or “Testbed Digitale Bewaring”, was a 3-year project at the start of the millennium that aimed to “help achieve the lasting accessibility of government information in digital form”<sup>[3]</sup>.

The Dutch testbed’s experimental process comprised 12 generic stages which were worked through in order to create an experiment design which can be used to perform one or more related digital preservation tests. The first 5 stages are general stages which are concerned with determining the broad requirements for the experiment (test) design, while the remaining 7 stages are concerned with the specifics of designing and implementing an individual experiment. The 12 stages are:

1. *Define exploration area*
2. *Prepare experiment*
3. *Define requirements*
4. *Develop experiment design*
5. *Specify resources*
6. *Go / no-go decision*
7. *Develop experiment*
8. *Test experiment*
9. *Go / no-go decision*
10. *Run experiment*
11. *Evaluate experiment*
12. *Consider results*

Each stage specifies what is being done, why it is being done, who is doing it, where and when, and how it will be done, as well as a description of any inputs required and outputs expected, and, what will be done with the resulting output. Specifying and documenting this for each stage is the responsibility of a particular member or subset of the project team (and who has responsibility for which stage is given in the paper).

One of the more interesting experimental results (tests) that came out of this work was to compare the outputs of performing migrations of an early Microsoft Word document through several different versions of the format to the current version, and then comparing that result with the single migration of the same document over the different versions. The results demonstrated that in some cases a better representation of the original digital object could be achieved by jumping over the versions rather than performing multiple transformations from one parent – child version through to the next.

### 3.1.1.2 DELOS Network of Excellence

The DELOS Network of Excellence also developed a testbed methodology as part of its work. Their approach used a combination of Utility analysis and the Testbed Digitale Bewaring’s experimental protocol. They considered that the Testbed Digitale Bewaring’s experimental protocol has a “detailed definition of the environment and the experiment basis”, and that Utility Analysis has a “clear hierarchical structuring of the preservation objectives, which documents the requirements and the goals for an optimal preservation solution”<sup>[4]</sup>.

While they use the same 12-stage experimental process as the Testbed Digitale Bewaring, DELOS introduces a metric for testing and comparing various preservation strategies, which is based on Utility Analysis. This uses an objectives tree to identify those characteristics of a digital object that are important to preserve and the other relevant factors (e.g. costs) for comparing digital preservation tools and techniques. Utility analysis attempts to provide a quantitative method for estimating the direct benefit

achievable for each possible solution by assigning value to those characteristics and these objectives are then measured and the measurements combined according to a pre-determined set of weights in order to come up with an objective measure of how well a particular digital preservation tool or technique meets a pre-determined set of requirements.

This introduces cost modelling as part of the outputs from a testbed as the “utility” is measured in monetary terms. Other objectives may also be given weights in order to assess the overall success or failure of a preservation approach, but it is interesting to note that this has similarities to the later work undertaken in systems like PLATO<sup>[5]</sup>.

### 3.1.1.3 Planets Testbed

The Planets Testbed based its methodology based on the DELOS approach, but modified to fit in with its goal of automation. The experiment workflow used in Planets is taken directly from their deliverable documentation<sup>[6]</sup> as follows:

#### Design Experiment

1. *Define basic properties*
2. *Design experiment*
3. *Specify Resources*
4. *Go / no-go decision*

#### Run Experiment

5. *Run experiment*

#### Evaluate Experiment

6. *Evaluate experiment*

Their internal deliverable describes in more detail the testing process and the criteria for evaluating the outcome of each experiment.

The Planets testbed was used to benchmark characterisation and migration tools on specific digital objects, and as such the infrastructure was built to automate these types of experiment. Although it was made available to the general digital preservation community to use while the Planets project was in progress, the public instance of the testbed is no longer easily accessible. It is possible to download the software from Sourceforge and build a private instance with some effort, but this will not have the corpus of test data that was available for use whilst the Planets project was active. The Open Planets Foundation (OPF), a not-for-profit organisation founded to extend the results from the Planets’ work, might make available a publicly-accessible instance, however at the time of writing this document that was not the case. See OPF’s website for latest developments<sup>[7]</sup>.

### 3.1.1.4 DCC Test Methodology

The DCC (Digital Curation Centre) created a test methodology based on the Planets methodology, but focussing on testing digital preservation tools and techniques in context. The context was provided by a set of use cases, which were expected to identify:

1. *What digital resource is being used (what)*
2. *For what purpose is it being used (why)*
3. *In what way is it being used (how)*
4. *By whom (e.g. researcher, student) is it being used (who)*
5. *Within which designated community is it being used (where)*

6. *When (e.g. is it regularly being used or is it being used only one-time) it is being used (when)*

The experiment workflow for those experiments derived from the use cases was intended to be run on the Planets testbed and so is similar to the Planets' experiment workflow:

1. *Develop use case scenarios and model use cases, including what, why, how, who, where, when as described above;*
2. *Define basic properties of the experiment, such as name, description, purpose and focus;*
3. *Design the experiment;*
4. *Specify the required outcomes i.e. quality criteria or characteristics that need to be maintained;*
5. *Go / no go decision based on whether the experiment is feasible;*
6. *Run the experiment in the PLANETS testbed environment;*
7. *Validate the results of the experiments by implementing use case process on results,*
8. *Record the results of 4 and 5 in a DCC report, and the database of the testbeds.*

Further details of the DCC methodology can be obtained from the DCC website <sup>[8]</sup>.

### 3.1.2 Designated Community based Test Environments (DCT)

An alternative to the significant properties approach to the assessment of digital preservation techniques and strategies is to consider **who** currently uses the digital object currently and **who** will want to use that same information object at some time later on in the future. These groups of users represent the information object's designated community, the identified groups of potential consumers who should be able to understand the information contained within the digital object over its entire lifetime. Note, conceptually this approach focuses on the information to be conveyed to the various consumers, not the technical specifics of how the information object is maintained or rendered, and this technical aspect may well vary over time as usage of the object also changes. We know from recent experience that designated communities are also likely to change over time as the information either becomes no longer useful, or else finds new user groups that can aggregate the information and repurpose its data with other information to build new knowledge. Some foresight is therefore required by the users preserving the data in order to provide all of the Representation Information required in order to preserve the context and understanding about why the object was created in the first place.

CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) was an EC funded digital preservation project which ran from 2006-2009. It implemented a set of designated community test environments to verify its approach in a number of different digital data preservation scenarios. The CASPAR framework uses the following validation metrics to assess preservation strategies:

- Demonstrate a sound theoretical basis for the preservation strategy taken, including ensuring compatibility with the OAIS Reference Model and related standards. The alignment with standards that have been extensively peer reviewed by practitioners of digital preservation in a great number of areas demonstrates the strategy is in keeping with current best practice.
- Use practical demonstrations by means of what may be regarded as "accelerated lifetime" tests. These involved demonstrating the ability of the digital information to survive changes in:
  - environment (including software, hardware)
  - the designated communities and their knowledge bases



Members of the appropriate designated community validated the various types of digitally encoded information from various disciplines as well as the information to do with digital rights, authenticity and auditability, for usability and understandability.

The results were a collection of evidence from the domains of science (STFC and ESA), cultural heritage (UNESCO) and Contemporary Performing Arts (CIANT, IRCAM, INA and ULeeds). This evidence was made available in the CASPAR project's publically available Evaluation/Validation report D4104<sup>[9]</sup>.

There are no specific software tools to install and use with this approach. The CASPAR testbed was a test methodology with a collection of preservation techniques and strategies, with worked examples that interested readers could follow or tailor to their own specific preservation needs, and which could be used as a framework to structure the testing of any preservation technique for any digital object. CASPAR provided a general approach to define how to perform accelerated lifetime tests for all types of digital objects. Of particular note is that the focus was on the designated community's ability to understand and use the digitally encoded information, and hence must bring in the semantics needed, for example the language, special terminology or other information needed to understand the context of the information object, which is very clearly needed in the case of scientific data but also most if not all digitally encoded information.

Many other preservation and/or testing techniques are applicable to types of digital objects, which can be migrated, for example to simple human readable forms (like PDF/A). It is always beneficial to take into account the semantics of the language used, as described in the OAIS standard. To be more general, one needs also to be able to also preserve those digital information objects which are not normally just rendered for human interpretation – commonly called data objects. Besides transforming the objects, another technique is to add Representational Information in order to adequately protect the future use of the information object. This Representation Information could be of many different kinds including for example emulation software, and must certainly include the semantics of the object.

### 3.1.3 Transformational Information Properties based Test Environments (TIPT)

The concept of identifying significant properties for file format objects that can be used for assessing the efficacy of a preservation action was questioned by Giaretta *et al*<sup>[10]</sup> as to whether the same concept could be carried through to more complex scientific data objects. (Databases, scientific formats, derived data sets). They proposed that Significant Properties were open to the interpretation of the individuals identifying them and had become specific to the capabilities of the formats in which the information was maintained rather than conceptually belonging to the information object that needed to be preserved. A concept based on the definition of **Information Properties** that describe aspects of a digital objects particular purpose was developed so as to better capture Representational Information properties related to scientific digital objects. This concept was then extended to define **Transformational Information Properties** (TIPs) that are used when transforming digital objects.

The paper defined a **Transformational Information Property** as an information property whose preservation is regarded as being necessary but not sufficient to verify that the non-reversible transformation has adequately preserved information content. This could be important as contributing to evidence about authenticity. Such Information Properties will need to be associated with specific representation information, including semantic information, to denote how they are encoded and what they mean.

Maintaining specific "Transformation Information Properties" are a necessary but not a sufficient condition for full preservation, but on the basis of them a curator can assure him/herself, and others, that a transformation is adequate in terms of regarding the transformed digital object as authentic representations of the original. Note: there is some recognition of information loss following a preservation action in this concept.

This extended concept was carried forward into the revised version of OAIS<sup>[11]</sup> and defined as "Transformational Information Properties".

Tessella's Preservica (SDB) digital preservation system can offer itself as one such example of such a system. Although it has a pedigree based on the developments of SPT systems and as such extracts physical properties pre & post transformation, it can also maintain 'conceptual object' properties throughout this process due to its domain and data models. Physical properties are mapped to conceptual properties and then comparing and assessing these conceptual properties before and after migrations (transformations) at the information object level are in keeping with the TIPT concept.

A good example of this behaviour can be demonstrated by archiving a website that contains many links to embedded text and image based file formats. A transformation of the image formats from BMP to JPG may result in the website still being rendered with the same conceptual information available to the end user, they can still interact with the embedded images in the same way, yet the underlying image formats themselves have changed. The Preservica system also provides users with a component comparison screen where the Transformational Information Properties both before and after transformation are presented to the system user for assisting them with their assessments of transformation quality.

The APARSEN test framework provides the ideal vehicle to challenge this viewpoint and to seek to prove or disprove the validity of this classification as part of its ongoing assessment work.

### 3.1.4 Multi-Valent Test Environments (MVT)

The SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) project was a 4-year EC funded project that started in 01-DEC-2007 <sup>[12]</sup>. During the course of the project it implemented three prototype demonstrator test environment systems where the SHAMAN digital archiving framework was used to demonstrate and validate the SHAMAN approach to digital archiving for typical real-world scenarios through the prototyping of technologies and services to provide solutions to these domain areas. The SHAMAN project validated its software by using standard software validation protocols, such as referencing the design back to the requirements and the requirements back to the use cases and then basing the system testing on the use cases and requirements.

The preservation approach taken by the SHAMAN framework was to develop a Data Format Description Language "DFDL" to describe the various bitstreams that represent digitally encoded information objects so that objects to be preserved can be expressed in a form that is separate from the technology that is used to implement it in its current form. Once an information object is expressed in DFDL it can then be rendered in virtualised environments that implement the display of DFDL described objects so that the information objects themselves are then displayed in the same way as when they were originally created. These presentation tools are considered to be "Multivalent" in that they can be used to present digital objects of many file formats through the interpretation of the technology independent language that describes them. Preservation approaches therefore require not only the migration of the original digital objects to a DFDL format, but also the migration of the behaviours exhibited by the presentation tools that are used to render those formats. Both parts are essential to the long term accessibility of the information objects being maintained by this preservation approach.

### 3.1.5 Commentary on previous test environment approaches

Some members of the digital preservation community have raised concerns that the concept of Significant Properties is ill-defined for data centric objects, as what constitutes a significant property is open to subjective interpretation and may be missing the vital semantic representation information required to guarantee long term preservation and authenticity of the digital objects <sup>[10]</sup>. Also a good number of the test environment systems developed during previous projects are no longer online or available for use anymore, further reinforcing the fragility of digital information research and development, and the need for these types of project to plan for their longer term sustainability.

### 3.2 APARSEN TEST ENVIRONMENT SYSTEMS

During the project's inception, the consortium partners realised that there was a need to provide access to Test Environment facilities to generate the evidence needed to critically appraise emerging preservation strategies, however the constraints of the project's available funding would limit our scope to sourcing systems that were either provided by the consortium partners, it would not be possible for the consortium members to try to resurrect test environment software from previous projects. We investigated whether the test environments and testbeds from previous projects were being actively maintained and operated by their founding members, however we did not find this to be the case, and our recent email discussions with prominent members of those projects have re-affirmed that these older systems are no longer accessible.

On that basis, the following test environment systems were offered by APARSEN partners for the purposes of providing further testing of our understanding of implementing preservation strategies on specific digital information objects. These test environments include:

- Tessella's Preservica ([SDB](#)) digital repository service that includes all aspects of the OAIS functional model including Preservation Planning functional workflows that incorporate common Characterisation tools such as DROID, JHOVE, Registry (a.k.a PCR / PRONOM) and a variety of wrapped migration tools from both the commercial and public domain sectors. Demonstrator systems are used by Tessella to re-assure prospective clients of the types of workflows and tools that can be deployed in order to demonstrate and initial assess the capabilities of implementing various preservation policies and strategies with their typical digital objects.
- DNB are to provide their [koLibRI](#) (kopal Library for Retrieval and Ingest)-tool. koLibRI represents a library of Java tools that have been developed for the interaction with the DIAS system of IBM within the kopal project. It has been designed with the intention to be re-usable as a whole or in parts within other contexts, too. It has been enhanced by further digital preservation projects (DP4lib, textGrid, LuKII, SHAMAN) including integration with the FITS characterisation tool, and has been used as a test environment for these project. Currently, one instance of koLibRI runs at the DNB on a Debian GNU/Linux Server with connection to the internet.
- STFC / APA will make the CASPAR test strategy and the data in the CASPAR collection available for re-use by the APARSEN partners.
- SCAPE test environments: Several members of the APARSEN project (KB, BL, APA, STFC, ONB, MRL) are also involved in the EC funded SCAPE project that will investigate SCALable Preservation Environments, and will seek to establish several testbeds as part of that project. The SCAPE DoW states "The main goal of the testbeds is to assess the large scale applicability of the SCAPE Preservation Platform and the preservation components developed within the project. Using these software components, it creates test environments for the different application scenarios [Web Content, Scientific Data Sets, Large Scale Digital Repositories -BS] and complex large scale preservation workflows which will shed new light on existing preservation services and the improved or new components developed within the project". The project's website now provides access to the emerging deliverable documents that detail the approach to be taken by the three testbeds and the types of data and preservation scenario that is considered to present an operational challenge when run at scale <sup>[13]</sup>. Following our recent contact with the SCAPE project team responsible for their project's test environments, we will be seeking to share resources amongst the two projects and include assessments of the SCAPE testbeds in future revisions of this work. Investigations will start once the SCAPE testbeds become operational after commissioning for use by September 2013.

Note: Although the SHAMAN Integration & demonstration SubProjects (ISP) have been developed by APARSEN partners, by design they are highly focussed towards performing specific preservation actions



on select types of digital objects. Although the SHAMAN preservation approach could be applied to a wide number of other digital object types within other scenarios of use, the SHAMAN project has now completed and there is little opportunity for facilitating these changes as part of the APARSEN work plan.

During the identification work undertaken as part of task 1410 we also looked at the possibility of including Testbed LDP, that was financed by the European Regional Development Fund (ERDF) and the County Council of Norrbotten in Sweden<sup>[14]</sup>. The project is a first step towards a flexible testbed, built on loosely coupled modules which are easily exchanged as technology evolves. The testbed is to reflect the entire OAIS-model, i.e. cover testing of digital information from delivery to e.g. an archive to management of the information in the archive to a future customer gaining access to the information. Reviewing the maturity and the capability of the system it was decided that at present we would note this as a possible future candidate test environment, but would not include it in our initial set of test environments.

### 3.3 THREATS TO DIGITAL MATERIAL

The PARSE.Insight project identified, through a series of extensive surveys that generated thousands of responses from across many disciplines around the world, a set of 7 threat types to digital material:

- 1) Users may be unable to understand or use the data e.g. the semantics, format, processes or algorithms involved.
- 2) Non-maintainability of essential hardware, software or support environment may make the information inaccessible.
- 3) The chain of evidence may be lost and there may be lack of certainty of provenance or authenticity.
- 4) Access and use restrictions may make it difficult to reuse data, or alternatively may not be respected in future.
- 5) Loss of ability to identify the location of data.
- 6) The current custodian of the data, whether an organisation or project, may cease to exist at some point in the future.
- 7) The ones we trust to look after the digital holdings may let us down.

The ability of different preservation systems, tools, and techniques to preserve different types of digital material from these threat types will be tested as part of our work within this project. Should additional threat types be identified through our ongoing developments, this list will be extended to include these threat types as well.

## 4 TEST ENVIRONMENTS EVALUATION FRAMEWORK

Having categorised the types of digital preservation test environment that have been proposed by the community (SPT, TIPT, DCT and MVT), the next step is to set out a common approach for classifying and evaluating the tests that are performed on these systems so that the APARSEN community and its subsequent Virtual Centre of Excellence can fairly and without bias “evaluate the efficacy and applicability of proposed preservation tools, techniques and services” irrespective of the type of test environment used to preserve the digital object.

The APARSEN community also believe that the digital preservation community is currently fragmented into various silos of interest based on the types of information that those specialised communities seek to preserve, and this is reflected in the tools, techniques and/or services that they promote. We want to disseminate best practice amongst all interested parties irrespective of personal or commercial interests, and provide unbiased recommendations to new user groups that are becoming aware of the need to take action to preserve their digital information assets for the longer term.

By developing a process to evaluate preservation capability through tangible measurements of performance, independent of the underlying technology, we hope to provide a resource that provides clear evidence when making any future recommendations of what preservation policies and/or strategies may be useful for new or future preservation scenarios that are presented to the VCoE in the form of requests for help or assistance. We also hope to highlight any gaps in currently available preservation policies and strategies so that additional research and development effort can be targeted towards these areas, demonstrating a more common understanding and converging vision in the provision of future preservation orientated solutions.

### 4.1 DEVELOPING A COMMON APPROACH

Test environments provide their users with the ability to construct tests (known as cases) that specify the inputs and operating pre-conditions for a specific state of a system, so that when the system is allowed to react to these influences, its resulting behaviour can be assessed and verified against its expected performance. Each test case is defined in enough detail so as to make the whole test process repeatable, without unexplained variations occurring in the resulting behaviour of the system.

As demonstrated by our review of previous work, the digital preservation community have to date developed test environments that allow users to construct test cases relating to the preservation of specific digital objects that are generally of interest to their select interest groups mainly due to practical reasons such as applicability or limitations of funding.

A clear gap in understanding relating to the preservation of more complex data formats exists, and this view is supported by the lack of internal and external file signature data currently built into tools like DROID, or the number of property extraction tools that exist for common scientific data formats. This means that formats like those created by scientific research equipment like gas chromatography machines, spectrophotometers, CAD, have little if any supporting digital preservation best practice advice available to their users.

Our aim within WP14 of APARSEN is to bring together results from the wide variety of known and currently available preservation tools and techniques in to whatever test environments are currently available within the project’s consortium of partners. As a minimum this will include providing project partners with access to a test instance of Tessella’s Preservica Preservation as a Service (PAAS) offering. These systems will be used to demonstrate the applicability of the APARSEN common approach for evaluating the effectiveness of a specific preservation tool or technique. Our goal will be to provide a consolidated independent body of evidence to support the Virtual Centre of Excellence when assessing and advising on any future preservation strategy enquiries made of it.

The approach sought should provide us with a simple method for qualitatively assessing the capabilities of a digital preservation test tool or technique to preserve different types of digital information objects irrespective of the test environment’s underlying technology. The sum of these works should also provide

a means to assess the capability of the digital preservation community to provide real world preservation solutions.

This approach needs to be extensible so that as and when new and emerging digital preservation tools/techniques/environments are created, the approach can be updated and new conclusions can be drawn. By periodically performing and updating the assessment of the capabilities of the preservation community to provide preservation solutions we anticipate that we can demonstrate how the capabilities within the digital preservation landscape are evolving over time, adapting to provide solutions to new challenges or utilising new technology to overcome more difficult well known issues.

Collating together the results from a variety of test environments, we will be able to compare and contrast the corresponding success or failure of specific preservation techniques to a wide variety of known preservation challenges.

Specifically, we want to use independent evidence to be able to determine objectively which preservation techniques can protect which types of specific digital objects from which sorts of threats.

By performing this analysis we hope to develop a common shared approach for assessing the capabilities of various digital preservation test environments. As the Virtual Centre of Excellence matures, we can see this approach forming part of our knowledge base when consulting on digital preservation issues to the wider public community.

## 4.2 IMPLEMENTATION STRATEGY

To achieve the work package objectives and the aims as set out above, the work package team needed to agree an implementation strategy for the work package. Much debate and discussion lead to the conclusion that the best way forward would be to adopt a process of classification of the object types that need to be preserved, a classification for the types of threats that may be imposed on those objects, and the identification of the tools and techniques used to provide the preservation solutions to mitigate those threats. It is important that we 'throw a wide net' across the digital preservation community space as we want to gather enough data points to build up our 'picture' of preservation capability, and whilst doing that we also need to make sure that we do not end up repeating known results, concentrating on just one small aspect, or else failing to identify whole sections of the digital preservation landscape.

The assumption was therefore that similar objects could most likely be assumed to be preserved by following preservation strategies that had been shown to be successful for objects of that type. In fact we wanted to challenge this assumption as part of the ongoing assessment of our approach.

By mapping out the various combinations of object / threat / technique combinations we are seeking to find the combinations where there is no adequate preservation solution, thus helping us to identify where the gaps are in our provision of services.

### 4.2.1 Classification

To support our approach above, we agreed to develop classification schemes for:

- Digital objects
- Threats to digital objects
- Preservation tools and techniques to combat the threats to digital objects

**Digital objects** hold the information object(s) (or their components) that need to be preserved. Conceptually, an information object may be self-contained in a single file (e.g. an image file, or a text file) or may be distributed across an aggregation of many digital file objects (e.g. a website, or database, where the index.html or table1.dat files respectively are likely to be just one component file of the many files needed to access the bigger parent information object). Using this conceptual approach helps us to decouple the information that needs to be preserved from the physical implementation of the digital objects (files of various formats) that are used to represent that information on the storage media.

**Threats to digital objects** come in many forms, however the Parse.Insight project was able to define 7 major types of threat that could potentially impact on every digital object. These major types of threat are stated in section 3.3 on page 17. The project team decided to use these threat types in the first instance, on the basis that these seven were comprehensive in their coverage and that the approach that was being sought would be extensible if new threats emerge.

**Preservation tools and techniques** (or strategies) are the processes that can be implemented to remove or else reduce the threat of loss of access or re-use of the information object due to changes in hardware, software, environment and tacit knowledge.

Data-centric objects are more likely to be processed by some consuming system (e.g. by combining with other objects) and generally require explicit semantics as well as more extensive format information; whereas document-centric objects are generally used with specific software which embed the structural semantics (although not the document content's semantics), than scientific data formats. This assumption is however changing as even text based document software applications become more complex in terms of the features that can be specified by their users for describing the appearance during rendering of the data objects, requiring ever more complex file format structures be developed.

In addition, the designated community that is interested in the data centric objects will change over time and may follow the trend for data 'mash ups' where the results from one designated community are re-used by another to further extend and influence their own analysis, further emphasising the need to capture the environment that surrounds these types of object in may be more detail than that based on static imagery and text. That said, we can now start to see how dynamic web 2.0 sources are also beginning to challenge this view that document centric information is also static.

#### 4.2.2 Defining a Digital Preservation Landscape

To develop the digital preservation landscape view we need to capture information about the capabilities of current and previous test environment deployments, based on what they were used for and how successful they were at doing it. To facilitate this data collection process we decided to:

- Examine the current coverage by looking within APARSEN partners at the
  - types of digital objects available as examples
  - types of threats that impact the future accessibility of those types of digital objects
  - types of test techniques and tools used to reduce or negate those threats.
- Have a mechanism to identify what types of digital objects our examples do not cover.
- Create a matrix for each test environment to highlight the techniques that it can implement for each digital object type and the types of preservation technique that could be applied to it.
- Seek out additional resources from outside the APARSEN project community to further broaden our view of the digital preservation landscape.
- Analyse the collated results from within APARSEN and externally, to ascertain to what extent the current preservation techniques have been tested and/or adopted and identify where any gaps in test coverage may exist. This data should help influence the direction of future digital preservation research.
- The final outputs from this WP will be the array of test objects and the preservation and test techniques as well as a summary of what has been discovered about the efficacy of the current preservation techniques.

### 4.3 CAPTURING PRESERVATION SCENARIOS

A preservation scenario provides a description of how a user will interact with a system to achieve a single specific function or objective. It defines the context of why a user wants to perform the actions that they do, and defines the expected behaviour and outcomes of that system for that user.

Sets of preservation scenarios can be created to define the majority of the actions a user may have with a system. These sets of preservation scenarios can then be applied to each test system in order to evaluate its ability to respond to these tasks, and help us evaluate the capability of one system against another in a qualitative way.

In WP14 we decided to develop a preservation scenario driven approach to capture information relating to the types of digital objects that had been identified by our project participants as typical examples of digital material that needs to be preserved, the type of test environments that may have been used in the past to analyse and test the preservation of these objects, and the associated preservation techniques that can be applied to those objects.

Aligning our data collection process with a preservation scenario approach should allow external non-technical staff to be able to make enquiries of the VCoE about their own digital content collections. The VCoE representatives can use the data provided by them to recommend those approaches that already have good documentary evidence to confirm that they provide an acceptable solution, or else, have good similarity to an existing solution that may help to define a preservation strategy that meets some if not all of their needs, or could be extended to fully meet their needs.

Collating together preservation scenarios from real world digital preservation user communities allows us to define our classification knowledgebase, and form a ‘back catalogue’ of known workable solutions. Unlike some projects where there is a specific limit to the scope of applicability, APARSEN needs to provide to the VCoE with a framework that can assist with any type of preservation activity that is brought to the VCoE’s attention. Initially the WP will document Preservation Scenarios provided by APARSEN partners during WP14, and as the work matures over time, the APARSEN participants will look at taking a wider view by obtaining the preservation scenarios from associate partners and external organisations and projects as part of our later work in WP11 and WP44.

In addition to describing preservation scenario examples of known digital preservation challenges, we also sought to obtain sample data sets of the digital content where appropriate to do so. There are a number of publicly available websites and repositories that hold test corpora of digital content. Where it is not possible to obtain links to the data described within a preservation scenario (e.g. due to access control reasons, etc) then alternative pseudo data sources may be provided so that test environment providers can test the type of content that is referenced to within the preservation scenario.

It will not always be possible to provide test environments with access to test data objects, either for security reasons or because it is dependent on an unavailable technology stack object like obsolescent storage media or operating systems. Note: Provision of data examples was not mandatory for a preservation scenario to be included within our catalogue, just very much appreciated.

#### 4.3.1 Preservation Scenario Template

A preservation scenario template was developed (see **Fehler! Verweisquelle konnte nicht gefunden werden.**) to maintain a consistent structure throughout the process of capturing all of the preservation scenarios provided by project partners for use in the WP14. This template was maintained online using our APARSEN project wiki, allowing project partners to clone and modify it to create their own lists of preservation scenarios.



<b>Preservation Scenario ID</b>	<i>Unique Identifier for the Preservation Scenario</i>
<b>Author</b>	<i>Name and Organisation details</i>
<b>Background</b>	<i>A brief description of "what, why, where, when, and who" relating to the Preservation Scenario</i>
<b>Type of digital information</b>	<i>Describe the material that needs preserving</i>
<b>Link to sample data</b>	<i>Provide a link to the samples of data that are available for testing within the Testbeds</i>
<b>Threat(s) to the data</b>	<i>Define the problems encountered/expected in keeping this data accessible</i>
<b>Preservation Technique</b>	<i>Define the preservation technique(s) that can be used to maintain access and re-use of the information</i>
<b>Designated Community</b>	<i>Define the Designated Community which should be able to use and understand the digitally encoded information</i>
<b>Usage</b>	<i>Brief description of the users and how they use the data</i>
<b>Success Criteria</b>	<i>Describe how you would hope to validate whether the digital information has been preserved against you identified threats</i>

**Table 1 Scenario template**

### 4.3.2 Collection of Preservation Scenario data

The WP14 project participants drew upon their own experiences of preserving digital information objects to develop a broad set of Preservation Scenarios that between them demonstrate a need to preserve a variety of digital objects of many forms, from both the document and data centric domains. Links to test data were provided where available to do so, so as to make it possible for partners with test systems to create test submission information packages (SIPs) to be ingested in to their test environments.

As there is a large amount of detail relating to each preservation scenario, these have been replicated from the project's wiki in to Oof this document, and edited to maintain the readability of the preservation scenario being described.

## 4.4 ANALYSIS OF PRESERVATION SCENARIOS

The purpose of obtaining preservation scenarios is three fold. In the first instance, it helps us develop a body of knowledge on what techniques and tools for digital object preservation have been used previously by experienced practioners within the digital preservation community. This will reveal useful sources of information such as documentary evidence about how successful the technique or tool was in achieving the stated preservation objective.

Secondly, it provides us with the data that we need to validate our approaches for classifying digital objects, the threats on those objects, and the preservation techniques and tools that can be implemented in the available test environments.

Thirdly, it becomes possible to filter out the duplication that naturally appears in the above two processes to develop a test environment capability matrix to 'snapshot' the current capabilities within the digital preservation communities landscape. This matrix can be created by extracting the unique combinations of digital information objects, the threats to their future access, and the preservation techniques and tools (unique test case instance) from the body of collected preservation scenarios.

Ideally these test case descriptions should be independent of any associated implementing technology as far as is technically possible in order to maintain their value as platform independent test cases and indicators of applicability for the preservation solution that they promote.
















The preservation scenarios that were captured by the project's partners demonstrate the wide diversity of file format types and uses of digital objects that are threatened by digital obsolescence issues. It is fair

comment to say that the preservation scenario data that has been captured by our work is heterogeneous in its nature, as it has been obtained from a wide number of sources that do not all have the legal authority to share the actual test data with the consortium. However, the learning experiences captured in these partial preservation scenario descriptions can still be insightful, explaining what preservation strategies were employed to overcome some of the technical issues relating to that type of digital object under threat. These insights should help to advise and influence future digital curators that have not experienced these types of data structures before, and may assist them in constructing adaptations to current processes to implement the solutions that were only partially presented in this work. A strict adherence to data quality in this instance may be detrimental to the objective of the task – to capture knowledge relating to the preservation of digital objects.

#### 4.4.1 A method for assessing the capabilities of Test Environments

By building up sets of tests that represent the challenges facing the wider digital preservation community it becomes possible to perform a general qualitative assessment of the capability of each test environment to preserve digital objects of varying types. We can envision the creation of a test corpus of digital objects that can be presented to test environments in order to qualitatively assess its ability to provide preservation capabilities for all of those objects in that test set. A simple assessment of whether the test environment was able to provide a solution or not for each test object/threat/strategy combination (marked as either available, partially available, or not available and where “strategy” equates to a known tool or technique) can provide a quick reference guide on the capability of the evaluated test environment.

Table 2 below provides an example of what such a matrix based evaluation may look like, based on some preliminary test case data, rating SDB and 2 other anonymous test environments. This is purely for demonstration purposes of the approach, and is not part of the final research outcomes for WP14, merely to assist with understanding the concept that we are developing.

Object	Threat	Strategy	SDB	TE #2	TE ..n
Image (*.tif)	Long term cost of storage	Migrate to .JP2 using tool: ImageMagick			
WordPerfect (*.wp)	Cannot read document	Emulate using Dioscuri			
3D CAD (*.prt)	Cannot access CATIAV4 files	Migrate using LOTAR techniques			
Error flag image (*.?)	Semantics of Image flags may be lost	Define usage by adding more Representation Information			
Database (*.mdb)	Cannot access DB file format	Use SIARD Suite to normalise to the .SIARD format			

**Table 2: Example of a Test Environment Capability Matrix**

(Green Tick = solution available; Amber Circle = partial solution available; Red Cross = no solution available)

#### 4.4.2 Monitoring the evolution of a Test Environment over time

To demonstrate the usefulness of capturing the capability of a test environment at various points in time, we have used the above approach to develop a capability matrix for Tessella’s SDB Digital Preservation product.

Tessella have made a test instance of Preservica (SDB, offered as a Service in the Cloud) available for the APARSEN community to access and use without restriction other than it must be for work relating to the APARSEN project. Although this instance is not a testbed, in that we are not mandating a test methodology like those defined in section 3.1 of this document, it can be considered a test environment in that the system is available to ingest any content whatsoever presented to it as part of an APARSEN test, and Tessella has made all the default file format migration tools that come with SDB by default accessible to this test instance. (e.g. Imagemagick, OpenOffice, etc) Should any of the other WPs in APARSEN implement a new service or capability that would benefit from being realised within a Digital Preservation System workflow, then this system is available to be used to implement and demonstrate any such capability if this is required by the APARSEN project team.

Using SDB, we thought about what capabilities this system had at the start of our work on APARSEN, and then thought it would be interesting to see how we could graphically represent this capability in a simple to evaluate way. We came up with the idea of creating a matrix that lists the objects that needs to be preserved as Rows, and the OAIS functional aspects of a test environment as columns. For each combination we assess whether the test environment can demonstrate full capability ('Y' = Yes), no capability ('N' = No) and finally where the test environment can demonstrate some partial aspects of achieving that capability but is not currently mature enough to say this is complete ('P' = Partial).

We have applied a "traffic light" colour code to each of these ratings, so as to build up a matrix that can easily demonstrate its current levels of capability.

Here we have summarised the digital preservation objects at risk in to high level types, where we are not so worried about the exact file format type, but more the typical nature of the object, and based on the assumptions that the digital preservation community would accept the preservation approach being applied to those objects as being current best practice.

The large PARSE.Insight survey <sup>[16]</sup>, showed that, from information provided by 1366 researchers, the following distribution of formats of digital objects to be preserved. The report notes that:

*...not surprisingly, office documents are most often used by the respondents. What is a bit surprising perhaps is that still 6% of the respondents do not use office documents. The other two of the top three most used data formats are: network-based data (web sites, e-mail, chat history, etc.) and images (such as JPEG, JPEG2000, GIF, TIF, PNG, SVG, etc). For both data formats 79% of the respondents claimed to use them.*

*What is rather more surprising is that almost half of respondents have source code, software applications, raw data and databases. It is likely that these forms of digital objects offer significant challenges in terms of usability and understandability, beyond those of documents and images.*



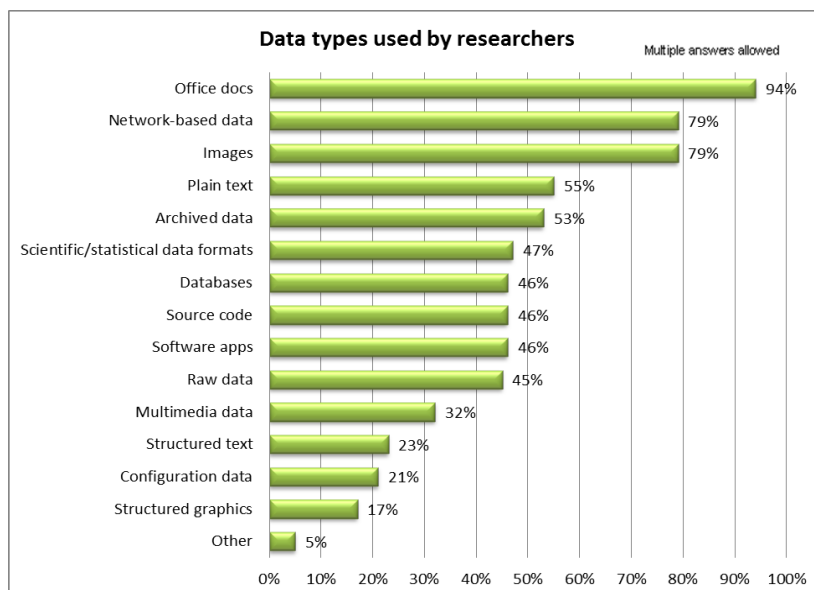


Figure 1 PARSE.Insight survey results showing data types used by researchers (from 1366 responses)

There are many ways to classify the objects used in our examples but we summarise the object types as follows:

- **Byte Streams:** Any file byte stream as maintained on a modern Operating System file store. This however only results from passive preservation of the bytes on storage media and does not include any use of preservation tools or techniques to migrate the object.
- **Document:** Typical office file formats like Word (DOC), Excel (XLS), Powerpoint (PPT)
- **Images:** Well known formats like TIFF, JPG, GIF, PNG, JP2, BMP.
- **AudioVisual:** Tools like FFMPEG support a number of formats like AVI, MPG, etc, and allow some migration between file formats, however the added complexity of CODECs means that the preservation of these objects is only partially defined
- **Websites (IIPC Style):** Web Archive tools like Heritrix can generate container format files that hold a complete snapshot of a given website, in formats like WARC, ARC.GZ
- **CAD:** Autocad (DXF) Catia (CAT\*), STEP (STEP), I-DEAS (UNV)
- **Databases:** Oracle, DB2, SQL Server, MySQL, Postgres
- **'Expressions' of Objects:** Planets based term for describing independent, non-derived, file formats of the same information object. E.g. A 3D CAD model file and a 3D cloud of points file of a solid object. A conceptual entity for modelling multiple views of the same complex information object.

OAIS Function ⇨		Ingest	Storage	Access	Preservation
Object Type	Byte Streams	Y	Y	Y	P
	Document	Y	Y	Y	Y
	Images	Y	Y	Y	Y
	Audiovisual	Y	Y	Y	P
	Websites (IIPC Style)	P	N	N	N
	CAD	N	N	N	N

Databases	N	N	N	N
'Expressions' of objects	P	P	P	P

**Figure 2: Capability of SDB at the start (M1) of the APARSEN project**

OAIS Function ⇒		Ingest	Storage	Access	Preservation
Object Type	Byte Streams	Y	Y	Y	P
	Document	Y	Y	Y	Y
	Images	Y	Y	Y	Y
	Audiovisual	Y	Y	Y	Y
	Websites (IIPC Style)	P	P	N	P
	CAD	N	N	N	N
	Databases	N	N	N	N
	'Expressions' of objects	Y	Y	Y	Y

**Figure 3: Capability of SDB now (M24) of the APARSEN project**

Reviewing the two matrix plots shown in Figure 2 & Figure 3 for differences reveals that the addition of the FFMPEG migration tool for audio-visual formats has improved the capability of SDB to preserve these types of digital object. It also shows that SDB is maturing in its development towards fully supporting Archives of websites in compression formats.

Although the above example demonstrates a rather simplistic prototype of the method, it shows that the method does not need to be overly complicated in its approach to provide a meaningful assessment of the capabilities of a given Test Environment. More complex versions can be derived by further drilling down in to the specifics of each information object category under examination and explicitly listing the different threats that may be present for each of those objects in to their own unique separate rows within the matrix. This however will make comparisons less easy to make visually as the number of managed items grows, but can be managed by applying suitable filters to the data that is collated together to form the matrix.

#### 4.4.3 Assessing the capabilities of Tests Environments to preserve digital objects

It is anticipated that the Virtual Centre of Excellence will want to develop and maintain such a matrix in the future so as to provide itself with the tools to be able to quickly locate references to known preservation solutions for specific objects, or to draw parallels to similar objects to make judgements on what suitable alternative approaches may be taken and promote these suggestions to its end user communities.

Using the preservation scenarios that have been collected to date, it is possible to develop a more complex matrix that lists all of the digital objects that have been identified within our preservation scenarios as rows, and then further specify the detail of the preservation task by adding additional columns of data to further exact nature of the test case instance. Additional data include identifying the primary preservation threat to the object, the preservation technique that was employed (if data is available) and the tool (if any) that was used to perform the preservation action.

Once the specifics of the test case instance have been entered in to the matrix, the columns in the matrix then change functionality to indicate the status of each test environment in its ability to implement the solution described by all the previous column specifications. If the test environment does implement a solution for this test case instance, then the status is marked as 'Y' for Yes, or 'N' for No, and 'P' for Partial, if the test environment can demonstrate some capability towards this solution. These status values can then be mapped back to our traffic light analogy by being coloured Green, Red or Amber respectively.

The additional columns (rendered, static, simple, passive), are described in section 4.5.1, show the coverage of the digital landscape we currently have.

Additional columns of data can be added or removed (hidden – never removed) from the matrix, and filters applied, so as to build up a searchable list of known preservation actions that can provide evidence to demonstrate the effectiveness of a specific tool or technique has been shown to preserve these types of object over the longer term from changes in hardware, software, environment and the knowledgebase of the designated communities.

So far this initial work is in a proof of concept phase whilst the VCoE assesses whether providing such a resource will be useful for its work. The spreadsheet that has been developed to date will be maintained through the remainder of the project by work within WP11: “Common Vision”, where the processes required to establish the VCoE are being developed, with the vision that this may at some point in the future be implemented in a scalable web based architecture.

A screenshot of some of the content maintained in the spreadsheet is shown in Figure 4 for illustrative purposes.

Work Package 14 - Assessment of Test Environments Capabilities to Preserve Digital Objects											Test Environment				
Digital Object details											Capability to Implement the Tool/Technique				
Digital Object Name (EXT)	Digital Object Encoding Name	Information Type	Rendered	Static	Simple	Passive	Digital Information Threat (Passive Insight List)	Preservation Scenario (1st encounter)	Preservation Technique	Tool	SDB	KoLiRI	PLATO	CASPAR	SCAPE
DBF	dBase Database Table files	Database	Non-rendered	Dynamic	Complex	Passive	2) Inaccessible form	PS-001	Migration	MIXED	N	N	N	Y	
MDB	MS Access	Database	Non-rendered	Dynamic	Complex	Passive	2) Inaccessible form	PS-001	Migration	MIXED	N	N	N	Y	
XLS	MS Excel	Database	Non-rendered	Dynamic	Complex	Passive	2) Inaccessible form	PS-001	Migration	OpenOffice	P	N	Y	Y	P
STR	DataPerfect	Database	Non-rendered	Dynamic	Complex	Passive	2) Inaccessible form	PS-001	Migration	MIXED	N	N	N	Y	
PDF	Portable Document Format	Document	Rendered	Static	Complex	Passive	2) Inaccessible form	PS-002	Migration	BFO Java PDF	P	P	P	Y	
HTML	HyperText Markup Language	Website	Rendered	Static	Complex	Passive	2) Inaccessible form	PS-002	Migration	OpenOffice	Y	P	Y	Y	
JPG	Jpeg Image	Image	Rendered	Static	Simple	Passive	2) Inaccessible form	PS-002	Migration	ImageMagick	Y	Y	Y	Y	P
PDF/A-1b	Archive version of PDF	Document	Rendered	Static	Complex	Passive	4) Access restrictions	PS-003	Migration	OpenOffice	P	Y	N	Y	
TIFF	TIFF image	Image	Rendered	Static	Simple	Passive	2) Inaccessible form	PS-004	Migration	ImageMagick	Y				P
J2K	Jpeg 2000	Image	Rendered	Static	Simple	Passive	2) Inaccessible form	PS-005	Migration	ImageMagick	Y				
LZW	Lempel-Ziv-Welch Compression	Compression	Non-rendered	Static	Complex	Passive	1) Misunderstand data	PS-006	Migration	Unknown	N				Y
BWF	Broadcast Wave Format	Audio	Rendered	Static	Simple	Passive	2) Inaccessible form	PS-007	Migration	Unknown	N				Y
FLAC	Free Lossless Audio Codec	Audio	Rendered	Static	Simple	Passive	1) Misunderstand data	PS-007	Reinfo	Registry	N	N	N	Y	
WavPack	WavPack compression	Audio	Rendered	Static	Complex	Passive	1) Misunderstand data	PS-007	Reinfo	Registry	N	N	N	Y	
MSST	Binary Radar data	Data	Non-rendered	Static	Complex	Passive	1) Misunderstand data	PS-008	Reinfo	Registry	N	N	N	Y	
XML	eXtensible Markup Language	Data	Non-rendered	Static	Complex	Passive	1) Misunderstand data	PS-008	Reinfo	Registry	P				Y
WARC	Web Archive file	Website	Non-rendered	Static	Complex	Passive	1) Misunderstand data	PS-008	Reinfo	Registry	P				Y
VICAR	Video Image Communications and Retrieval	Image	Non-rendered	Static	Complex	Passive	2) Inaccessible form	PS-009	Migration	Unknown	N	N	N	Y	
FITS	Flexible Image Transport System	Image & Data	Non-rendered	Static	Complex	Passive	1) Misunderstand data	PS-009	Migration	Unknown	N	N	N	Y	
TEXT	ASCII Text File	Data	Rendered	Static	Simple	Passive	1) Misunderstand data	PS-009	Reinfo	Registry	P	Y	Y	Y	
FITS (IMAGE)	A Calibration (offset) image	Data	Non-rendered	Static	Complex	Passive	2) Inaccessible form	PS-009	Migration	Unknown	N	N	N	Y	
VOTable	VOTable (XML)	Data	Non-rendered	Static	Complex	Passive	1) Misunderstand data	PS-010	Reinfo	Registry	N	N	N	Y	
CDF	Computable Document Format	Data	Non-rendered	Dynamic	Complex	Active	2) Inaccessible form	PS-011	Migration	EAST	N	N	N	Y	
EAST	EAST data description language	Data	Non-rendered	Static	Simple	Passive	1) Misunderstand data	PS-011	Migration	Unknown	N	N	N	Y	
CSD	Coordinate 3D	3D Geometry	Non-rendered	Static	Complex	Passive	2) Inaccessible form	PS-012	Reinfo	Unknown	N	N	N	Y	
TRC	Tracked Row Character	3D Geometry	Non-rendered	Static	Complex	Passive	2) Inaccessible form	PS-012	Reinfo	Unknown	N	N	N	Y	
BVH	BioVision Hierarchical Data	3D Geometry	Non-rendered	Static	Complex	Passive	2) Inaccessible form	PS-012	Reinfo	Unknown	N	N	N	Y	
HTR	Hierarchical Translations and Rotations	3D Geometry	Non-rendered	Static	Complex	Passive	2) Inaccessible form	PS-012	Reinfo	Unknown					Y
SHP	Shape files	3D Geometry	Non-rendered	Static	Complex	Passive	2) Inaccessible form	PS-013	Reinfo	Unknown					Y
ODF	Open Document Format	Document	Rendered	Static	Complex	Passive		PS-014		Unknown					Y
AIFF	Audio Interchange File Format	Audio	Rendered	Static	Simple	Passive		PS-014	Migration	FFmpeg					Y
MP3	Mpeg-1/2 Layer - 3	Audio	Rendered	Static	Complex	Passive		PS-014	Migration	FFmpeg					Y
AAC	Advanced Audio Coding	Audio	Rendered	Static	Complex	Passive		PS-014		Unknown					Y
WMA	Windows Media Audio	Audio	Rendered	Static	Complex	Passive		PS-014	Migration	FFmpeg					Y
WAV	Waveform Audio File Format	Audio	Rendered	Static	Complex	Passive		PS-014	Migration	FFmpeg					Y
AVI	Audio video interleave	Audio	Rendered	Static	Complex	Passive		PS-014	Migration	FFmpeg					Y
MPEG-2	Moving Pictures Export Group v2	Video	Rendered	Static	Complex	Passive		PS-014	Migration	FFmpeg					Y
MPEG-4	Moving Pictures Export Group v4	Video	Rendered	Static	Complex	Passive		PS-014	Migration	FFmpeg					Y
MOV	Quicktime	Video	Rendered	Static	Complex	Passive		PS-014	Migration	FFmpeg					Y
WMV	Windows Media Video	Video	Rendered	Static	Complex	Passive		PS-014	Migration	FFmpeg					Y
EPS	Encapsulated Postscript	Image	Rendered	Static	Complex	Passive		PS-014		Unknown					Y
GIF	Graphics Interchange Format	Image	Rendered	Static	Simple	Passive		PS-014	Migration	ImageMagick					Y
PNG	Portable Network Graphics	Image	Rendered	Static	Simple	Passive		PS-014	Migration	ImageMagick					Y
PPT	MS Powerpoint	Document	Rendered	Static	Complex	Passive		PS-015	Migration	OpenOffice					Y
PST	Outlook - Personal Storage Table	Email	Non-rendered	Dynamic	Complex	Passive		PS-015		Unknown					Y
TXF	ASCII Text File	Document	Rendered	Static	Simple	Passive		PS-016	Reinfo	Registry					Y
RTF	Rich Text Format	Document	Rendered	Static	Complex	Passive		PS-016	Reinfo	Registry					Y
CSV	Comma Separated Variable	Data	Non-rendered	Static	Complex	Passive		PS-016	Reinfo	Registry					Y
SAV	Save File - Proprietary Binary forms	Data	Non-rendered	Static	Complex	Passive		PS-017	Reinfo	Registry					Y
RAV	ISIS RAW file	Data	Non-rendered	Static	Complex	Passive		PS-018	Migration	MAINTID	Y	N	N	Y	P
NXS	Nexus	Data	Non-rendered	Static	Complex	Passive		PS-018		Unknown					Y
SPV	SPSS, Statistical Package for Social Science	Data	Non-rendered	Static	Complex	Passive		PS-019		Unknown					Y
DTA	STATA	Data						PS-019		Unknown					Y
SAS	Statistical Analysis Software	Data	Non-rendered	Static	Complex	Active		PS-019		Unknown					Y

Figure 4 Test Environment capability matrix by Digital Object

Note: Figure 4 is a screenshot of Googledocs based online table that is maintained regularly by the APARSEN partners and is intended to be handed over to the VCoE at the end of the project.

## 4.5 CLASSIFICATION SCHEMES

The analysis of the preservation scenarios helped us to generate data from which to formulate classification schemes for digital objects, preservation techniques and tools, and identify any additional threats that need

to be taken into consideration as demonstrated in the work described above. To understand the approaches that were taken by WP14 to perform these classifications, we now look at each of these in turn.

#### 4.5.1 Types of digitally encoded information object

There are clearly many types of digital objects (byte streams) containing many types of information, encoded in many ways. There are of course many ways to try to classify these, and each has its merits.

During the development of the WP14 objectives one proposal for classifying digital objects that arose out of these discussions was suggested by David Giarretta<sup>[15]</sup> and took the approach that digital objects may be roughly classified according to orthogonal types of digitally encoded information, trying as far as possible to be (1) independent of the discipline to which the information is related and (2) probably useful in terms of digital preservation.

Note that this was not intended to be a way to uniquely classify all digital objects. Instead it was proposed just as a checklist to make sure we think about a whole variety of examples in our considerations rather than remain within the comfort zones of our own research domains and areas of interest. In this way we have at least some way of identifying the possible gaps in our coverage of preservation challenges. The key advantage is that if someone asks the VCoE to recommend a preservation tool to use for a completely new type of digital object then at least one can use this scheme to see whether there is some other example in the same group which we can compare with in terms of tools and techniques which have been used, and the degree of success.

The divisions proposed are:

1. *Rendered vs non-rendered*
2. *Static vs dynamic*
3. *Simple vs composite*
4. *Passive vs active*

Rendered objects are those for which one would be satisfied if in a number of years one could print, display or play i.e. render the digitally encoded information for human consumption, an example would be a JPEG file; non-rendered objects are all other kinds.

Static objects are those which do not change over time – again an image is one example; non-static are all others. (Note: Documents can be versioned but for a regulatory repository or national archive and would thus be considered static and immutable if they are to have any meaningful provenance. Context can have as much influence on the classification as the underlying technology.)

Simple objects are those which one normally treats as a whole – for example an image; composite objects are all others. (Note: There are exceptions that need to be resolved in order to validate and test this division: You can save Outlook e-mails in a single PST file or in many MSG files. Is one simple and the other not?)

Passive objects are ones to which things are done – for example a piece of software is used to display an image given a JPEG file; active objects are the others which normally do things using other inputs, for example the software which displays the JPEG file.

Based on this classification process a JPEG image may be described as being of type: Rendered, Simple, Static and Passive.

Note that it is not a classification cast in stone because any digital objects may be used in a variety of ways. There is nothing to prevent one from rendering any digital object, no matter how complex, as 1's and 0's. Or a text document could be regarded as a data file for the purposes of computer text analyses.

The purpose of the classification was simply to ensure that we force ourselves to think about different types of objects covering a wide range – it is **not** to provide a definitive classification of any given digital object.

Other approaches can be taken, as demonstrated by the output matrix shown in figure 3 where a simplified approach to object type classification was used. In this instance the primary nature of the information

object that is typically held in that file format was used to categorise the digital file objects. These 11 non-orthogonal categories were:

- 3D Geometry 3D collections of data points that represent spatial coordinates of objects
- Audio Music and sound files
- Compression Container formats that are used to reduce the size of the objects within them
- Data Typically binary data types requiring details of the associated semantics in order to be understood and used.
- Database Relation Database system data formats
- Document Office environment type files (word processing, spreadsheets, presentations)
- Email electronic messages
- Image 2D pictorial data
- Image & Data 2D pictorial data that has been enhanced with embedded metadata
- Video Moving images, films, TV broadcasts
- Website Snapshots of websites using tools like Heritrix, WinHTTrack etc

Our results demonstrate that there is no simple solution to how best to classify digital objects for preservation purposes at this time. This issue was discussed at length during this work and there was no clear 'winner' in terms of being able to identify a unifying classification scheme to be used for all digital objects. Moreover that is not the purpose of coming up with a classification scheme. The purpose is really to help the VCoE to provide advice about how to preserve objects in the case where we do not have an exact match in our evidence base.

We adopted the approach shown in Figure 4 where we include 2 such digital object classification schemes within our assessment matrix. The way the assessment matrix has been constructed allows multiple classification systems to co-exist in the data so that we can test the suitability of each proposed classification scheme as the data set evolves. Alternative classification schemes can be proposed and added relatively simply in to the matrix data set, allowing an assessment to be made of their ability to fit to the known set of file format data points captured to date. The consortium will make this data set available so that any additional classification scheme suggestions can be evaluated in this way.

#### 4.5.2 Types of preservation techniques

The digital preservation community recognises two well published techniques for providing long term access to the information objects within digital media, these are **Migration** and **Emulation**. A lesser recognised yet equally important approach should be that of proving adequate **Representation Information** about the context of the object, in order to allow designated communities of the future to understand and contextualise the “who what why where when and how” that resulted in the creation of the object in the first place.

In OAIS terms the Migration here is actually a Transformation, and Emulation is a type of Representation Information. One can therefore broadly summarise the options as (1) Transform and/or (2) add Representation Information.

(Note: in OAIS Terminology, Representation Information can take the form of almost any additional information that may be required to further document the object to make it accessible to either its original designated community or potentially any future designated community. As such, it becomes a ‘catch all’ for describing information that may be required for future re-use. No additional information may actually be necessary to further implement a specific preservation technique if enough detail has already been provided. )

### 4.5.3 Common Test Data Sets

As part of WP16 of the APARSEN project, the VCoE will make publically available a collection of datasets that the community may find useful as a test resource. There are a number of test corpora for this purpose hosted by organisations around the world. If time allows, it may be useful to assess these collections for data objects that may also be used to add further data points in to our matrix of unique test cases. In the interim, the test data sets referenced in the preservation scenarios were used by the project partners to perform tests within the APARSEN Test Environments that were made available for use within the project.



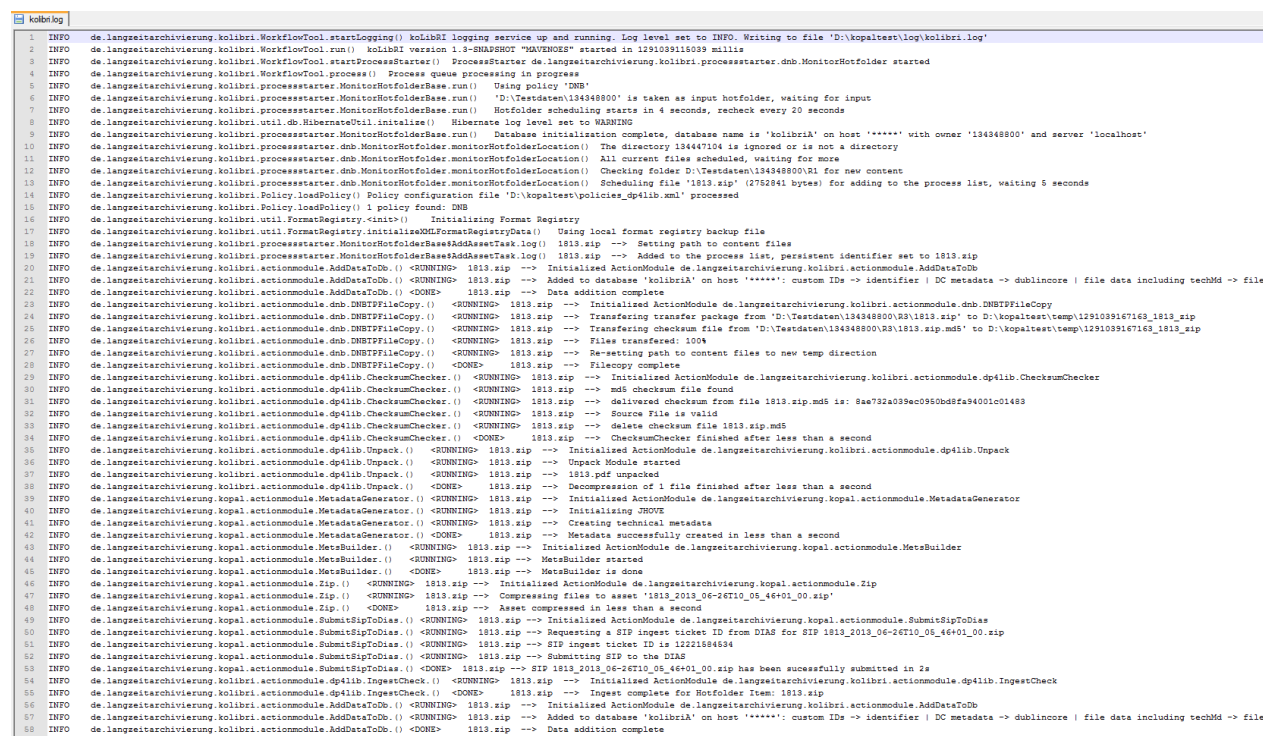
## 5 TESTING COMMON TEST DATA SETS IN THE APARSEN TEST ENVIRONMENT SYSTEMS

Once sources of common test data became available as part of the work to create Preservation Scenarios, the project partners that had test environment capabilities made these systems available for use by the project team to undertake tests and investigations using those data sets within their systems. The Preservica cloud based test Environment was also made available to all project partners in WP14 so that they could undertake their own test ingests, preservation actions, metadata edits, data access request, etc.

The projects partners performed tests to populate and further validate the analysis of the capabilities of Test Environments to preserve digital objects, presented in section: 4.4.3. The tests environments used in this activity are now discussed in more detail.

### 5.1 koLibRI (DNB)

koLibRI - (kopal Library for Retrieval and Ingest) represents a library of Java tools that have been developed for the interaction with the DIAS system of IBM within the kopal project [18]. koLibRI is a command line based tool (without a GUI), so to demonstrate its operation we have included in Figure 5 an extract from one of the INFO log files that was generated during a test ingest.



```
1 INFO de.langestarchivierung.kolibri.WorkflowTool.startLogging() kolibri logging service up and running. Log level set to INFO. Writing to file 'D:\kopaltest\log\kolibri.log'
2 INFO de.langestarchivierung.kolibri.WorkflowTool.run() kolibri version 1.9-08A0800T "MANUHOPE" started in 1291039115039 millis
3 INFO de.langestarchivierung.kolibri.WorkflowTool.startProcessStarter() ProcessStarter de.langestarchivierung.kolibri.processstarter.dnb.MonitorHotFolder started
4 INFO de.langestarchivierung.kolibri.WorkflowTool.process() Process queue processing in progress
5 INFO de.langestarchivierung.kolibri.processstarter.MonitorHotFolderBase.run() Using policy 'DNB'
6 INFO de.langestarchivierung.kolibri.processstarter.MonitorHotFolderBase.run() 'D:\Testdaten\134348800' is taken as input hotfolder, waiting for input
7 INFO de.langestarchivierung.kolibri.processstarter.MonitorHotFolderBase.run() Hotfolder scheduling starts in 4 seconds, recheck every 20 seconds
8 INFO de.langestarchivierung.kolibri.util.db.HibernateUtil.initialize() Hibernate log level set to WARNING
9 INFO de.langestarchivierung.kolibri.processstarter.MonitorHotFolderBase.run() Database initialization complete, database name is 'kolibriA' on host '*****' with owner '134348800' and server 'localhost'
10 INFO de.langestarchivierung.kolibri.processstarter.dnb.MonitorHotFolderBase.run() The directory 134447104 is ignored or is not a directory
11 INFO de.langestarchivierung.kolibri.processstarter.dnb.MonitorHotFolderBase.run() All current files scheduled, waiting for more
12 INFO de.langestarchivierung.kolibri.processstarter.dnb.MonitorHotFolderBase.run() Checking folder D:\Testdaten\134348800\RI for new content
13 INFO de.langestarchivierung.kolibri.processstarter.dnb.MonitorHotFolderBase.run() Scheduling file '1813.zip' (2752841 Bytes) for adding to the process list, waiting 5 seconds
14 INFO de.langestarchivierung.kolibri.Policy.LoadPolicy() Policy configuration file 'D:\kopaltest\policies\dp4lib.xml' processed
15 INFO de.langestarchivierung.kolibri.Policy.LoadPolicy() 1 policy found: DNB
16 INFO de.langestarchivierung.kolibri.util.FormatRegistry.<init>() Initializing Format Registry
17 INFO de.langestarchivierung.kolibri.util.FormatRegistry.initialize() Using local format registry backup file
18 INFO de.langestarchivierung.kolibri.processstarter.MonitorHotFolderBase.run() Added to the process list, persistent identifier set to 1813.zip
19 INFO de.langestarchivierung.kolibri.actionmodule.AddDataToDb() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kolibri.actionmodule.AddDataToDb
20 INFO de.langestarchivierung.kolibri.actionmodule.AddDataToDb() <RUNNING> 1813.zip --> Added to database 'kolibriA' on host '*****': custom IDs --> identifier | DC metadata --> dublincore | file data including techM4 --> file
21 INFO de.langestarchivierung.kolibri.actionmodule.AddDataToDb() <DONE> 1813.zip --> Data addition complete
22 INFO de.langestarchivierung.kolibri.actionmodule.dnb.DNBTFFileCopy() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kolibri.actionmodule.dnb.DNBTFFileCopy
23 INFO de.langestarchivierung.kolibri.actionmodule.dnb.DNBTFFileCopy() <RUNNING> 1813.zip --> Transferring transfer package from 'D:\Testdaten\134348800\RI\1813.zip' to D:\kopaltest\temp\1291039167163_1813.zip
24 INFO de.langestarchivierung.kolibri.actionmodule.dnb.DNBTFFileCopy() <RUNNING> 1813.zip --> Transferring checksum file from 'D:\Testdaten\134348800\RI\1813.zip.md5' to D:\kopaltest\temp\1291039167163_1813.zip
25 INFO de.langestarchivierung.kolibri.actionmodule.dnb.DNBTFFileCopy() <RUNNING> 1813.zip --> Files transferred: 100%
26 INFO de.langestarchivierung.kolibri.actionmodule.dnb.DNBTFFileCopy() <RUNNING> 1813.zip --> Re-setting path to content files to new temp direction
27 INFO de.langestarchivierung.kolibri.actionmodule.dnb.DNBTFFileCopy() <DONE> 1813.zip --> Filecopy complete
28 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.ChecksumChecker() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kolibri.actionmodule.dp4lib.ChecksumChecker
29 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.ChecksumChecker() <RUNNING> 1813.zip --> md5 checksum file found
30 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.ChecksumChecker() <RUNNING> 1813.zip --> delivered checksum from file 1813.zip.md5 is: Sae732a039ec095bdf84901c01483
31 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.ChecksumChecker() <RUNNING> 1813.zip --> Source File is valid
32 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.ChecksumChecker() <RUNNING> 1813.zip --> delete checksum file 1813.zip.md5
33 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.ChecksumChecker() <DONE> 1813.zip --> ChecksumChecker finished after less than a second
34 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.Unpack() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kolibri.actionmodule.dp4lib.Unpack
35 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.Unpack() <RUNNING> 1813.zip --> Unpack Module started
36 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.Unpack() <RUNNING> 1813.zip --> 1813.pdf unpacked
37 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.Unpack() <DONE> 1813.zip --> Decompression of 1 file finished after less than a second
38 INFO de.langestarchivierung.kopal.actionmodule.MetadataGenerator() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kopal.actionmodule.MetadataGenerator
39 INFO de.langestarchivierung.kopal.actionmodule.MetadataGenerator() <RUNNING> 1813.zip --> Initializing JHOVE
40 INFO de.langestarchivierung.kopal.actionmodule.MetadataGenerator() <RUNNING> 1813.zip --> Creating technical metadata
41 INFO de.langestarchivierung.kopal.actionmodule.MetadataGenerator() <DONE> 1813.zip --> Metadata successfully created in less than a second
42 INFO de.langestarchivierung.kopal.actionmodule.MetsBuilder() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kopal.actionmodule.MetsBuilder
43 INFO de.langestarchivierung.kopal.actionmodule.MetsBuilder() <RUNNING> 1813.zip --> MetsBuilder started
44 INFO de.langestarchivierung.kopal.actionmodule.MetsBuilder() <RUNNING> 1813.zip --> MetsBuilder is done
45 INFO de.langestarchivierung.kopal.actionmodule.Zip() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kopal.actionmodule.Zip
46 INFO de.langestarchivierung.kopal.actionmodule.Zip() <RUNNING> 1813.zip --> Compressing files to asse: '1813_2013_06-26T10_05_46+01_00.zip'
47 INFO de.langestarchivierung.kopal.actionmodule.Zip() <DONE> 1813.zip --> Files compressed in less than a second
48 INFO de.langestarchivierung.kopal.actionmodule.SubmitSipToDias() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kopal.actionmodule.SubmitSipToDias
49 INFO de.langestarchivierung.kopal.actionmodule.SubmitSipToDias() <RUNNING> 1813.zip --> Requesting a SIP ingest ticket ID from DIAS for SIP 1813_2013_06-26T10_05_46+01_00.zip
50 INFO de.langestarchivierung.kopal.actionmodule.SubmitSipToDias() <RUNNING> 1813.zip --> SIP ingest ticket ID is 12221884634
51 INFO de.langestarchivierung.kopal.actionmodule.SubmitSipToDias() <RUNNING> 1813.zip --> Submitting SIP to the DIAS
52 INFO de.langestarchivierung.kopal.actionmodule.SubmitSipToDias() <DONE> 1813.zip --> SIP 1813_2013_06-26T10_05_46+01_00.zip has been successfully submitted in 2s
53 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.IngestCheck() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kolibri.actionmodule.dp4lib.IngestCheck
54 INFO de.langestarchivierung.kolibri.actionmodule.dp4lib.IngestCheck() <DONE> 1813.zip --> Ingest complete for Hotfolder Item: 1813.zip
55 INFO de.langestarchivierung.kolibri.actionmodule.AddDataToDb() <RUNNING> 1813.zip --> Initialized ActionModule de.langestarchivierung.kolibri.actionmodule.AddDataToDb
56 INFO de.langestarchivierung.kolibri.actionmodule.AddDataToDb() <RUNNING> 1813.zip --> Added to database 'kolibriA' on host '*****': custom IDs --> identifier | DC metadata --> dublincore | file data including techM4 --> file
57 INFO de.langestarchivierung.kolibri.actionmodule.AddDataToDb() <DONE> 1813.zip --> Data addition complete
```

Figure 5 koLibRI Ingest Log

The sample ingest workflow that was run included the following steps:

1. TPFileCopy – copy the transfer package from hotfolder (FTP-location) into the koLibRI working area
2. ChecksumChecker – check the transfer package against a md5 checksum (ensure bitstream integrity)
3. Unpack – decompress the transfer package (normally a zip or tar container)
4. MetadataGenerator – generate technical metadata by using JHOVE
5. MetsBuilder – create mets.xml, a metadata file that contains the technical metadata and information about the content and the structure of the SIP

## 6. Zip - Compressing files to a SIP

7. SubmitSipToDias – transfer the generated SIP to our long term archive (DIAS), DIAS is located in Göttingen (not in Frankfurt)

8. IngestCheck – check whether DIAS has accepted the transferred SIP

9. AddDataToDb – store some administrative and workflow data into a database

The koLibRI system was tested with the test data from the DNB Preservation Scenarios, where Normalisation of ingested file formats to PDF/A is a commonplace activity.

## 5.2 Preservica (Tessella)

A test instance of Preservica, a cloud based digital preservation software as a service solution provided by Tessella, was deployed and made available to the project partners. Unlike koLibRI, this has a feature rich User Interface to allow an Archivist user to perform activities and functions according to the OAIS functions. The system comes preloaded with a wide variety of external third party tools to assist with the identification, characterisation, and property extraction for a good number of file format types, as well as a growing number of file migration tools.

This system was used to perform tests using the test data that is now publically available through the referenced links within our Preservation Scenarios. Figure 6 shows the Ingest workflow for data relating to the MIXED test data set from KNAW DANS, as provided by the link from Preservation Scenario PS-001.

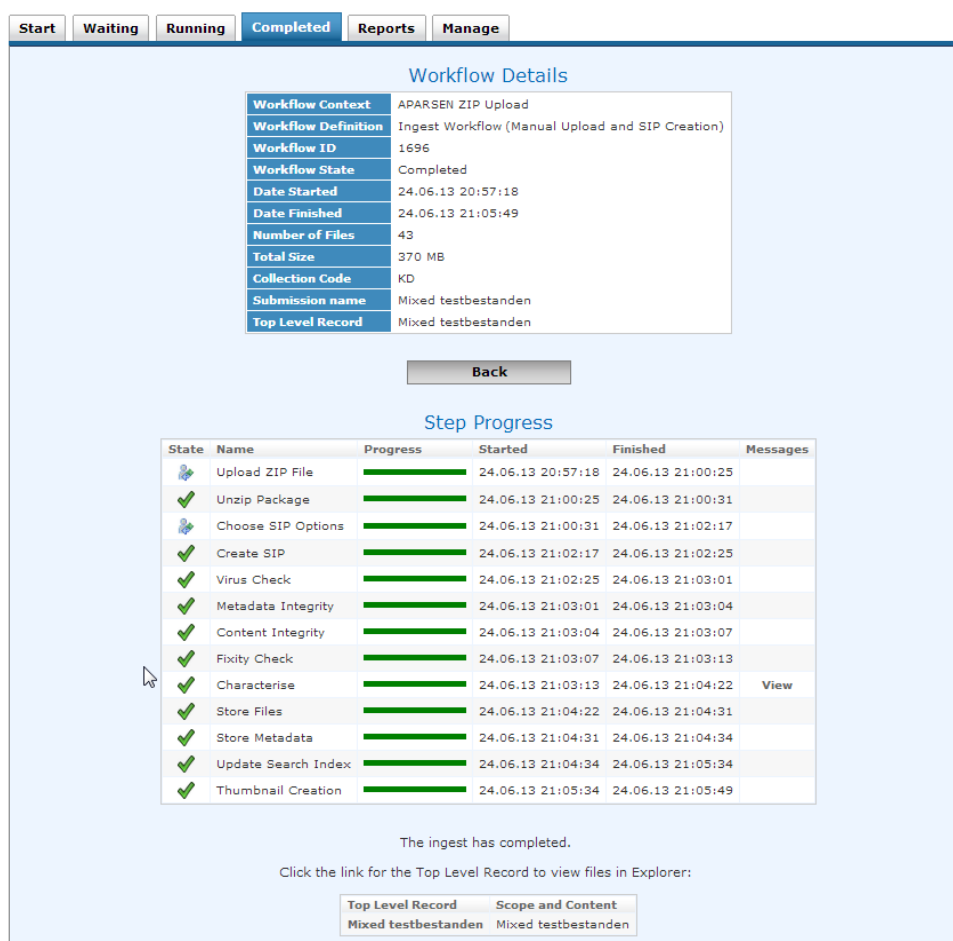
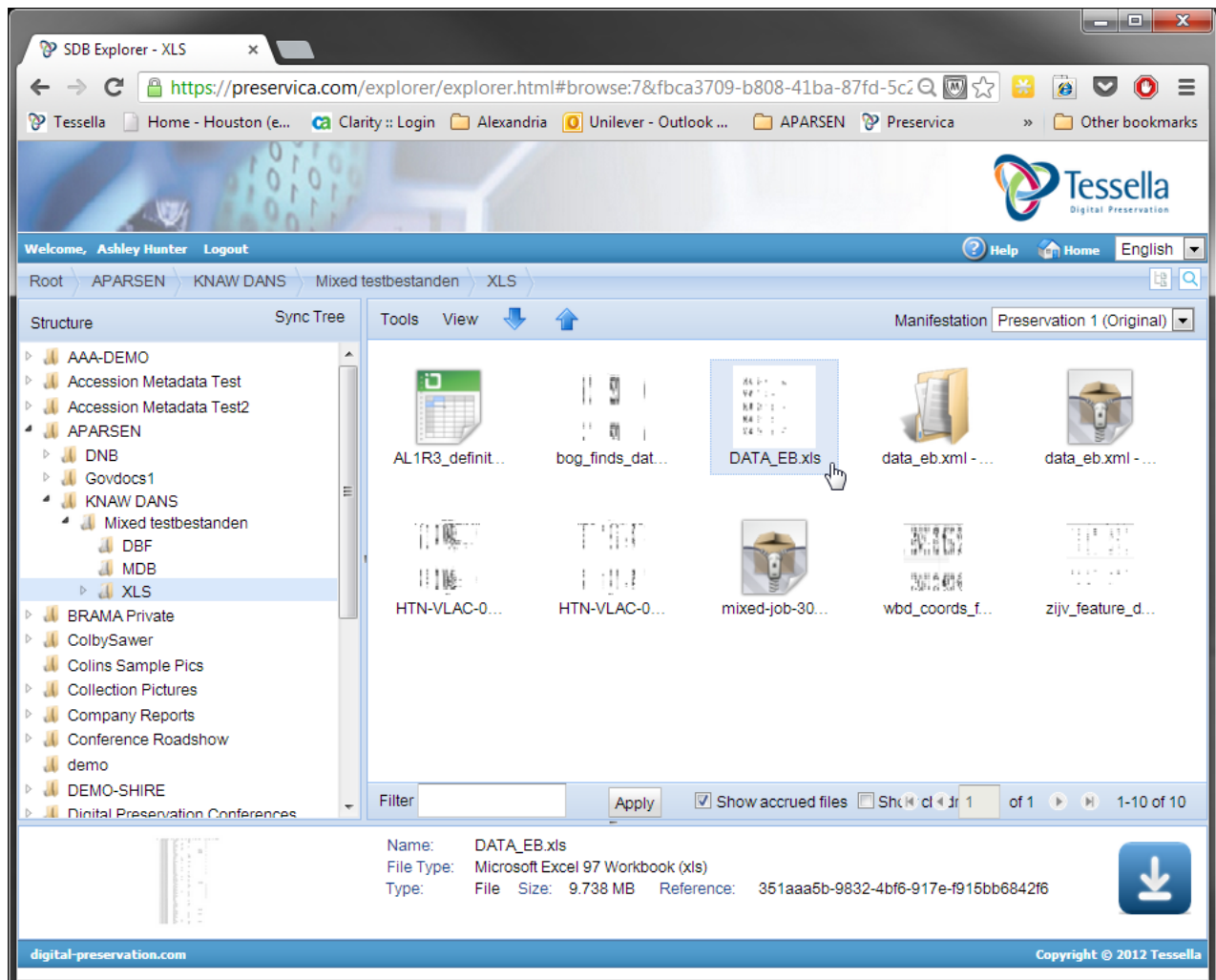


Figure 6 Preservica Ingest Workflow of MIXED Test Data



Once successfully ingested in to Preservica, it is possible to then browse to the content within the repository, as shown in Figure 7 below.



**Figure 7 Browsing the MIXED Test Data set in Preservica**

Once the digital objects have been successfully ingested in to the digital repository, it is then possible to run a preservation action on the content. In the example below, we decided to perform a migration action to transform the Excel objects (of varying Microsoft Excel version formats) to PDF/A objects.

The Preservica system supports the use of the Open Office XLS to PDF/A migration tool, so the user identifies the AIP that needs to undergo transformation, in this case, the top item in the File Set list shown in Figure 8. The migration pathways that are available for digital objects of file format type Microsoft Excel 97 Workbook (xls) 8 (PUID: fmt/61) are in this case the same as the later version, Microsoft Excel for Windows 2007 (xlsx) (PUID: fmt/214) and include the options to migrate to either PDF1.4, PDF/A, or Open Office ODS formats.

In this example, we select the Open Office PDF/A migration.(PUID: fmt/95).

### Preservation Plan Details

<b>Transformation Name</b>	Convert XLS to PDF
<b>Description</b>	Migrate .xls files to .pdf
<b>Date Created</b>	Friday, June 28, 2013 9:36 AM
<b>Number Of File Sets</b>	0
<b>Migration Pathway Role</b>	Preservation
<b>Usage</b>	Production

#### Select the File Sets for Transformation

Select	Collection Title	Catalogue Ref	Description	Manifestation	Number of Files	Files at Risk	Size	Size of Files at Risk	Details
<input checked="" type="checkbox"/>	KIAW DANS	Mixed testbestanden	Mixed testbestanden	1	43	7	388.74 MB	15.11 MB	<a href="#">Details</a>
<input type="checkbox"/>	Sample File Formats	Spreadsheets	Spreadsheets	1	0	0	0 B	0 B	<a href="#">Details</a>
<input type="checkbox"/>	BRAMA Private	documentation	documentation	1	0	0	0 B	0 B	<a href="#">Details</a>
<input type="checkbox"/>	BRAMA Private	documentation	documentation	1	0	0	0 B	0 B	<a href="#">Details</a>
<input type="checkbox"/>	Hagley Large Test	TV - 1	TV - 1	1	0	0	0 B	0 B	<a href="#">Details</a>
<input type="checkbox"/>	Company Reports	HSBC	HSBC	2	0	0	0 B	0 B	<a href="#">Details</a>
<input type="checkbox"/>	Financial Authorities	HK Securities and Futures Commission	HK Securities and Futures Commission	1	0	0	0 B	0 B	<a href="#">Details</a>
<input type="checkbox"/>	Govdocs1	732	732	1	0	0	0 B	0 B	<a href="#">Details</a>
<b>Totals</b>	<b>1</b>				<b>43</b>	<b>7</b>	<b>388.74 MB</b>	<b>15.11 MB</b>	

#### Select Transformation Pathways

Original Format	Resulting Format
Microsoft Excel 97 Workbook (xls) 8 (fmt/61)	<input type="button" value="Acrobat PDF/A - Portable Document Format 1a (fmt/95) - Open Office PDF/A"/> <input type="button" value="Acrobat PDF 1.4 - Portable Document Format 1.4 (fmt/18) - Open Office PDF"/> <input type="button" value="Acrobat 4 PDF/A - Portable Document Format 1.4 (fmt/18) - Open Office PDF/A"/> <input type="button" value="OpenDocument Spreadsheet 1.2 (fmt/295) - Open Office ODS"/> <input type="button" value="Do not migrate"/>
Microsoft Excel for Windows 2007 onwards (fmt/214)	

Figure 8 XLS to PDF/A Migration Plan

The Preservica system then retrieves the digital file objects from the repository store, performs the file format migration using the approved tool, and then re-ingests the new digital file objects in to the repository as a new manifestation of the original objects. The workflow for this process is shown in Figure 9 below:

Welcome, Ashley Hunter [Inbox \(0\)](#) [Logout](#) **SDB Digital Archive: Preservation** [Preservation](#) [Help](#) [Home](#) [English](#)

### Workflow Details

<b>Workflow Context</b>	Migration Workflow
<b>Workflow Definition</b>	Preservation Workflow
<b>Workflow ID</b>	1720
<b>Workflow State</b>	Completed
<b>Date Started</b>	28.06.13 09:36:32
<b>Date Finished</b>	28.06.13 10:08:22

### Step Progress

State	Name	Progress	Started	Finished	Messages
	New Details	<div style="width: 100%;"></div>	28.06.13 09:36:32	28.06.13 09:37:37	
	Pick Formats at Risk	<div style="width: 100%;"></div>	28.06.13 09:37:37	28.06.13 09:37:58	
	Pick File Sets	<div style="width: 100%;"></div>	28.06.13 09:37:58	28.06.13 09:51:55	
	Ready	<div style="width: 100%;"></div>	28.06.13 09:51:55	28.06.13 09:53:12	
	Migrate AIPs	<div style="width: 100%;"></div>	28.06.13 09:53:12	28.06.13 10:07:07	<a href="#">View</a>
	Launch Ingest Workflows	<div style="width: 100%;"></div>	28.06.13 10:07:07	28.06.13 10:07:07	

### Preservation Plan Details

<b>Transformation Name</b>	Convert XLS to PDF
<b>Description</b>	Migrate .xls files to .pdf
<b>Number Of File Sets</b>	1
<b>Migration Pathway Role</b>	Preservation
<b>Usage</b>	Production

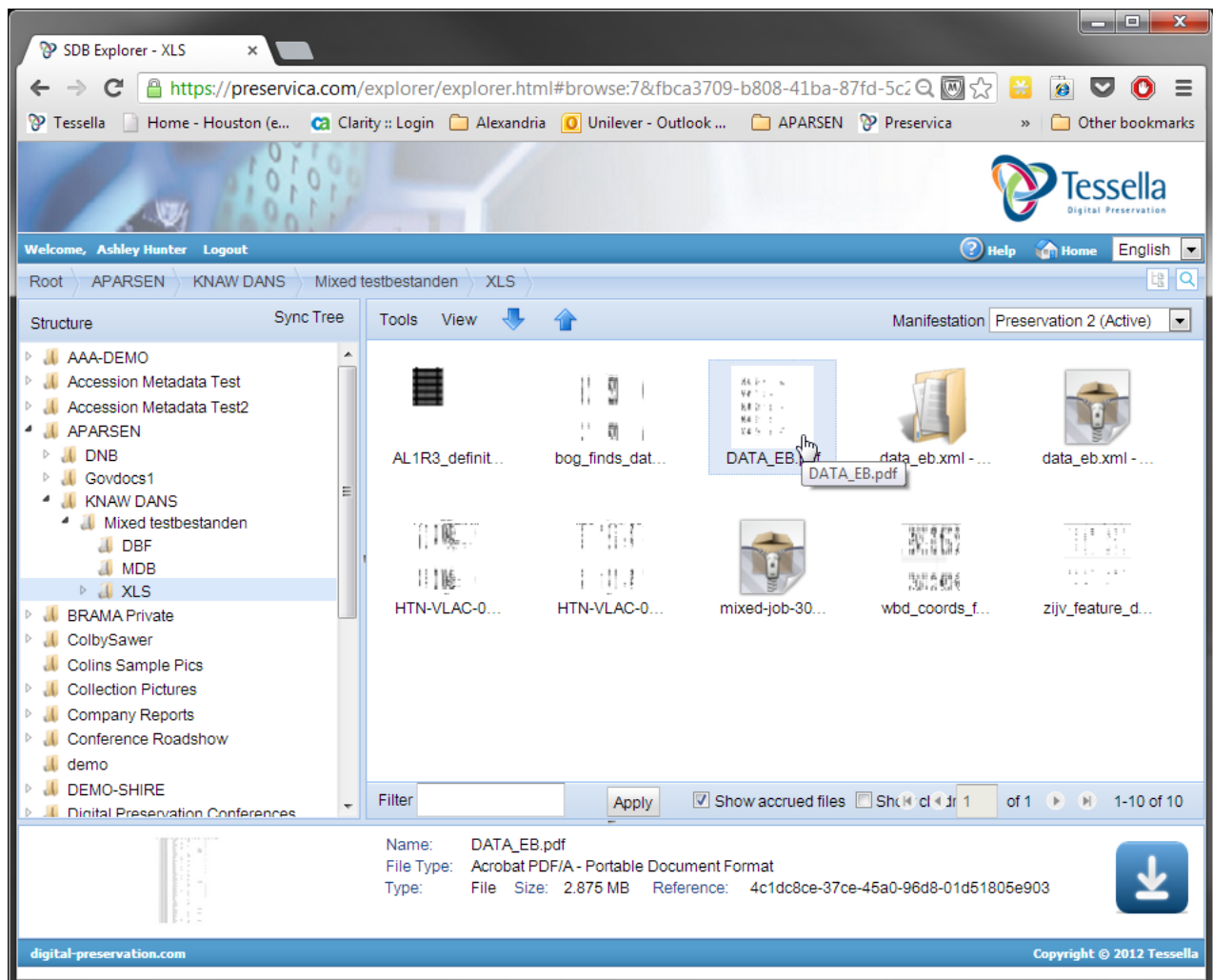
The transformation has completed.

Top Level Record	Collection Title	Outcome
Mixed testbestanden	KIAW DANS	Transformation Complete

digital-preservation.com Copyright © 2012 Tessella

Figure 9 Preservation Service workflow operating on MIXED test data

Once the new manifestations of the object have been created and re-ingested in to an AIP, it is then possible to go back in to the repository Explorer application and browse to the migrated objects. In Figure 10 you can now see that digital object “DATA\_EB.xls” shown in Figure 7 has now been migrated to “DATA\_EB.pdf”, demonstrating that the same digital information object has been transformed by the preservation action process.



**Figure 10 Migrated MIXED Test Data (XLS now PDF/A)**

The technical properties that have been extracted at each phase can also be compared to provide some qualitative assessment of the transformation process so that an Archivist or Digital Curator can validate the appropriateness of the transformation performed. As the transformation takes place in a managed repository, it is always possible to request to view the original manifestation, and perform additional preservation actions on the object to allow a comparison of the relative merits between two competing preservation action strategies.

The Preservica system is deployed with the ability to perform both ‘Test’ transformations, so the migrated files are created but not re-ingested back in to the Archive making them available for a limited period of review, and ‘Production’ transformations where the migrated objects are committed to the repository and are added to an appropriate AIP in the repository storage areas.

## 6 CONCLUSIONS

We have reviewed the test environment methodologies that have been used to date to demonstrate the effectiveness of tools and techniques to preserve digital information objects, and used this knowledge to generate a consensus view as to what a test environment must be amongst the consortium partners.

Through this process we initially encountered differences in opinion that in the majority were actually arising out of differences in our technical language relating to digital preservation, so to address this we developed the project glossary <sup>[17]</sup> that provides the project now with a single vocabulary of terms to use when communicating with each other about digital preservation issues and concepts.

The concept of preservation scenarios has been developed and used to capture the efficacy of tools and techniques for digital preservation, against changes in hardware, software, environment and knowledgebase of the Designated Communities.

These have been further extended to detail what new preservation scenarios are being envisioned by our community so as to provide a more general awareness of who has an active interest in which area, allowing us to further cooperate and unify our continuing work within APARSEN.

A simple matrix based assessment tool approach (described in section 4.4.1) has been developed to provide a generic testing methodology that allows digital preservation tools (and similarly digital preservation techniques) to be tested in a rigorous, impartial and replicable evidence based manner such that the results are comparable.

A process for being able to periodically snapshot the capabilities of one or more test environments has been demonstrated, allowing the VCoE to demonstrate both the maturing evolution of these test systems as they are enhanced to improve their preservation capabilities, and to use the combined results to highlight any areas for improvements where preservation techniques (and/ or tools) need to be added or further refined to provide better preservation strategies for those types of digital objects.

A version of the matrix assessment approach has been prototyped (see section 4.4.3) using test data provided by the consortium's partners to demonstrate the current tools and techniques that are in use for preserving digital objects now within the APARSEN community. It is anticipated that this approach will be further developed by the Virtual Centre of Excellence (VCoE) as one of the services that such a VCoE should offer to its end user community to assist with answering questions of a technical nature such as requests for evidence based advice about ways (tools or techniques) to preserve certain digital objects..

## APPENDIX A: PRESERVATION SCENARIOS

The preservation scenarios that have been provided to date by project participants during our work in WP14 have been reproduced and edited below for reference. The originals remain on our project wiki.

In summary, these Preservation Scenarios are (Click to follow link in document):

- KNAW-DANS: Migration of Datasets via an “intermediate XML format”
- DNB: Migration of static online publications (theses) from PDF to PDF/A
- DNB: File format validation of static online publications (newspapers in PDF/A-b)
- DNB: Migration of a sample of digitization master files from TIFF to Jpeg 2000
- DNB: Migrations of audio files from BWF to FLAC and WavPack
- DNB: Detection of protection measures in online publications
- DNB: Format identification and format validation of online publications
- STFC: Add Structural & Semantic Representation Information to Scientific Data
- STFC: Migration of binary data to a more commonly used & accessible format
- STFC: Migration to a more open format and adding more Representation Information
- STFC: Migrate binary data format to a standardised data description language (EAST)
- STFC: Migration of Contemporary Performing Arts data in to alternative file formats
- UNESCO: Migrate various World Heritage Site data into composite open formats
- Finnish National Digital Library: Evaluation of various ‘Office’ type migration paths
- Tessella: Preservation of Born-Digital Government Departmental Records
- Tessella: Digitisation of Paper based Government Records for future access
- Tessella: Scientific Datasets (ISIS Neutron and Muon Facility, STFC)
- UKDA Migration of Social Science and Humanities Data between statistical formats

### A.i.KNAW-DANS: Migration of Datasets via an “intermediate XML format”

Preservation Scenario ID	PS-001
Author	Valentijn Gilissen - DANS / Rene van Horik - DANS
Background	The DANS repository contains scientific datasets. A number of datasets deposited in the past are formatted in file formats that are now considered as obsolete. The task of curation of these datasets is now the responsibility of DANS. The preservation scenario consists of the curation activities to guarantee that the datasets are usable both now and in the future. DANS developed a framework to migrate binary formats into an "intermediate XML format" (for this the so-called MIXED framework was developed). Currently DANS is in the process of developing a stable, supported MIXED service as a web based service. It was agreed that the MIXED approach represents a developing preservation tool or technique that may be applicable to a wide number of formats.
Type of digital information	The sample datasets are in obsolete file formats (one could argue whether these formats are truly obsolete or not, but in any case they are not part of the list of "preferred formats" as supported by DANS). File formats include: <ul style="list-style-type: none"> <li>• dBase</li> <li>• MS Access</li> <li>• Excel</li> <li>• Dataperfect</li> </ul>
Link to sample data	A sample dataset consisting of files in dBase, MS Access and Excel format. DANS has also files on other obsolete formats, such as Dataperfect. <a href="https://www.dropbox.com/sh/9iyq658qas2uk97/Rper3C835e">https://www.dropbox.com/sh/9iyq658qas2uk97/Rper3C835e</a>

Threat(s) to the data	The content and structure of the files in obsolete formats are not easily accessible to researchers any more.
Usage	The files are used by researchers, and need to be accessible over the long term future to support the research efforts of future researchers.
Success Criteria	Content and structure of the datasets in obsolete binary formats are usable on current platforms, allowing the researchers to access and make structured queries of the data.

### A.ii. DNB: Migration of static online publications (theses) from PDF to PDF/A

Preservation Scenario ID	PS-002
Author	Sabine Schrimpf, DNB; Attila Zabos, DNB
Background	<p>The DNB has the legal mandate to collect, catalogue and archive publications published in Germany and in the German language. Since 2006, the legal mandate includes online publications. So far, DNB has focused mostly on static online publications that are collected in PDF format. We accept and archive any PDF version that is submitted to us.</p> <p>In this test scenario, we would like to convert a sample of our static online publication from PDF to PDF/A to evaluate the extent of information loss. Information loss is almost inevitable because PDF/A leaves out PDF features that are not suited for long term archiving. If possible, we would like to compare a) different converters and b) the results of migrations to PDF/A-1 vs. PDF/A-2.</p> <p>We can use our collection of online theses, because the data is openly available and has already been used in the SHAMAN context. It would be the easiest if the data could be manipulated on the SHAMAN test side, which contains 200-300 sample theses (see link to sample data).</p>
Type of digital information	<p>Doctoral theses and teaching theses of German universities (as opposed to theses that are released by publishers). <b>Size of collection:</b> ca. 102.000 objects. <b>Mean, Median, and Maximum size of digital object in the collection:</b> mean: ca. 22,83 MB, max. more than 1 GB. <b>Media formats included in the collection:</b> ca. 95% PDF plus a couple of thousand ZIP containers, mostly including PDFs and all kinds of various data file formats as attachments, e.g. .html, .xls, .jpg etc.).</p> <p>File formats include:</p> <ul style="list-style-type: none"> <li>• PDF</li> <li>• HTML</li> <li>• XLS</li> <li>• JPG</li> </ul>
Link to sample data	<a href="https://portal.dnb.de/opac.htm?query=catalog%3Ddnb.hss+AND+location%3Donlinefree&amp;method=simpleSearch">https://portal.dnb.de/opac.htm?query=catalog%3Ddnb.hss+AND+location%3Donlinefree&amp;method=simpleSearch</a> , <a href="http://services.dnb.de/googleScholarService/sitemap.html">http://services.dnb.de/googleScholarService/sitemap.html</a> ATTENTION: large file; The google sitemap is easier to process automatically, By appending "/34" to the object link, you can directly download the object
Threat(s) to the data	Due to the continuous technology and file format development, the main threat evaluated by this scenario is digital file format obsolescence. As a response to digital file format obsolescence, archived objects need to be migrated to enable long-term preservation. This test case scenario proposes to migrate PDF to PDF/A, which explicitly leaves out features of the standard PDF format that are not suited for long term archiving.
Usage	DNB reading room users (researchers, students, general public) that look for one particular thesis or that look for a particular topic, possibly covered by different theses. Readers are interested in all parts of the thesis, so correct preservation of embedded images, tables, indexes etc. is critical.
Success Criteria	Page breaks unchanged; Page structure unchanged; Images unchanged; Fonts embedded ; Colour profiles embedded. !! Tools for automatic verification of migration results and to report on migration quality are rare, therefore success of the migration process would have to be checked manually.

### A.iii. DNB: File format validation of static online publications (newspapers in PDF/A-b)

Preservation Scenario ID	PS-003
Author	Sabine Schrimpf, DNB; Robert Stoldt, DNB
Background	DNB has shifted its newspaper collection (comprising ca. 300 German newspaper titles) from print delivery to digital delivery. An external service provider harvests e-papers from the homepages of the publishing houses,



	converts them to PDF/A-1b and delivers them to the DNB. Routinely, several delivered issues fail the format validation due to configuration of or objects within the PDF-documents that only the publishers of the original content could solve. We are interested in looking at a subset of invalid PDF/A-1bs in more detail. We want to start with the originals of the erroneous files and convert them to PDF/A-1b in a controlled environment to find out which errors we can avoid under which conditions.
Type of digital information	German newspapers in PDF/A-1b. <b>Size of the collection:</b> 75.000 now, with an estimated growth of the collection by 90.000 objects per year. <b>Mean, Median, and Maximum size of digital object in the collection:</b> Mean = 4 MB, Median = 20 MB; max = 200 MB. <b>Media formats:</b> PDF/A-1b. For the APARSEN test case, we would define a reasonably sized sample of the collection. File Formats include: <ul style="list-style-type: none"> <li>• PDF/A-1b</li> </ul>
Link to sample data	Restricted access: This collection can only be accessed within the sites of the DNB.
Threat(s) to the data	Erroneous PDF/A-1b files may not be presented correctly in the future. Due to failed font embedding, inconsistent font widths entries, transparency in images and device-specific or corrupt colour profiles, the newspapers may look differently. DNB need to pick up these deficiencies in the format migration process as part of the Quality Assurance measures whilst the original formats are still accessible.
Usage	DNB reading room users (researchers, students, general public), some with targeted interest (search articles on a specific topic), some with a more general interest (how did a certain newspaper look like at a specific time?)
Success Criteria	Page breaks unchanged; Page structure unchanged; Images unchanged; Fonts embedded; Colour profiles embedded. !! Tools for automatic verification of migration results and to report on migration quality are rare; therefore success of the migration process would have to be checked manually.

#### A.iv. DNB: Migration of a sample of digitization master files from TIFF to Jpeg 2000

Preservation Scenario ID	PS-004
Author	Sabine Schrimpf, DNB; Robert Stoldt, DNB
Background	DNB has set up a small scale digitization workflow for books. Our target format is TIFF. The hierarchical structure of the TIFF files (reflecting the order of pages in the book) is expressed in METS. We are interested in converting a sample of digitization master files from TIFF to Jpeg 2000 to evaluate: <ul style="list-style-type: none"> <li>a) if we can save storage space without visible information loss and</li> <li>b) if the structure information stored in METS will remain useable.</li> </ul>
Type of digital information	Digitized books. <b>Size of collection:</b> 100 books (= 21.000 image files). <b>Media format:</b> TIFF, LZW-compressed, Metadata in METS container File Formats include: <ul style="list-style-type: none"> <li>• TIFF</li> <li>• JP2</li> <li>• LZW</li> </ul>
Link to sample data	<a href="http://d-nb.info/998814423/34">http://d-nb.info/998814423/34</a> This collection could be used outside the DNB. All books within it are out of copyright. They were digitized for special, prominent presence within The European Library.
Threat(s) to the data	a) Migration of digitized objects may become necessary in the future due to format obsolescence or to reduce storage space because new formats with higher lossless compression capabilities may come up. b) It is unclear how the structure information stored in METS containers, which arrange the pages of a scanned book in the correct order, will cope with such a migration.
Usage	General public with an interest in German literature.
Success Criteria	No information loss in the converted images; Colour profiles (if there are any) correctly transferred; No loss of structure information within the METS file.

#### A.v. DNB: Migrations of audio files from BWF to FLAC and WavPack

Preservation	PS-005
--------------	--------

Scenario ID	
Author	Sabine Schrimpf, DNB; Robert Stoldt, DNB
Background	The DNB hosts the German Music Archive, which has the legal mandate to collect all music published in Germany. We are in the process of gradually digitizing all audio CDs of the German Music Archive since 1983. We are interested in converting our audio files from uncompressed Broadcast wave (BWF) to FLAC and WavPack. The (potential) loss of metadata stored in the header of the audio file during migration is an important factor in this scenario.
Type of digital information	Digitized music. <b>Size of the collection:</b> 18.000 now, about 430.000 by the end of 2014 with an estimated growth of the collection by 100.000 objects per year. <b>Mean, Median, and Maximum size of digital object in the collection:</b> Median = 525 MB; max = 800 MB. <b>Media formats:</b> Uncompressed Broadcast Wave. For the APARSEN test case, we would define a reasonably sized sample of the collection. File Formats include: <ul style="list-style-type: none"> <li>• BWF</li> <li>• FLAC</li> <li>• WavPack</li> </ul>
Link to sample data	Restricted access: This collection can only be accessed within the sites of the DNB.
Threat(s) to the data	Similar as with the digitized books, we might be forced to migrate our audio data in the future and it is unclear to us how the metadata stored in the .wav header will cope with the migration.
Usage	DNB reading room users (researchers, students, general public), some with a more general interest (which reading errors occurred within the migration process of the CD?)
Success Criteria	All metadata (including the BWF-metadata in the extension chunk) is transferred from the .wav header to the new format.

## A.vi. DNB: Detection of protection measures in online publications

Preservation Scenario ID	PS-006
Author	Stefan Hein, DNB; Sabine Schrimpf, DNB
Background	The DNB has the legal mandate to collect, catalogue and archive publications published in Germany and in German language. Since 2006, the legal mandate includes online publications. So far, DNB has focused mostly on static online publications that are collected in PDF format. We accept and archive any PDF version that is submitted to us.  In this test scenario we would like to examine a sample data set of online publications in PDF with regard to the existence of protection measures like password protection and print or copy restrictions. The DNB has made the experience that these measures are often used by publishers when they deliver online publications to the library. Our current approach for the ingest workflow is to refuse "DRM suspicious" material and to give the publisher or the delivering institution the possibility to remove the protection for a second delivery. We do not modify the publication ourselves.
Type of digital information	Doctoral theses and teaching theses of German universities (as opposed to theses that are released by publishers). <b>Size of collection:</b> ca. 102.000 objects. <b>Mean, Median, and Maximum size of digital object in the collection:</b> mean: ca. 22,83 MB, max. more than 1 GB. <b>Media formats</b> included in the collection: ca. 95% PDF plus a couple of thousand ZIP containers, mostly including PDFs and all kinds of various data file formats as attachments, e.g. .html, .xls, .jpg etc..- File Formats include: <ul style="list-style-type: none"> <li>• PDF</li> <li>• HTML</li> <li>• XLS</li> <li>• JPG</li> </ul>
Link to sample data	<a href="https://portal.dnb.de/opac.htm">https://portal.dnb.de/opac.htm</a>
Threat(s) to the data	Protected publications are likely to cause difficulties throughout the preservation process and access now and in the future. A password protected file is hard to convert in another format. Also the success of preservation actions by emulation is threatened.
Usage	DNB reading room users (researchers, students, general public) that look for one particular thesis or that look



	for a particular topic, possibly covered by different theses. Here the DNB is willing to ensure a comfortable and barrier-free access to its publications – now and of course in the future.
Success Criteria	Successful detection of protected publications and report about the kind of the protection measure.

### A.vii.DNB: Format identification and format validation of online publications

Preservation Scenario ID	PS-007
Author	Stefan Hein, DNB; Sabine Schrimpf, DNB
Background	<p>The DNB has the legal mandate to collect, catalogue and archive publications published in Germany and in German language. Since 2006, the legal mandate includes online publications. So far, DNB has focused mostly on static online publications that are collected in PDF format. We accept and archive any PDF version that is submitted to us and also all kinds of various data file formats as attachments, e.g. .html, .xls, .jpg etc.</p> <p>In this test scenario we would like to identify and validate the format of a file that is taken from a sample data set of online publications.</p>
Type of digital information	<p>Doctoral theses and teaching theses of German universities (as opposed to theses that are released by publishers). <b>Size of collection:</b> ca. 102.000 objects. <b>Mean, Median, and Maximum size of digital object in the collection:</b> mean: ca. 22,83 MB, max. more than 1 GB. <b>Media formats</b> included in the collection: ca. 95% PDF plus a couple of thousand ZIP containers, mostly including PDFs and all kinds of various data file formats as attachments, e.g. .html, .xls, .jpg etc..-</p> <p>File Formats include:</p> <ul style="list-style-type: none"> <li>• PDF</li> <li>• HTML</li> <li>• XLS</li> <li>• JPG</li> </ul>
Link to sample data	<a href="https://portal.dnb.de/opac.htm">https://portal.dnb.de/opac.htm</a>
Threat(s) to the data	Without file format identification the archiving institution has no information what kinds of objects are archived. This information is essential to establish an efficient migration strategy. This is the only way to address the objects which are concerned by a format migration at all. Moreover, format validation helps to sort out erroneous objects which may not be presented in the future.
Usage	DNB reading room users (researchers, students, general public) that look for one particular thesis or that look for a particular topic, possibly covered by different theses. Here the DNB is willing to ensure a comfortable and barrier-free access to its publications – now and of course in the future.
Success Criteria	Successful identification of the file format. Report the file format by using controlled namespaces like PRONOM or mimeType. Identification and report of violations against the specification of the identified file format by using validation tools like JHOVE.

### A.viii. STFC: Add Structural & Semantic Representation Information to Scientific Data

Preservation Scenario ID	PS-008
Author	David Giaretta STFC/APA
Background	Scientific laboratories collect and archive data from various sources. It is important for other scientists to be able to use that data to, for example, reprocess to confirm some published results or, probably more frequently, to analyse in new ways and/or to combine with data from other sources.
Type of digital information	<p>MSST radar data</p> <p>File formats include:</p> <ul style="list-style-type: none"> <li>• Binary Data (Proprietary file formats)</li> <li>• XML (or other structured text based format for describing structure &amp; semantics)</li> <li>• Website Archives (may be in compressed formats such as WARC, etc)</li> <li>• Software Applications (.exe and any supporting .ini / .cfg files etc)</li> </ul>

<b>Threat(s) to the data</b>	(1) Existing software libraries to access the data may be unusable (2) the structure of the data i.e. the format, may be forgotten (3) the semantics i.e. the meaning of the individual numbers may not be understood e.g. this number is a temperature measured in degrees C, measurement from (...) using a type of thermometer and the raw values were turned into degrees C using this (...) calibration curve.
<b>Designated Community</b>	Scientists involved in atmospheric physics
<b>Preservation Technique</b>	Create a fairly complete Representation Information Network to analyse risks. Create additional Representation Information - Structure and Semantic. Save the BADC web site for information about the measurement instrument. Save the software and associated algorithms
<b>Usage</b>	
<b>Success Criteria</b>	Ask members of the Designated Community and those from a closely related discipline if they would be able to sensibly use the data, given that additional Representation Information

### A.ix. STFC: Migration of binary data to a more commonly used & accessible format

<b>Preservation Scenario ID</b>	PS-009
<b>Author</b>	David Giaretta STFC/APA
<b>Background</b>	<p>The International Ultraviolet Explorer (IUE) was an astronomical satellite which obtained UV spectra of tens of thousands of astronomical objects. The data for one object consists of an image which has one or more spectral orders (each as a band across the image). The raw data is processed through several stages, first correcting photometrically and geometrically and then extracting the spectrum.</p> <p>The original IUE processing software created what is called VICAR file formats; the VICAR files had binary (i.e. non text) header files followed by data with, in the processed files, various quality flags to show where pixels cannot be trusted.</p> <p>Since the launch of IUE the FITS astronomical format has become the accepted astronomical format for data. What actually happened was that it was decided to create the "IUE Final Archive" was a way to ensure that the scientific data collected by IUE would not be lost. In the process of doing this a new way to process the raw data in a more accurate was developed however putting this to one side there were a number of interesting considerations.</p> <p>* the binary header data encoded temperatures and voltages of the instrument - in FITS the headers are essentially text. Therefore what was done was to convert the various temperatures, voltages into physical into (degrees K or Volts) and these were put into the header as numerical data in characters. The FITS headers take the form of NAME = VALUE where NAME is limited to 8 characters. Therefore for each value a name must be created - limited to 8 characters.</p> <p>* The quality flags needed to be converted into separate images within the FITS file. The meaning of each quality image pixel value needed to be defined (e.g. 1 means pixel was saturated to has no meaningful data, 2 means the calculated value is affected by reseau marks and so should be regarded with suspicion...)</p>
<b>Type of digital information</b>	<p>Astronomical data</p> <p>File Formats include:</p> <ul style="list-style-type: none"> <li>• VICAR</li> <li>• FITS</li> <li>• Binary Data (Proprietary formats)</li> <li>• Text (accessible, but may be in an unknown form or structure)</li> <li>• Image formats holding flag based calibration data</li> <li>• XML (to maintain additional representation information)</li> </ul>
<b>Threat(s) to the data</b>	<p>1) Existing software libraries to access the data may be unusable,</p> <p>2) the structure of the data i.e. the format, may be forgotten</p> <p>3) the semantics i.e. the meaning of the individual numbers may not be understood e.g. this number is a temperature measured in degrees C, measurement from (...) using a type of thermometer and the raw values were turned into degrees C using this (...) calibration curve.</p>
<b>Designated Community</b>	Astronomers
<b>Preservation Technique</b>	<p>Transformation into a new format - FITS.</p> <p>However as noted above, even ignoring the processing algorithms, this is not a simple transformation. A great</p>

	deal of new semantics must be passed on to the users and the relationship between the various components within the new format file must be explained in order for the digitally encoded information.
Usage	Astronomers access the FITS file and use a variety of different suites of astronomical software to extract new astronomical information, perhaps combining with data from other sources.
Success Criteria	The success of the preservation activity can be seen from the fact that the IUE data is still used by astronomers, 33 years after the launch of the satellite and 13 years since the satellite was closed down.

### A.x. STFC: Migration to a more open format and adding more Representation Information

Preservation Scenario ID	PS-0010
Author	David Giaretta STFC/APA
Background	Astronomical data is often in the form of tables. These vary from simple text files, with a few lines of headers followed by columns of numbers and/text, to components in FITS files either as text or binary. Although the column headings often seem simple e.g. "VsubJ" - Johnson visual magnitude. However the accurate interpretation of the data values one needs to know the filter transmission curves. Moreover in some cases the names can be misleading.
Type of digital information	Tabular data containing data from various sources. File formats include: <ul style="list-style-type: none"> <li>FITS</li> <li>VOTable (XML)</li> </ul>
Threat(s) to the data	(1) Existing software libraries to access the data may be unusable (2) the structure of the data i.e. the format, may be forgotten (3) the semantics i.e. the meaning of the individual numbers may not be understood e.g. this number is a temperature measured in degrees C, measurement from (...) using a type of thermometer and the raw values were turned into degrees C using this (...) calibration curve.
Designated Community	Mostly astronomers
Preservation Technique	Several preservation techniques have been used. In many cases the original data format has been Transformed to FITS. However since XML became popular it was decided to create a table format which would be better suited to exchange via Web Services. This new format was called VOTable. It was also believed that as XML it would have some advantages for preservation.  In many cases a Uniform Column Descriptor (UCD) has been created to capture the semantics of the column to give some idea of which column could sensibly be combined.
Usage	The various encodings of data must be able to be understood and used. IN particular the various datasets must be combinable sensibly. One way this is done is to virtualise the various encodings into the Java AbstractTableModel and a variety of specialisations which capture additional information and semantics. This allows data from multiple sources and in multiple formats to be combined.
Success Criteria	Astronomers can use and understand the data that is encoded, and in particular can combine data from various sources.

### A.xi. STFC: Migrate binary data format to a standardised data description language (EAST)

Preservation Scenario ID	PS-011
Author	David Giaretta STFC/APA
Background	The European Space Agency launches a number of satellites which captured data and, after processing these were stored in Common Data Format (CDF). CDF was a format which originated in NASA to encode data which was repeatedly measured on a grid. The internal format is very complex and was not described anywhere except in the access libraries - which itself was very complex. However NASA decided that it would no longer support the CDF access software at some point in the future. ESA had what was, at that time, a huge amount of data in this format.

	Moreover the CDF file format had certain limitations and so a number of "conventions" were imposed on the CDF files which meant that analysis software had some built-in semantics which the "standard" CDF software did not know about.
Type of digital information	Solar Terrestrial Physics (STP) measurements obtained from a number of satellites. File formats include: <ul style="list-style-type: none"> <li>• CDF</li> <li>• Described content in the EAST data description language (Binary, Ascii, Database forms)</li> </ul>
Threat(s) to the data	1) Existing software libraries to access the data may be unusable, 2) the structure of the data i.e. the format, may be forgotten 3) the semantics i.e. the meaning of the individual numbers may not be understood e.g. this number is a temperature measured in degrees C, measurement from (...) using a type of thermometer and the raw values were turned into degrees C using this (...) calibration curve.
Designated Community	STP scientists
Preservation Technique	Transforming to another format was a possibility although the volume and the hidden semantics made this unattractive. Instead it was decided to ensure the long-term usability of the data by describing it using the EAST language. This first required the CDF software team to write a fairly full description of the CDF internal structures which could then be described in EAST. This gave ESA the confidence to continue to keep the data in CDF format for a considerable time.  Eventually technology changes in storage and the emergence of new analysis tools and formats meant that at least some of the data was transformed, but the hidden semantics had to be exposed.
Usage	The data was used in a variety of analysis tools.
Success Criteria	The ability of scientists to use historical data and combine this with new data to perform enhanced analyses.

#### A.xii. STFC: Migration of Contemporary Performing Arts data in to alternative file formats

Preservation Scenario ID	PS-012
Author	David Giaretta STFC/APA
Background	New contemporary performing arts composition must be able to be re-performed over time.
Type of digital information	Consists of musical composition perhaps PDF) plus software (known as patches) which changes the music e.g. adding reverberation etc in a complex workflow using proprietary software and hardware. The patches are essentially subroutines which run in the proprietary software. File Formats include: <ul style="list-style-type: none"> <li>• PDF</li> <li>• Coordinate 3D (C3D)</li> <li>• Tracked Row Character (TRC)</li> <li>• BioVision Hierarchical Data (BVH)</li> <li>• Hierarchical Translations and Rotations (HTR)</li> </ul>
Link to sample data	Sample Data may be available from the Cultural Heritage institutions involved in the original CASPAR work.
Threat(s) to the data	The interaction and timing of the interactions between the music and computer effects must be maintained. The computer generated effects must be maintained despite the lack of ability to run the software patches and the availability of the hardware.
Designated Community	Performers of this type of music plus their musical assistants.
Usage	The performer and musical assistant must be able to re-perform the music.
Success Criteria	The music must be able to be re-performed to the satisfaction of the composer, if available, or to the performer.

#### A.xiii. UNESCO: Migrate various World Heritage Site data into composite open formats

Preservation Scenario ID	PS-013
Author	David Giaretta STFC/APA
Background	World Heritage Sites (WHS) are documented using a variety of techniques including laser-scans, satellite observations, etc captures as a variety of data files. The state of the site at one point in time must be able to be compared with other measurements later on in order to determine if the site has deteriorated
Type of digital information	ESRI shape files. File formats included: <ul style="list-style-type: none"> <li>• 3D point clouds (imp, dxf, dwg)</li> <li>• Elevation grids (agr, bt)</li> <li>• 3D meshes (mdl, vrml, v3d)</li> <li>• Textured 3D models (max, pmr, ive, osg)</li> <li>• Satellite data (ers, ecw)</li> <li>• GPS data, maps (txt, apm, shp)</li> <li>• Digital images (targa, jpeg, tiff, png, psd, bmp, gif, dds)</li> <li>• an elevation grid of the site (agr / grd)</li> <li>• a map of the site contours (shp)</li> </ul>
Link to sample data	Sample data may be available on request from CASPAR participants
Threat(s) to the data	The ESRI software needed may become unavailable in future years.
Designated Community	UNESCO WHS experts.
Usage	The data from one time must be able to be compared with measurements taken with different instruments in order to see if the site has deteriorated.
Success Criteria	The measurements are successfully compared and the older data does, in spot checks, agree with the data values extracted by the original s/w.

#### A.xiv. Finnish National Digital Library: Evaluation of various 'Office' type migration paths

Preservation Scenario ID	PS-014
Author	Pekka Mustonen, CSC
Background	<p>The Finnish National Digital Library project - launched by the Finnish Ministry of Education and Culture in year 2008 - brings the achievements of culture and science to general public. The aims of the NDL project are improving availability and usability of the key national information resources of libraries, archives and museums in information networks, and the development of long-term preservation solutions for digital cultural heritage content data objects. The long-term preservation section of the NDL project has prepared a plan describing the model for centralized national long-term preservation solution for the digital objects of memory organisations responsible for the preservation of cultural heritage.</p> <p>In National Digital Library project file formats are divided into "acceptable for preservation" and "acceptable for transfer" -categories. For example, file formats used in MS Office suite are considered "acceptable for transfer" but these files will be converted into long-term preservation format before being archived.</p> <p>We want to study various migration paths from "acceptable for transfer" to "acceptable for preservation" to be able to instruct depositors in preservation planning.</p>
Type of digital information	<p>According to a recent study, the current number of digital objects to be deposited to the NDL long-term preservation system is roughly 687 000 000 (2500TB), and the size of the collection is estimated to be 1 458 000 000 objects (5700TB) in 2015 when the system will be in production (obviously only a tiny share of this will be available for testing).</p> <p>Possible migration paths can be any of the following (Note: also scenarios by DNB are very relevant to us):</p> <p>File Formats include:</p> <ul style="list-style-type: none"> <li>• "Text": <ul style="list-style-type: none"> <li>• Acceptable for transfer: Microsoft Word for Windows Document</li> <li>• Acceptable for preservation: Open Document Format (ODF), PDF for long-term preservation (PDF/A)</li> </ul> </li> <li>• Audio: <ul style="list-style-type: none"> <li>• Acceptable for transfer: Audio Interchange File Format (AIFF), Mpeg-1 layer-3, Mpeg-2 layer-3 (MP3), Mpeg-4 aac – advanced audio coding (AAC), Window media audio</li> </ul> </li> </ul>

	<ul style="list-style-type: none"> <li>Acceptable for preservation: Broadcast Wave Format (BWF), Waveform Audio Format (WAV), AIFF (PCM-coded), AAC,</li> <li>Video:</li> <li>Acceptable for transfer: Audio video interleave (AVI), Moving pictures expert group (MPEG-2), Moving pictures expert group (MPEG-4), Quicktime (MOV), Windows media video (WMV)</li> <li>Acceptable for preservation: JPEG 2000 MXF or Motion JPEG 2000</li> <li>Still images</li> <li>Acceptable for transfer: Encapsulated postscript (EPS), Graphics interchange format (GIF), Portable network graphics (PNG)</li> <li>Acceptable for preservation: Joint photographic experts group (JPEG), Joint photographic experts group jpeg 2000 (JP2), Tagged image file format (TIFF)</li> </ul>
Link to sample data	Not yet available.
Threat(s) to the data	Information loss during the conversion
Usage	(Mainly) producers
Success Criteria	Object properties are preserved with satisfying quality

## A.xv. Tessella: Preservation of Born-Digital Government Departmental Records

Preservation Scenario ID	PS-015
Author	Ashley Hunter, Tessella Ltd
Background	National Archives are founded on the basis that they must provide long term access to a wide variety of government records. Simple cases may include keeping the minutes of specific departmental meetings in PDF or DOC safe and available for access for a given period of time. More complex data types may be digital objects like databases, websites, CAD files etc. Each digital information object may have a defined closure period during which access permissions can only be granted to specific authorised individuals, but after this period, the record becomes open for wider or even public consumption. The digital records may well only reach the National Archive after a specific holding period (typically 10-20yrs) has expired within the issuing department's own Content Management System, setting the status of the material as "Archival". This further complicates the task of the National Archive as it strives to ensure that records do not become obsolete before they are even allowed to be transferred out of the issuing department to the Archival Repository. Records that are transferred to the Archive may then remain closed for significant periods (e.g. 50yrs, etc) or until such a reasonable time has expired that anyone referenced in the record is likely to be deceased.
Type of digital information	<p>File formats include:</p> <ul style="list-style-type: none"> <li>MS Word documents(DOC)</li> <li>Excel Spreadsheets(XLS)</li> <li>Powerpoint presentations (PPT)</li> <li>Outlook PST files</li> <li>Simple Text Files (TXT)</li> <li>Presentation copies of printed documents (PDF, PDF/A)</li> </ul>
Link to sample data	Tessella to provide a test set of typical files (not real, made up - likely from the DROID test file corpora) for testing purposes
Threat(s) to the data	<p>The main purpose of the Archive is to provide long term search and access of records for the various approved user communities, and to this end the Archive must defend against passive preservation issues, relating to:</p> <ol style="list-style-type: none"> <li>(1) keeping the digital objects safe from 'bit rot' or 'data decay' on the primary storage media.</li> <li>(2) Complex container &amp; compression formats, or files with embedded digital certificates &amp; signatures can also introduce further risk to the ability to access the information objects in the future, as access may be dependent on having a tool available that knows how to unpack or uncompress the objects, or be dependent on performing a verification process with a 3rd party entity that may not be available anymore.</li> <li>(3) The Archive may want to provide multiple versions of its digital objects, rendering them in different formats for different access purposes (free low grade manifestation, and commercially available full high-resolution manifestations), and to this end Archives can use Active Preservation methods including file format migration to provide these various manifestation types.</li> <li>(4) Representation Information about the record's provenance and authenticity is provided through arrangement of the digital objects in to a hierarchy of collections, with Descriptive Metadata provided according to agreed</li> </ol>



	metadata schemas. Commonly these are bespoke to each Archive, rather than using standard Schemas such as METS, MODS, EADS, PREMIS, etc.
<b>Designated Community</b>	Initially the submitting Government Department and any other affiliated departments and organisations will have access to the material and over time this access will be relaxed for some of the records leading to wider or even public access to records for use by policy investigators, historians, and the general public.
<b>Preservation Technique</b>	<p>(1) The standard approach to defend against bit rot is to keep multiple copies of the AIP on different storage media, and to periodically test the integrity of these objects against their known fixity/checksum values at the time of ingest. Where corruptions are found, the system should notify the Archive staff so that the corrupted file can be replaced from one of the other AIP copies that has recently passed the integrity test.</p> <p>(2.a) During ingest, unpack and uncompress all digital objects and characterise these objects as individual digital objects in their own right, extracting descriptive, administrative, technical metadata and structural information in relation to any other extracted digital objects.</p> <p>Remove (2.b) Digital certificates and signatures from objects where possible and re-assert provenance and authenticity from within the Archival system itself (i.e. make it too have a discernable and guaranteed provenance and authenticity) whilst maintaining references to the original certification system.</p> <p>(3) Use file migration techniques to provide digital information in alternative file formats. This process may in itself be a lossy process, so some process of quality assessment and validation needs to be applied to ensure that the appropriate level of quality is maintained following each format migration. Several migrations may be required over time to provide the Designated Community with the information that they want to access in a format suitable to their needs.</p> <p>(4) Provide methods for translating between metadata schema definitions via code or XSLT, to enable exchange of networks of representation information. Facilitate further integration with catalogue collection systems through the use of OAI-PMH exchange protocols.</p>
<b>Usage</b>	Archive staff, submitting government department staff via access requests, public access request information
<b>Success Criteria</b>	Digital objects remain available to their designated community through 'Search' &/or 'Browse' functionality, and are directly accessible to these users in forms and ways that are meaningful to them and can facilitate their re-use if allowed. (e.g. secure downloads, scheduled reader-room deliveries, public internet, etc)

#### A.xvi. Tessella: Digitisation of Paper based Government Records for future access

<b>Preservation Scenario ID</b>	PS-016
<b>Author</b>	Ashley Hunter, Tessella Ltd
<b>Background</b>	Collections of Government records may span several years of accessions, during which time the producers of the material moved from creating paper based records to digital records. To maintain accessibility across the 'divide' archives are digitising the paper manifestations in order to make these accessible along with the born-digital material. Typically, the paper based record will be scanned (or photographed) to create a high resolution Preservation Manifestation, and at the same time create a lower grade presentation copy along with other files including OCR'd text where available, and specific technical metadata extracted during the digitisation process (Page number, camera specification, creator, etc).
<b>Type of digital information</b>	<p>Preservation Manifestation formats include, but not limited to:</p> <ul style="list-style-type: none"> <li>• TIF</li> <li>• JP2</li> <li>• RAW (Where this is a specific format to the camera manufacturer)</li> <li>• Presentation formats including JPG, JP2, PDF</li> <li>• OCR output in the form of TXT, RTF, CSV</li> <li>• Metadata in XML formats.</li> </ul>
<b>Link to sample data</b>	Tessella to provide a test set of typical files (not real, made up - likely the DROID test file corpora) for testing purposes
<b>Threat(s) to the data</b>	<p>The main purpose of the Archive is to provide long term search and access of records for the various approved user communities, and to this end the Archive must defend against passive preservation issues, relating to:</p> <p>(1) keeping the digital objects safe from 'bit rot' or 'data decay' on the primary storage media.</p> <p>(2) The Archive may want to provide multiple versions of its digital objects, rendering them in different formats for different access purposes (free low grade manifestation, and commercially available full high-resolution manifestations), and to this end Archives can use Active Preservation methods including file format migration to provide these various manifestation types.</p> <p>(3) Representation Information about the record's provenance and authenticity is provided through arrangement of the digital objects in to a hierarchy of collections, with Descriptive Metadata provided according to agreed</p>

	metadata schemas. Commonly these are bespoke to each Archive, rather than using standard Schemas such as METS, MODS, EADS, PREMIS, etc.
<b>Designated Community</b>	Initially the submitting Government Department and any other affiliated departments and organisations will have access to the material and over time this access will be relaxed for some of the records leading to wider or even public access to records for use by policy investigators, historians, and the general public.
<b>Preservation Technique</b>	<p>(1) The standard approach to defend against bit rot is to keep multiple copies of the AIP on different storage media, and to periodically test the integrity of these objects against their known fixity/checksum values at the time of ingest. Where corruptions are found, the system should notify the Archive staff so that the corrupted file can be replaced from one of the other AIP copies that has recently passed the integrity test.</p> <p>(2) Use file migration techniques to provide digital information in alternative file formats. This process may in itself be a lossy process, so some process of quality assessment and validation needs to be applied to ensure that the appropriate level of quality is maintained following each format migration. Several migrations may be required over time to provide the Designated Community with the information that they want to access in a format suitable to their needs.</p> <p>(3) Provide methods for translating between metadata schema definitions via code or XSLT, to enable exchange of networks of representation information. Facilitate further integration with catalogue collection systems through the use of OAI-PMH exchange protocols.</p>
<b>Usage</b>	Archive staff, submitting government department staff via access requests, public access request information
<b>Success Criteria</b>	Digital objects remain available to their designated community through 'Search' &/or 'Browse' functionality, and are directly accessible to these users in forms and ways that are meaningful to them and can facilitate their re-use if allowed. (e.g. secure downloads, scheduled reader-room deliveries, public internet, etc)

#### A.xvii. Tessella: Scientific Datasets (ISIS Neutron and Muon Facility, STFC)

<b>Preservation Scenario ID</b>	PS-017
<b>Author</b>	Ashley Hunter, Tessella Ltd
<b>Background</b>	<p>ISIS, the Neutron and Muon Source at Rutherford Appleton Laboratory in the UK, and part of STFC, wanted to preserve their instrument data that has been generated over the many years of operation. The instrument data has evolved over time as the instruments themselves have been updated and enhanced, resulting in various file formats specific to each instrument. A common data structure was developed for these known as ISIS RAW. Additional ad-hoc files may also be present to describe additional metadata about the instrument or its operation, and temporary files are generated during the cyclic operation of the instruments for backup purposes (SAV file formats).</p> <p>A separate catalogue system maintains representation information relating to the setup of each instrument and why it was used in a specific experiment and by whom. This remains confidential for a 2 year period so that results can be derived by the initiating investigator before the results are made publically available for use by the wider research community.</p>
<b>Type of digital information</b>	<p>Instruments generated files in the following formats:</p> <ul style="list-style-type: none"> <li>• SAV (Proprietary binary formats)</li> <li>• ISIS RAW (RAW)</li> <li>• Nexus (NXS).</li> <li>• Ad-hoc text formats (TXT &amp; XML)</li> </ul>
<b>Link to sample data</b>	Link to publically available historic instrument data to be added here
<b>Threat(s) to the data</b>	<p>The main purpose of the Archive is to provide long term search and access of records for the various approved user communities, and to this end the Archive must defend against passive preservation issues, relating to:</p> <p>(1) keeping the digital objects safe from 'bit rot' or 'data decay' on the primary storage media.</p> <p>(2) ISIS may want to provide multiple versions of its digital objects, like aggregating older ISIS RAW and other ad-hoc files in to one combined NXS format.</p> <p>(3) Network of Representation Information about the instrument data is held in the iCAT cataloguing system.</p>
<b>Designated Community</b>	Initially the Research scientists that commission the work, but later this will become public access to all the instrument data after the 2 year closure period has passed.
<b>Preservation Technique</b>	(1) The standard approach to defend against bit rot is to keep multiple copies of the AIP on different storage media, and to periodically test the integrity of these objects against their known fixity/checksum values at the time of ingest. Where corruptions are found, the system should notify the ISIS support staff so that the corrupted file can be replaced from one of the other AIP copies that has recently passed the integrity test.

	<p>(2) Use file migration techniques to provide digital information in alternative file formats. The MANTID software has been used to provide this migration pathway from the older ISIS RAW data formats to the newer NXS formats.</p> <p>(3) Provide methods for translating between metadata schema definitions via code or XSLT, to enable exchange of networks of representation information. Facilitate further integration with catalogue collection systems through web services.</p>
Usage	ISIS Research staff, Client sponsoring research staff, and wider academic community following release as public access
Success Criteria	Research datasets remain available to their designated community through 'Search' &/or 'Browse' functionality, and are directly accessible to these users in forms / formats that are meaningful to them and can facilitate their re-use when allowed.

#### A.xviii. UKDA Migration of Social Science and Humanities Data between statistical formats

Preservation Scenario ID	PS-018
Author	Sharon Bolton (UK Data Archive), Contact Hervé L'Hours (UK Data Archive)
Background	The UK Data Archive is curator of the largest collection of digital data in the social sciences and humanities in the United Kingdom. The Archive holds several thousand datasets relating to society, both historical and contemporary. For conversion to standard preservation (and access) formats, a standard approach is required at ingest to ensure that processing staff and, subsequently, users have confidence in the conversion output.
Type of digital information	<p>A large proportion of our files are quantitative data files which are deposited primarily in SPSS but we also receive STATA, SAS and Excel format.</p> <p>File formats include:</p> <ul style="list-style-type: none"> <li>• SPSS</li> <li>• STATA</li> <li>• SAS</li> <li>• Excel (XLS)</li> </ul>
Link to sample data	An ESDS Government sample dataset can be made available on request
Threat(s) to the data	<p>Various proprietary statistical software packages manage files in a variety of formats to support general and software-specific functions. Differences between packages and their formats and the large and complex nature of the data sets present a significant risk that some information will be lost (through error or truncation) or damaged on conversion with no record produced of the changes incurred. These are ongoing problems with ingest conversion but similar issues exist with obsolescence of a particular format version of a statistical software package.</p> <p>Threats include:</p> <ol style="list-style-type: none"> <li>(1) Truncation of variables (to a reduced number of decimal places)</li> <li>(2) Truncation of labels (to a reduced number of characters)</li> <li>(3) Non-identical feature sets supported (Not all of format A included in format B or some elements of format B left blank because not present in format A)</li> <li>(4) Different approaches to the application of weighting (and truncation of decimal places may occur with weighting variables)</li> </ol>
Designated Community	Social Scientists working with quantitative data
Usage	Statistical information used for secondary analysis or replication of results
Success Criteria	<p>Processing users will either be presented with confirmation that the outcome of the format conversion is content-identical or with clearly flagged difference between the two files, preferably with verbose explanations.</p> <p>The end users will either be presented with content-identical statistics or detailed explanations of any variation from the originally deposited material sufficient to replicate any analysis made on that material</p>

## APPENDIX B: Example classification of digital objects

The following table provides one interpretation of how a classification of each digital object type could be implemented. The APARSEN consortium have not agreed on this as yet, this is more for internal reference to allow WP14 Partners to visualise one expression of this information in order to stimulate further input on this topic.

Rendered?	Static?	Simple?	Passive?	Example
Y	Y	Y	Y	JPEG image
Y	Y	Y	N	PHP script
Y	Y	N	Y	Word file
Y	Y	N	N	Word file with macros
Y	N	Y	Y	Active Blog
Y	N	Y	N	no examples
Y	N	N	Y	no examples
Y	N	N	N	no examples
N	Y	Y	Y	Time series
N	Y	Y	N	Executable file
N	Y	N	Y	FITS file
N	Y	N	N	Executable file
N	N	Y	Y	Growing time series
N	N	Y	N	no examples
N	N	N	Y	Java compiled file
N	N	N	N	Transactional database with built-in procedures

**Table 3: Example classification of digital objects**

## APPENDIX C: REFERENCES

- [1] Hanks, S., Pollack, M. E., & Cohen, P. R. (1993). "Benchmarks, test beds, controlled experimentation and the design of agent architectures." *AI Magazine*, 14(4), 17-42
- [2] Guttenbrunner, M. "Challenges for the Evaluation of Emulation",  
[http://publik.tuwien.ac.at/files/PubDat\\_180727.pdf](http://publik.tuwien.ac.at/files/PubDat_180727.pdf)
- [3] Potter, Mauren. "Researching Long Term Digital Preservation Approaches in the Dutch Digital Preservation Testbed (Testbed Digitale Bewaring)", Digital Preservation Testbed, Netherlands. *RLG DigiNews*, Vol. 6, No. 3 (15 June 2002):  
<http://www.imaginar.org/dppd/DPPD/141%20pp%20Researching.pdf>
- [4] Hofman, Hans; Verdegem, Remco, *et al.* DELOS WP6 (Deliverable D6.1.1). "Framework for Testbed for digital preservation experiments", 11 November 2004  
[http://www.dpc.delos.info/private/output/DELOS\\_WP6\\_D611\\_finalv2\(0\)\\_den Haag.pdf](http://www.dpc.delos.info/private/output/DELOS_WP6_D611_finalv2(0)_den Haag.pdf)
- [5] PLATO Website: <http://www.ifs.tuwien.ac.at/dp/plato/intro.html>
- [6] Helwig, P.; Rog, J.; van Wilk, C.; Nicchiarelli, E.; Thaller, M.: Planets TB/3-D2. "Test methods for Testbed", 30 November 2007  
[http://www.planets-project.eu/docs/reports/Planets\\_TB3-D2\\_MethodsForTesting.pdf](http://www.planets-project.eu/docs/reports/Planets_TB3-D2_MethodsForTesting.pdf)
- [7] Open Planets Foundation website link to the proposed community testbed: <http://wiki.opf-labs.org/display/KB/The+OPF+Testbed>
- [8] Jones, S.; McCann, P.; Kim, Y.; A practical guide to implementing the DCC testbed methodology:  
[http://www.dcc.ac.uk/sites/default/files/TestbedMethodologyManual\\_1.2%5B1%5D.pdf](http://www.dcc.ac.uk/sites/default/files/TestbedMethodologyManual_1.2%5B1%5D.pdf)
- [9] CASPAR Evaluation/Validation Report available from  
[http://www.alliancepermanentaccess.org/filestore/CASPAR-deliverables/CASPAR-4104-RP-0101-1\\_0.pdf](http://www.alliancepermanentaccess.org/filestore/CASPAR-deliverables/CASPAR-4104-RP-0101-1_0.pdf)
- [10] Giaretta, David; Matthews, Brian; Bicarregui, Juan; Lambert, Simon; Guercio, Mariella; Michetti, Giovanni; *et al.* (2009). "Significant Properties, Authenticity, Provenance, Representation Information and OAIS Information.", *iPRES 2009 Proceedings*. UC Office of the President: California Digital Library.: <http://www.escholarship.org/uc/item/0wf3j9cw.pdf>
- [11] Reference Model for an Open Archival System (OAIS), Draft Recommended Standard, CCSDS 650.0-P-1.1 (Pink Book) Issue 1.1 August 2009. Retrieved from  
<http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf> on 16 January 2012. The original Blue Book (CCSDS 650.0-B-1) Reference Model for an Open Archival Information System version and ISO standard (ISO 14721:2002) is available from <http://public.ccsds.org/publications/archive/650x0b1.pdf> (last retrieved on 16 January 2012).
- [12] SHAMAN website link to demonstrator system: <https://shaman-ip.eu/shaman/Demonstrator%20for%20Memory%20Institutions/>
- [13] SCAPE project website: <http://www.scape-project.eu/news/d18-1-first-evaluation-report-draft>
- [14] Testbed LDP website link: <http://www.ltu.se/research/subjects/Data-och-systemvetenskap/Langsiktigt-digitalt-bevarande/Avslutade-projekt>
- [15] David Giaretta, *Advanced Digital Preservation*, Springer-Verlag 2011
- [16] Survey report, December 2009, is available from [http://www.parse-insight.eu/downloads/PARSE-Insight\\_D3-4\\_SurveyReport\\_final\\_hq.pdf](http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf) (last retrieved on 30 January 2013)
- [17] APARSEN Glossary: <http://www.alliancepermanentaccess.org/index.php/knowledge-base/dpglossary/>

[18] Kopal project website: <http://kopal.langzeitarchivierung.de/>