Final Demonstration Report

Authors
Sven Schlarb, Kristin Dill (Austrian National Library), Opher Kutner (Ex Libris Ltd.), Bolette Jurik, Rune Ferneke-Nielsen (State and University Library Denmark), William Palmer (The British Library), Leila Medjkoune (Internet Memory Foundation), Ivan Vujic (Microsoft Research), Catherine Jones (Science and Technology Facilities Council)

July 2014

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

Executive Summary

In the SCAPE project, the role of the demonstrations is to showcase the *SCAPE Platform*[1] technologies as well as *Preservation Component*s in the context of the *User Stories* developed by the *Demonstrating Institutions* to third parties, both working within the partner organisations and in external institutions.

This document gives an overview of the scope of demonstrations carried out in the SCAPE project's *Testbeds* work package. First, it shows the different demonstration environments that were created by the partners involved in demonstration activities. Second, it outlines the demonstration assets that were able to be demonstrated at the individual institutions. Finally, it reports on the demonstration activities that took place in the first half of 2014.

[1] Terms either defined in the glossary or in its own section within this document are marked in italics.

Table of Contents

1 Introduction

The final demonstration report documents the outcomes of the demonstration activities carried out as part of the completion of the *Testbeds*[2] sub-project. SCAPE environments, assets and preservation stories were demonstrated at participating institutions in the first half of 2014. The purpose was to reach non-participating members of the involved institutions as well as the interested public with the outcomes of the SCAPE project.

This report is organized into six chapters. The first three chapters, Chapter 2 "Demonstration Environments", Chapter 3 "Demonstration Assets" and Chapter 4 "Preservation User stories" consist of an updated version of the information included in the document D19.1, "Demonstration scope definition document"[3]. For the purpose of completeness, along with the changes to the demonstration environments, the demonstration assets and the preservation *user stories* that were a result of the dynamics of the ongoing work at the different institutions, relevant information previously reported in D19.1 is included in the final demonstration report as well. For this reason, there is significant overlap between the first three chapters of D19.2 and D19.1.

Chapter 5 is the main concern of D19.2 as it includes the reports of the demonstration events and visiting opportunities that took place at the individual institutions (BL, IM, ONB, SB, STFC, EXL, and MSR). The penultimate chapter on "Task Publicity" (Chapter 6) highlights the promotion activities carried out in the context of the demonstrations by TU.WP.1. Chapter 7 provides the report's conclusion.

2 Demonstration Environments

The *SCAPE Platform*[4] consists of a number of system entities, services and applications, which form a flexible infrastructure (including both hardware and software) that can be reconfigured on demand to support the different scenarios and workflows required by the unique needs of institutions concerning the scalable preservation of their content. Within the *Testbeds* subproject, the participating institutions have created *Local Instance*s, in which architecture and hardware specifications of the *SCAPE Platform* have been tailored to reflect the SCAPE outcomes and concepts relevant to their local environments and *user stories*. The preservation *user stories* associated with these environments will be explained further in Chapter 3 of this document.

Chapter 2 provides an overview of these SCAPE demonstration environments at the participating institutions. These institutions include content providers, who have implemented SCAPE technology in their institutional environments (British Library, Internet Memory Foundation, Austrian National Library, State and University Library in Denmark, and at the Science and Technology Facilities Council) and commercial partners, who have taken up SCAPE concepts in their commercial products (ExLibris and Microsoft Research).

In the context of the demonstrations task, information about the environments and preservation *user stories* was published on the SCAPE wiki[5] so that interested parties could contact institutions with environments and preservation stories similar to their own. The Wiki also contained information

[2] Terms either defined in the glossary or in its own section within this document are marked in italics.

[3] D19.1 was an internal project document with restricted access.

[4] For more information about the *SCAPE Platform* Architecture:
http://homepage.univie.ac.at/rainer.schmidt/publications/iPres12.pdf

[5] http://wiki.opf-labs.org/display/SP/Demonstrations

about the possibility of scheduling a visit to receive a demonstration of the SCAPE tools in the particular environment.

2.1 The British Library (BL) Hadoop Developer Platform

Figure 1 shows an overview of the architecture of the local instance at the British Library.

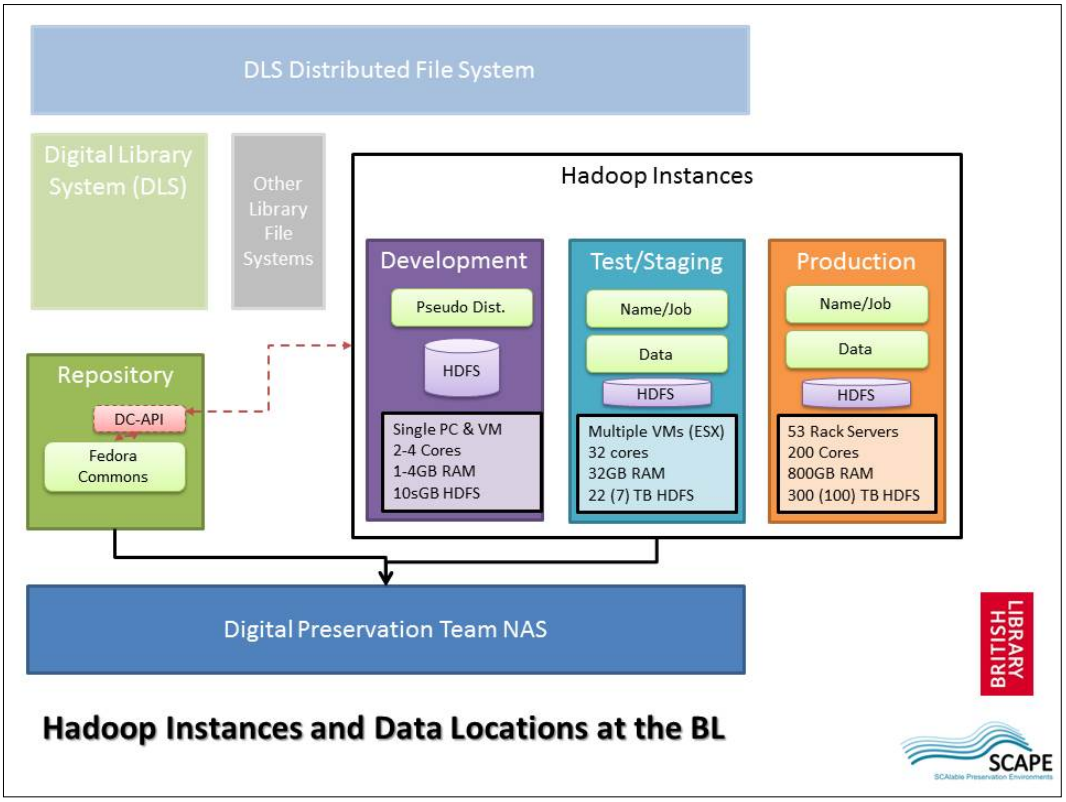

Figure 1 - Architecture overview on the local instance at The British Library

The British Library architecture is using an incremental development cycle, starting with a small data set being used on a Hadoop installation on a virtual machine, up to large scale processing using the BL's developer Hadoop cluster.

Testbed content is held either in the Digital Preservation Team's repository or on the NAS directly as appropriate. The development *Apache Hadoop* environment at the BL is a VMWare ESXi cluster with 32 CPUs, 224GB RAM and ~27TB HDD. It is currently configured to have 30 1CPU nodes; 1 manager, 1 *NameNode/JobTracker* and 28 *DataNode/TaskTrackers*, each with 1CPU/6GB RAM/500GB HDD.

2.2 Internet Memory Foundation (IM)

As Figure 2 - Distributed Crawler and Document Repository at shows, the *Local Instance* at Internet Memory Foundation consists of two distinct multi-node hardware clusters designed to retrieve and archive web content.

The first one is the Distributed Crawler cluster, which is used to retrieve content from the web.

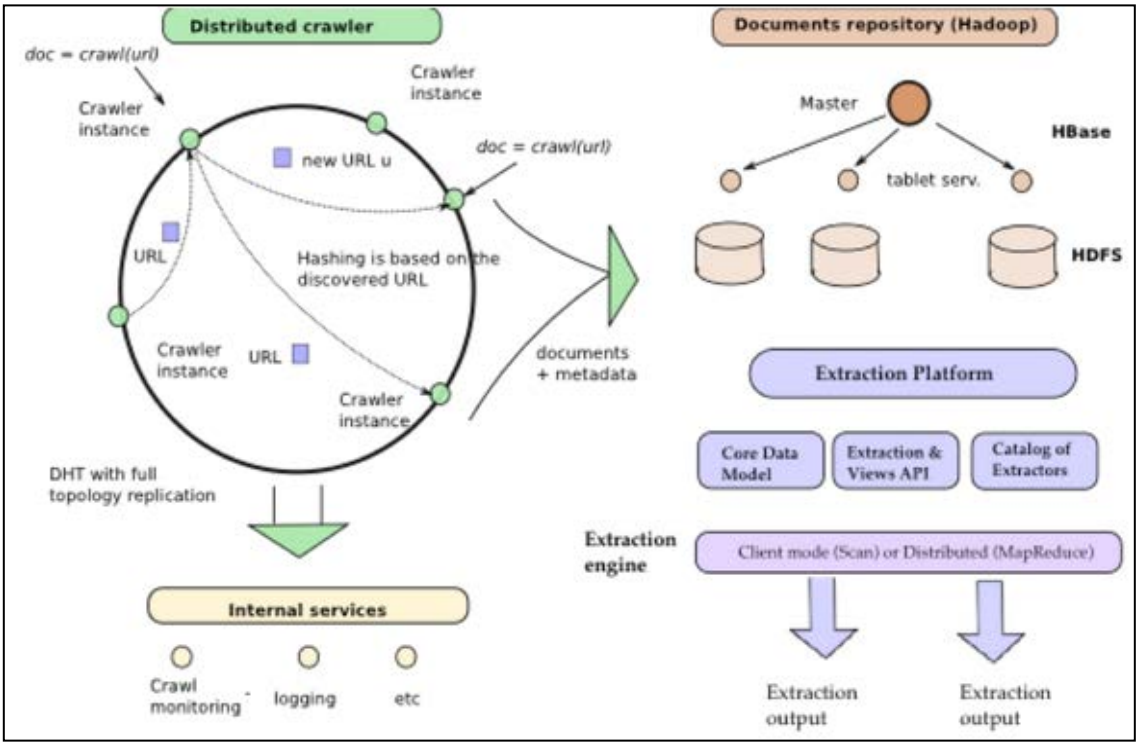

Figure 2 - Distributed Crawler and Document Repository at Internet Memory Foundation

The second cluster is the Document Repository which is dedicated to processing and storage of the crawled data. This system builds upon *Apache Hadoop* and *HBase*. The in-house developed extraction platform enables the user to create a data specific workflow of "extractions" used to derive information from the "raw" data (detect *MIME* type, extract plain text, and detect news articles). The output of one extractor can of course serve as an input for another extractor. When the processing of a workflow is finished, the data is ordered and stored in *HBase*.

2.3 Hadoop Cluster at the Austrian National Library (ONB)

At the Austrian National Library, a dedicated experimental cluster has been set up for the SCAPE project. First, the hardware for the cluster consisting of one controller and five worker nodes was selected and then the installation on the Ubuntu Server 10.04 LTS 64bit operating system together with a Cloudera Hadoop distribution (CDH)[6] as a basis for a *SCAPE Platform* installation was set-up.

Figure 3 shows the hardware of the Austrian National Library's *Apache Hadoop* cluster.

[6] https://ccp.cloudera.com/display/SUPPORT/CDH+Downloads

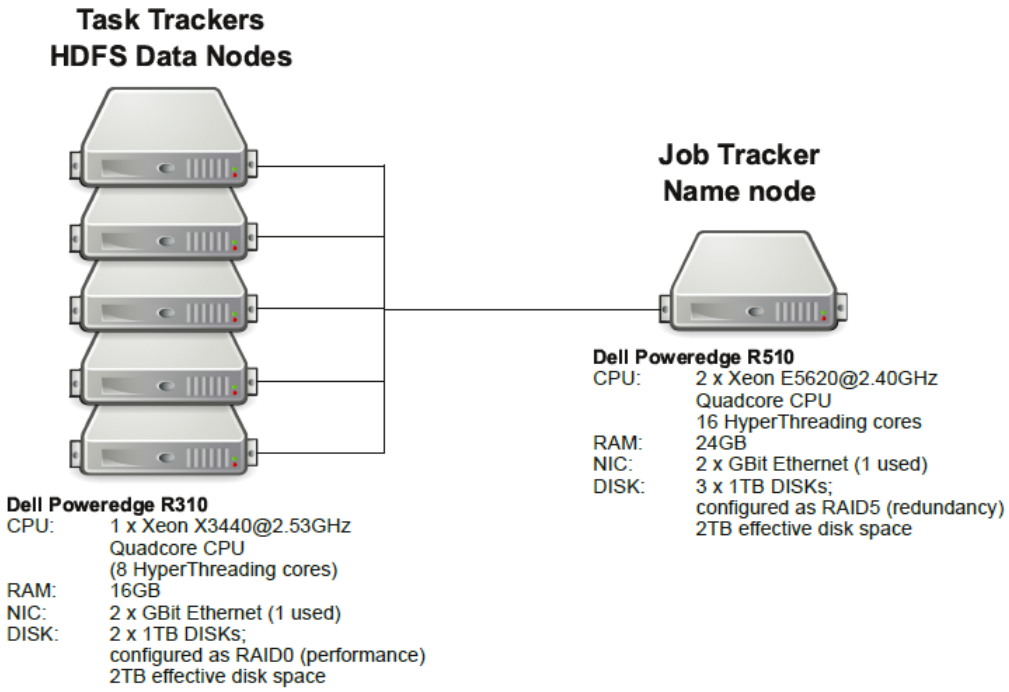

Figure 3 - Hadoop Cluster at the Austrian National Library

As shown in Figure 3 the cluster consists of 5 machines with the Ubuntu Server 64bit 10.04 LTS operation system. A Cloudera Hadoop distribution (CDH)[7] was chosen as a basis for a *SCAPE Platform* installation.

In the following, an overview on the hardware, software and the SCAPE system for workflow demonstration is presented according to the current planning. As an extension to the basic Hadoop installation, *Apache Pig* and *Apache Hive* were installed. These extensions have been added to make generating *MapReduce* jobs easier.

Additionally, on the master and worker nodes of this cluster the tools required for executing large-scale workflows of the various preservation *User Stories* (see Chapter 4) have been deployed. Version numbers of the tools are only indicated if they have not been installed from the Ubuntu Lucid (10.04LTS) package repository:

- Unix tool file
- Jpylyzer 1.10.1
- *FITS* 0.6.2
- JHove 1.4
- openjpeg-tools
- *C3PO* 0.1.0
- Exiftool
- ImageMagick
- *ToMaR* (branch filelists)

[7] http://www.cloudera.com/content/cloudera/en/products-and-services/cdh.html

- *ToMaR* 1.5.0

And Figure 4 Figure 4 shows an architecture diagram of the Austrian National Library's *SCAPE Platform Local Instance*.

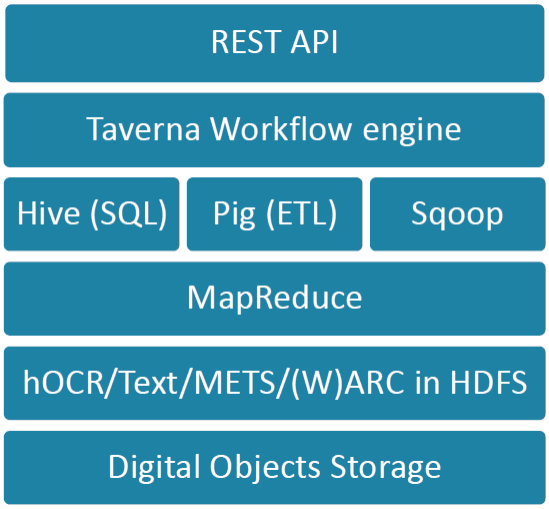

Figure 4 - Layered architecture diagram of the Austrian National Library's SCAPE Platform local instance

The base level of this layered architecture diagram is digital object storage, which is mounted to all worker nodes of the cluster as network attached storage (NAS). This way, *MapReduce* applications are able to access a file system path in a shared storage space. Due to the size limitation of the distributed storage (5 TB effective storage for experimental data and intermediate and final processing results), image data from digital books are not copied to the cluster, but only local file system paths pointing to the external NAS are used. Only "small files" (HTML, Text files, METS, XML, etc.) – files which have a file size below 5 MB – are stored and processed directly in HDFS. The MapReduce layer represents Java applications which implement MapReduce jobs based on Apache Hadoop. On top of this layer, *Apache Hadoop* is extended by framework extensions, such as *Pig*, *Hive*, and Sqoop[8] which, one the one hand, are used to generate *MapReduce* jobs based on an expression language (*PigLatin* for *Pig*) and (*HiveQL* for *Hive*) and, on the other hand, serve the purpose of integrating with existing relational data base systems such as MySQL (Sqoop). The REST API of the *Taverna Server* is used to start the workflow processing by submitting a workflow description document (created using the *Taverna Workbench*) and setting input parameter values.

2.4 Hadoop Cluster at the State and University Library, Denmark (SB)

The SCAPE environment at the State and University Library (SB) is shown in the architecture diagram shown in Figure 5 below. It shows three layers; at the bottom is SB's existing preservation platform, and on top of this is the *SCAPE Platform*. Both platforms are hosted at SB. There are two categories at the top level, *SCAPE Tools* (equivalent to *SCAPE Component*s) and *SCAPE Application*s. *SCAPE tool*s are used to enrich SB's *SCAPE Platform*s functionality. The category *SCAPE Application* consists of complex software applications created in the SCAPE Project.

The *Local Instance* of *the SCAPE Platform* at SB is on the basis of *Apache Hadoop* (Cloudera distribution), *Taverna* and *C3PO*.[9]

[8] https://sqoop.apache.org

[9] https://github.com/openplanets/c3po

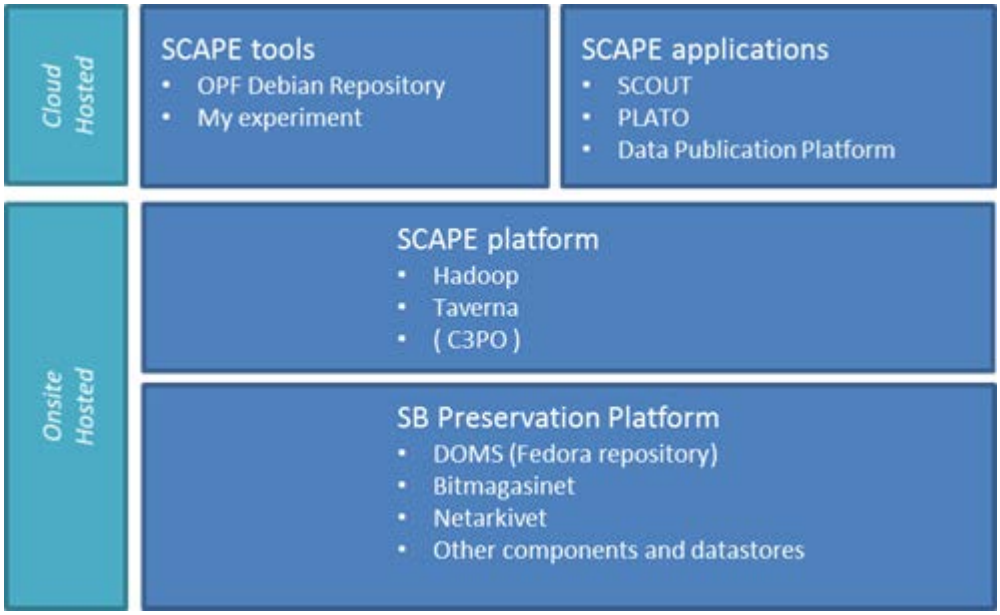

Figure 5 - SCAPE components architecture at SB

SB's *Local Instance* of the *SCAPE Platform*, represented by the middle layer in Figure 5, is running with the hardware configuration shown in Figure 6 below. The hardware consists of five workstations and a storage box from EMC with 20 TB of storage.

Apache Hadoop is running on a cluster with four nodes and has access to the storage box through a high speed connection. *Taverna* and *C3PO* are running on a separate machine and also connected to the storage box through high speed connection.

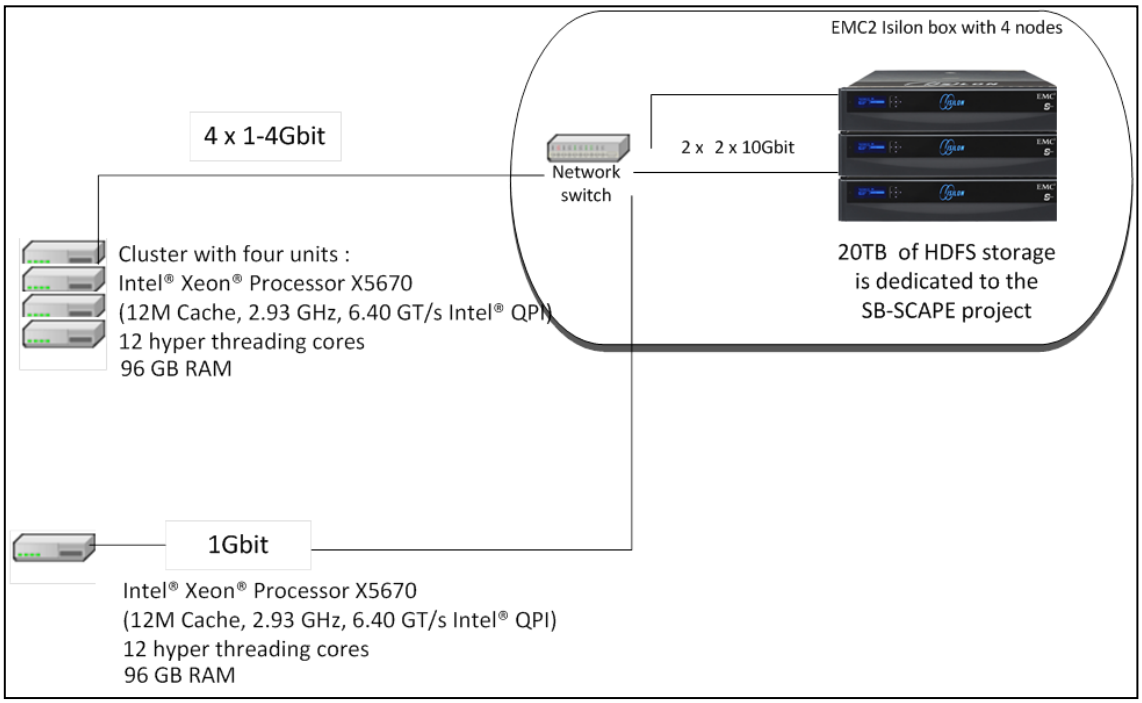

Figure 6 Hardware configuration of SB's local instance of the SCAPE Platform

2.5 Hadoop Cluster at the Science and Technology Facilities Council (STFC)

The *Apache Hadoop* cluster used in SCAPE at the Science and Technology Facilities Council is illustrated in Figure 7. It is maintained and provided by other members of the STFC Scientific Computing Department and is a test facility only available within the STFC firewall.

The Hadoop cluster had initially six slaves providing 70TB of HDFS storage and 24 *MapReduce* slots; these are managed by two virtual machine head-nodes, the *NameNode* and *JobTracker*. Each slave machine has both a *TaskTracker* and *DataNode* to enable the minimum movement of data to the compute resource. The operating system used is Scientific Linux which is a scientifically enhanced version of Red Hat Enterprise Linux.[10] At a later stage, the demonstrator environment (i.e. the computer cluster) was upgraded. There are now 8 nodes delivering a total of 96 CPU cores and 1024 GB RAM over the nodes. It is running Scientific Linux v6. There is 8 x 4TB of storage and 1Gbit/s network between the nodes.

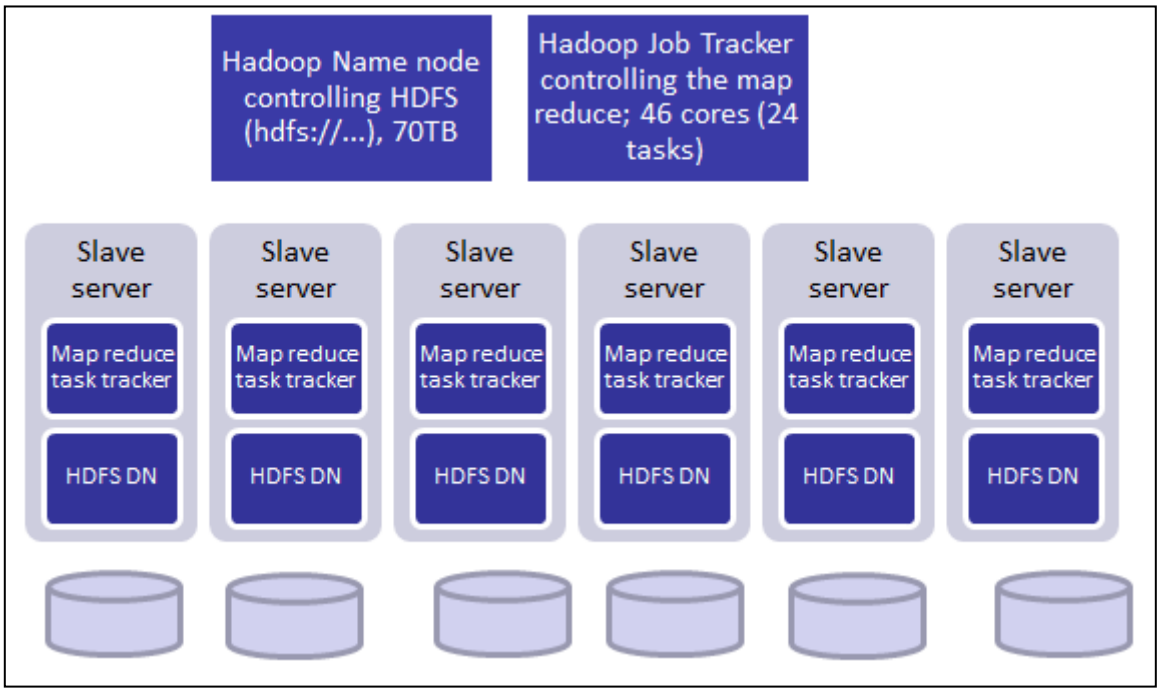

Figure 7 - SCAPE Demonstrator environment at STFC

2.6 Ex Libris Ltd (EXL)

For the purpose of this project the following small-scale Rosetta environment was built:

- OS: Linux RHEL 5.5 x86 64bit
- CPU: 4x Intel® Xeon® CPU E5530 @ 2.40GHzRAM: 16GB
- JBoss AS v.5.1 running on Java 1.6.0_30, 4GB heap size
- Oracle version: 11g, 1.5GB SGA

For the Testbed experiments, the project was also provided with temporary access to a multi-server environment with the following specifications:

- 3 application servers (all-in-one), each running
    - OS: Linux RHEL 6.3 x86 64bit

[10] https://www.redhat.com/products/enterprise-linux

- PU: 24x Intel® Xeon® CPU E5-2620 @ 2.00GHz
- RAM: 32GB
- JBoss AS v.5.1 running on Java 1.6.0_45, 16GB heap size

- 2 Oracle servers running
  - Linux RHEL x86_64
  - Oracle Real Application Cluster (RAC) version: 11g, 10GB SGA

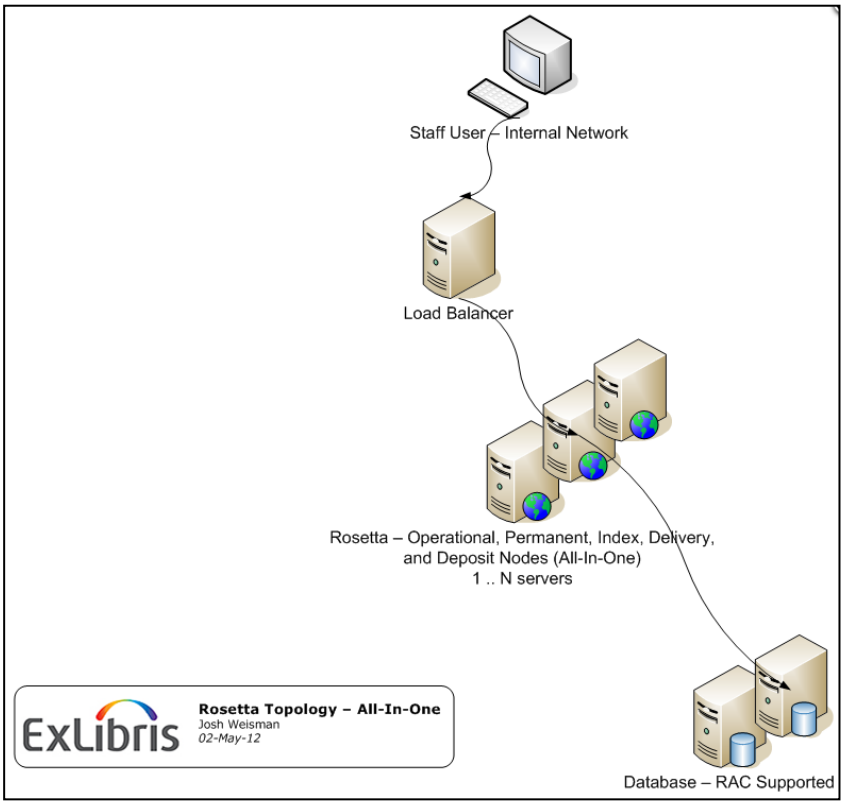

Figure 8 - The Rosetta Multi-server environment used for SCAPE

Figure 8 gives an overview on the Rosetta Software architecture.

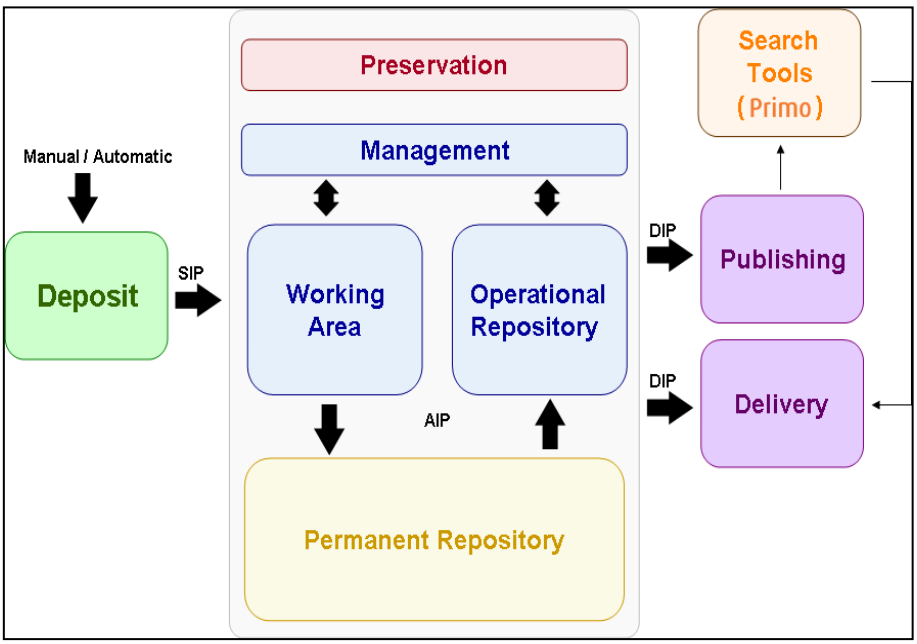

Figure 9 - Rosetta Architecture

The Rosetta environment is built on the OAIS model and includes several components (application roles) that can be deployed on any number of servers. Rosetta application servers can work in distributed mode (i.e. different servers are dedicated to one or more components) or in clustered mode (all components are deployed on all servers). Communication between the servers is based on Web Services and JMS and Oracle AQ. Rosetta API documentation is freely available via the Ex Libris Developer Network. Authentication is done via the Ex Libris Patron Directory Service application (PDS), which runs on an Apache HTTP Server and interfaces with a variety of user management systems. Rosetta can also manage internal users.

The following components exist in every Rosetta environment:

- Deposit: Handles manual and automated deposit in Repository: SIP processing, operational repository data processing, preservation action processing
- Permanent: The permanent repository (write once, read only)
- Delivery: Present objects via various viewers
- Index: Apache Solr[11] platform

SIP processing is comprised of the following process:

1. Validation Stack: A series of processes that add critical preservation information to the submitted data. This includes:
    a. Checksum validation
    b. Virus check

[11] https://lucene.apache.org/solr

c. Format identification

d. Technical metadata extraction

e. Risk extraction

2. Assessment: Staff members can review (all or a pre-defined sample of) deposited content, allowing them to accept, reject, manipulate SIPs – all in accordance with the institution policy.

3. Enrichment: A list of post-processing procedures related to digital asset management aspects. These can include, but are not limited to:

   a. Derivative copy generation

   b. Synchronization with an external CMS system (e.g. a library catalogue)

   c. Collection creation

   d. Commit: The SIP is committed to the permanent repository as one or more Intellectual Entities (IEs) and these will include full audit trails of all activities since submission. From here on any change will generate a new IE version with appropriate provenance metadata.

The SCAPE Rosetta instance is available at http://scape.exlibrisgroup.com:1801/mng. Authentication details can be provided on request.

2.7 Microsoft Research (MSR)

MSR has designed and implemented a Microsoft Azure-based architecture (see Figure 10 to get an overview).
Windows Azure is Microsoft's application platform for the public cloud. The platform can be used in many different ways. For instance, Windows Azure can be used:

- to build a web application that runs and stores its data in Microsoft datacentres,
- to store data, with the applications that use this data running on-premises (that is, outside the public cloud),
- to create virtual machines for development and test or to run *SharePoint* and other applications,
- to build massively scalable applications with lots and lots of users.

Considering the Hardware, Windows Azure is extensible. To illustrate, here are two configurations (instances) for processing large data sets:

- 8 computation cores with 60 GB of RAM
- and 16 compute cores with 120 GB of RAM

Both instance configurations are delivered on systems consisting of:

- Intel Sandy Bridge processors at 2.6 GHz
- DDR3 1600 MHz RAM
- 5 x 1 TB disks
- Two network connections
  - 10 GigE network for storage and internet access

- RDMA + InfiniBand (IB) 40 Gbps network for inter node communication

More detailed information is published on the Microsoft Windows Azure website.[12]
The following SCAPE Azure v.1.0 components are the building blocks of the implemented architecture:

- **Authentication** is in charge of User Authentication (e.g. user profile and authentication). With Service Authentication we want to ensure that external services can communicate securely with internal services currently running in SCAPE Azure (*SAZ)*.
- **SCAPE Azure Execution Layer** is responsible for running and managing all the operations and logging within SAZ.
- **Content Representation Layer** is a metadata layer which is describing Data, Reports, Logs and Workflows. It maps stored data and metadata in *SQL Azure* (an Azure specific SQL dialect).
- **Tools and Resources Layer** represents our *Action service*s and tools we are using for Characterization, Conversion, Comparison and Reporting.
- **Data store** is virtually unlimited storage in *BLOB (Binary Large Object).*

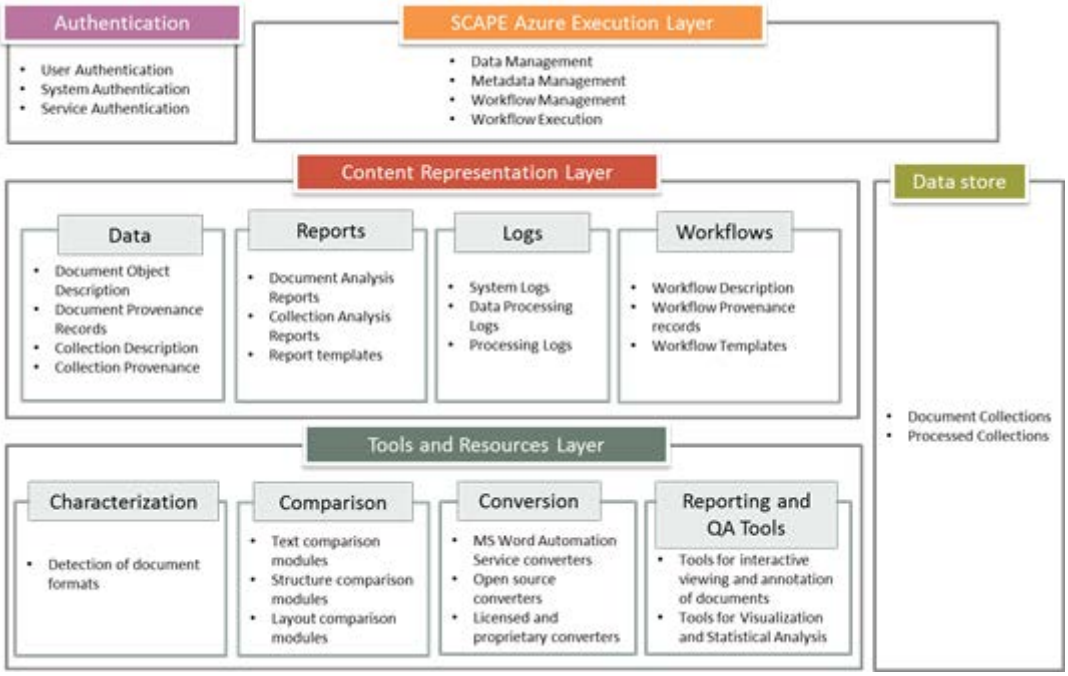

Figure 10 - Architecture components of SCAPE Azure v.1.0.

Figure 11 shows the details of the implemented architecture. Data is placed in the BLOB storage. Conversion and comparison functions are implemented as worker roles. SharePoint is placed in a VM (Virtual Machine) environment and the MS Word Automation Services are leveraged to convert document formats. Reporting services are under development. They will aggregate processing information, which includes system performance related to ingest, conversion, and comparison, and qualitative data about the quality of the conversion based on different techniques.

[12] http://www.windowsazure.com/en-us/home/features/big-compute/

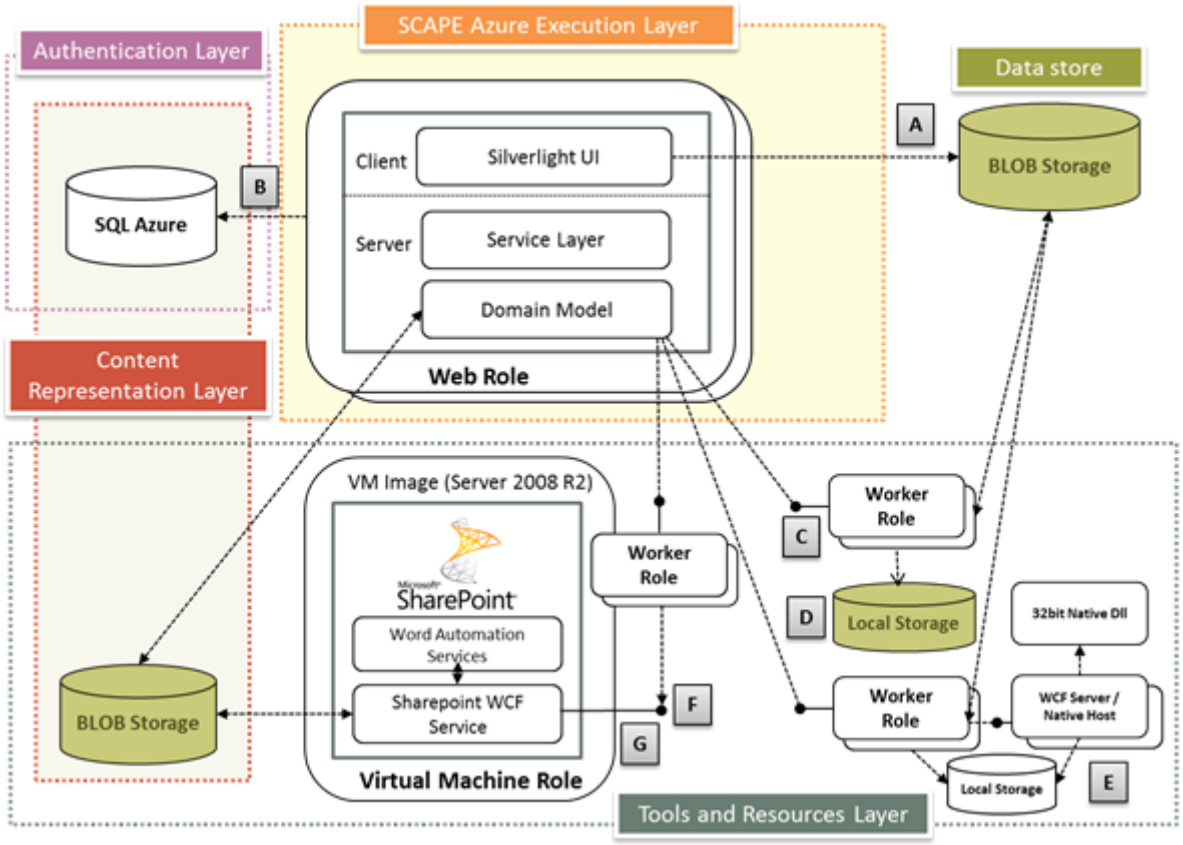

Figure 11 - System architecture of SCAPE Azure v.1.0.

3 Demonstration assets

The following chapter provides an overview of the demonstration assets created and/or utilized within the SCAPE project. These include the *SCAPE Platform*, SCAPE *Preservation Components*, Preservation Watch and Commercial Products. For each asset, the related preservation scenarios, *Demonstrating Institutions*, and main demonstration contact, and technical support (if need be) are indicated.

3.1 SCAPE Platform

The *SCAPE Platform* is a software environment to enhance the scalability of preservation activities. It provides a scalable architecture and technologies for parallel processing, automated data processing, and error recovery.

More concretely, in the context of the *Testbeds*, the *SCAPE Platform* was used as the blueprint for creating various local *SCAPE Platform* instances which consist of a basic *Apache Hadoop* Installation together with a set of related components from the *Apache Hadoop* ecosystem (such as Apache HBase, *Hive* and Pig) and of *SCAPE component*s providing additional interfaces and preservation specific capabilities.

3.1.1 Fedora 4[13]

The SCAPE digital objects repository (DOR) reference implementation is based on Fedora 4. It was used to show how to implement the three main APIs required by a DOR and to demonstrate the data structure (SCAPE Digital Object Model) support needed by digital objects repositories in order to facilitate ingest and access of Intellectual Entities.

3.2 SCAPE Preservation Components

The SCAPE *Preservation Component*s are divided into the following categories: Characterisation services, Action services and Quality assurance services. The *SCAPE Testbeds* used these components to build composite workflows to provide solutions according to the requirements derived from the preservation *user stories*.

3.2.1 Characterisation services

3.2.1.1 FITS[14]

This tool was mainly used as a legacy tool which supports "characterisation". This means that it is able to extract properties of files. This tool uses various other file format identification and characterisation tools underneath and allows the processing of single files as well as directories. For this reason it was a good starting point for studying the integration of this tool into workflows, such as the ones developed as part of the *User Story* "File Format Identification and Characterisation". Additionally, it played an important role related to the gathering of statistical data for Planning and Watch using the *C3PO* tool. It was applied by ONB and SB.

3.2.1.2 Nanite[15]

The *Nanite* project builds on *DROID* and *Apache Tika* to provide a rich format identification and characterization system. It aims to make it easier to run identification and characterisation at scale, and helps compare and combine the results of different file format identification tools.

3.2.2 Action services

3.2.2.1 arc2warc-migration-cli[16]

The arc2warc-migration-cli artefact is part of the Hawarp[17] collection of web archive record processing components based which are designed to work together with Apache Hadoop.

3.2.2.2 raw2nexus

The scientific data produced by the *ISIS* instruments at the Rutherford Laboratory (STFC) produce raw data which are stored together with and associated metadata and log files in a non-standardised way.

The *NeXus* format is a common data format for neutron, x-ray and muon science. It has been developed as an international standard by scientists and programmers representing major scientific facilities in Europe, Asia, Australia, and North America in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data.

[13] https://wiki.duraspace.org/display/FF/Fedora+Repository+Home
[14] https://code.google.com/p/fits/
[15] https://github.com/willp-bl/nanite
[16] https://github.com/openplanets/hawarp/tree/master/arc2warc-migration-cli
[17] https://github.com/openplanets/hawarp

Based on the Mantid[18] analysis software, which is able to analyse both raw & NeXus formats, software called raw2nexus was developed to migrate from raw to NeXus. This software is developed and applied by STFC and is fully tailored to the data sets being created by this institution.

3.2.3 Quality assurance services

3.2.3.1 Flint[19]

Flint is a framework to facilitate a configurable file format validation. Its underlying architecture is based on the idea that file/format validation almost always has a specific use-case with concrete requirements that may differ from a validation against the official industry standard of a given format.

3.2.3.2 Jpylyzer[20]

Jpylyzer is a validator and feature extractor for JP2 images (the still image format that is defined by JPEG2000 Part 1 - ISO/IEC 15444-1). Jpylyzer was specifically created to check if a JP2 file actually conforms to the format's specifications. Additionally Jpylyzer is able to extract the technical characteristics of each image.

3.2.3.3 Matchbox[21]

Matchbox is a very computing intensive quality assurance component which uses a *SIFT* (Scale Invariant Feature Transform) feature detector to determine key points in an image, which are then later used to compare the image with other images in the collection.

3.2.3.4 xcorrSound[22]

The xcorrSound package contains four tools for audio file analysis; 'overlap-analysis' detects overlap in two audio files; 'waveform-compare' compares two audio files and outputs the similarity; 'sound-match' finds occurrences of a smaller audio file (e.g. a jingle) within a larger audio file or an index of audio files; 'sound-index' builds an index for sound-match.

3.2.3.5 Pagelyzer[23]

Pagelyzer is a tool for comparing two versions of a web page in the context of for web archiving. The tool is based on a combination of structural and visual comparison methods, a visual similarity measure designed for Web pages that improves change detection, and a supervised feature selection method adapted to Web archiving. A Support Vector Machine model is trained with vectors of similarity scores between successive versions of pages. The trained model then determines whether two versions, defined by their vector of similarity scores, are similar or not.

[18] http://www.mantidproject.org/main_Page
[19] http://openplanets.github.io/flint/
[20] https://github.com/openplanets/jpylyzer
[21] http://www.scape-project.eu/leaflets/matchbox-the-duplicate-image-detection-tool
[22] https://github.com/http://openplanets.github.io/scape-xcorrsound/
[23] http://openplanets.github.io/pagelyzer/

3.3 Preservation Watch

3.3.1 Plato[24]

The preservation planning tool Plato is a decision support tool that implements a solid preservation planning process and integrates services for content characterisation, preservation action and automatic object comparison in a service-oriented architecture to provide maximum support for preservation planning endeavours.

3.3.2 C3PO[25]

C3PO is a tool which helps to visualise statistical information based on metadata extracted from digital objects. It consists of two parts; a CLI (Command Line Interface) application and a Web Application. The CLI app reads in and processes *FITS* metadata files and stores them in a document store. The Web Application offers visualisation, filtering, export of the data and much more.

3.4 Commercial products

3.4.1 ExLibris Rosetta[26]

The Rosetta demonstration for SCAPE is hosted at ExLibris and allows users viewing a newspapers sample that is represented as Intellectual Entities (IEs), one Intellectual Entity per newspaper. Each IE has 2 different Representations.

3.4.2 Microsoft Azure[27]

SCAPE Azure services is available as a web portal with functions that support a four step workflow for batch-mode document conversion: ingest and characterisation of document collections, conversion, comparison, and reporting.

4 Preservation user stories

The preservation scenarios collected and described in the context of the *Testbeds* sub-project explain preservation issues related to specific data sets that are used to derive requirements for the development and the evaluation of a number of demonstration key assets.
During the refinement of these preservation scenarios, the already existing preservation scenarios have been transformed into *User stories* - a short and succinct high-level statement of the preservation issue encountered by a partner institution - and Experiments - a unit of work that combines a dataset, one or more *Preservation Component*s, a workflow and a processing platform that can be used to evaluate SCAPE technology and provide evidence of scalable processing. The following provides an overview of the *User stories* for the different *SCAPE Testbeds*, which shall be implemented in the context of the described demonstrations.

4.1 Web content Testbed

The Web Content Testbed *user stories* represent the real world challenges in the area of web content preservation. The specific challenge of web archives compared to other application areas is the

[24] http://www.ifs.tuwien.ac.at/dp/plato/intro/
[25] http://ifs.tuwien.ac.at/imp/c3po
[26] http://www.exlibrisgroup.com/category/RosettaOverview
[27] http://azure.microsoft.com

heterogeneity of its content, especially the huge variety of digital objects of different file formats because a web archive contains text content, images, audio, and video content where only a minor part strictly follows the corresponding file format specifications. Therefore these *user stories* explain the high level requirements that ensure the future accessibility to the content.ARC to WARC Migration

4.1.1 ARC to WARC Migration

The ARC to WARC migration *User Story* relates to the question how web archive content should actually be stored for the long term. Originally, content was stored in the ARC format, a format developed by the *Internet Archive* together with the *Heritrix* Web Crawler software which produced these files as the default persistent storage file format for crawled web sites. The format was designed to hold multiple web resources aggregated in a single – optionally compressed – container file. But this format was not supposed to be an ideal format to store content for the long term, for example, it was lacking features that allow adding contextual information in a standardised way. For this reason, the new WARC format as an ISO Standard was created to provide additional features, especially the ability to hold harvested content as well as any meta-data related to it in a self-contained manner.

4.1.2 Comparison of web snapshots

Web Archiving means capturing web content, which is heterogeneous, complex and highly ephemeral. To assure the quality of a web archive means ensuring the ability to display archived web pages in an authentic manner. This concerns the process of collecting web resources and storing them in ARC and WARC container files as well as the environment which is used to provide access to archived web content. For this reason, it is crucial to enhance capture methods and to identify and solve issues related to displaying archived web pages.

4.1.3 File Format Identification and Characterisation of Web Archives

Web archive data is very heterogeneous. Memory institutions doing web archiving have an implicit or explicit policy that determines which type of material is collected. Therefore, data may be text documents in all kinds of text encoding, html content loosely following different HTML specifications, audio and video files that were encoded with a variety of codecs, etc.

In order to take any decisions in preservation, it is indispensable to have detailed information about the content in the web archive, especially those pieces of information that preservation tools depend on. This can lead to different views regarding the prioritisation of which type of content and which properties of that content needs special attention in what concerns preservation planning.

The main issue that we are dealing with in this deliverable is the question how these actions can be achieved using workflows that process large amounts of web archive content at scale.

4.2 Large scale digital repositories Testbed

The Large Scale Digital Repositories *user stories* (work package TB.WP.2) represent the real world challenges in the area of preserving large collections of digital objects in content repositories. The specific challenge of this Testbed is the large number of items contained in digital collections which are managed and preserved by a repository software with defined data ingest and data manipulation procedures.

4.2.1 Large Scale Audio Migration

An institution which holds a large audio collection needs a digital preservation system that can migrate large numbers of audio files from one format to another and ensure that the migration is a good and complete copy of the original.

4.2.2 Large Scale Image Migration

An institution which holds collections consisting of large numbers of image files, a digital preservation system is required which is able to migrate images from one format to another, ensuring that the migrated images conform to our institutional profile, that no image data is lost and that the migration is cost effective (saving storage for example).

4.2.3 Policy-Driven Identification of Preservation Risks in Electronic Document Formats

As a Digital Library holding a large number of electronic documents from various sources a digital preservation system is required which can help to identify preservation risks within these files to ensure that the institution is aware of preservation risks and can sustainably manage the content.

4.2.4 Quality Assurance of Digitized Books

As a cultural heritage institution, a digital preservation system is required that can identify books within a large digital book collection that contain duplicated book pages and inform us of the pages within those books that are duplicate images.

4.2.5 Validation of Archival Content against an Institutional Policy

A memory institution must be able to ensure that content in repositories conforms both to its file format specification and (where appropriate) the profile of that format as specified by the institutional policies. This is to ensure that the content conforms to existing preservation policies and also that content we ingest is acceptable within the bounds of those policies.

4.3 Research datasets Testbed

The Research Data Sets application area refers to a domain from the scientific community where massive amounts of data are being generated. In order to be able to refer to these data sets in publications, the scientific community is searching for ways to store (at least “significant” parts of) these data for the long term. In the SCAPE project, the Science and Technology Facilities Council (STFC) represents this community and provides data from physical experiments. The following *user stories* represent the main issues that have been raised to derive solutions for the development of large-scale data processing.

4.3.1 Migration from Local Format to Domain Standard Format

As the content holder/manager of scientific data held in a local format, it is desirable to hold data in a domain standard format to reduce the risks of losing the ability to read/use and reuse the data contained within the file format.

In this user story, the concrete goal is to migrate data in the RAW format together with associated metadata and information from log files to the NeXuS format which represents the domain standard format.

4.3.2 Preserving the Context and Links to Research Data or Preserving Research Objects

This *User Story* involves the building and maintenance of an Investigation Research Object with specific specialisations for STFC Research Data. The purpose is to bring all of the relevant objects together so that the Research Object as a whole can be understood and that the data object to remains useable/reusable over the long term.

For example, for data coming from the *ISIS* facility, the Investigation Research Object is an aggregation object which collects both the digital object under consideration and artefacts such as raw experimental data, metadata, software, derived data etc. This object is growing over time and the *User Story* is about the challenges to preserve context and links to external information entities in a sensible manner.

As this *User Story* is especially interesting to a wider public, it was selected as an appropriate candidate for demonstration, it will therefore be presented at the Digital Libraries 2014 conference in London (September 2014).

5 Demonstrations

For the purposes of this document, there are three aspects to the term "scope of demonstration":

1. the *Testbeds' Local Instance*s are the *demonstration environment*s,
2. the *preservation asset*s are the outcomes of the SCAPE project to be demonstrated, and
3. the *preservation stories* are the institutional context in which requirements for the development of solutions were derived during the first two years of the project.

The decision to build several instances for deploying and evaluating the *SCAPE Platform* together with selected *SCAPE component*s needed in the context of different *Preservation Stories* was made after the first two months of the project. At this point the decision was taken to build *Local Instance*s of the *SCAPE Platform* close to the data of the institutions hosting the content.

First of all, a major obstacle regarding the option of transferring large data sets from libraries and data centres to a central instance was the legal constraints related to collections which were protected by copy right law. Furthermore, there was the operational advantage that by choosing Local Instances, it was possible to avoid moving large amounts of data from each institution to a *central instance*, and instead bringing the *SCAPE Platform* to the data. Finally, the option of creating various *Local Instances* brought the advantage that it was possible to present a set of highly heterogeneous institutional environments, data, and preservation stories that show the possibilities of using the SCAPE outcomes in other institutions.

5.1 British Library

The British Library is leading the Large Scale Digital Repository Testbed (WP16/TB2) and participating in the Web Content Testbed (WP15/TB1) and Research Datasets Testbed (WP17/TB3). The British Library holds a large collection of different types of data and the se datasets are being used within its Testbed work.

The main preservation *user stories* that are being worked on in this environment are:

- 4.1.3: File Format Identification and Characterisation of Web Archives
- 4.2.2: Large Scale Image Migration
- 4.2.3: Policy-Driven Identification of Preservation Risks in Electronic Document Formats
- 4.3.4: Identification, validation and *checksumming* of a complex corpus

5.1.1 Planned Events and Visiting opportunities

5.1.1.1 Event One: Internal staff talk

An internal meeting was held at the British Library on 12 March 2014, which was open to all staff. An overview of the project was given along with details about the work undertaken by the British Library. Additionally, specific outputs that may be useful to staff, such as *Jpylyzer*[28], *Flint*[29], *Nanite*[30], *xCorrSound*[31], *C3PO*[32] and *Matchbox*[33] were outlined.

[28] See section 3.2.3.2
[29] See section 3.2.3.1
[30] See section 3.2.1.2
[31] See section 3.2.3.4
[32] See section 3.3.2

5.1.1.2 Event Two: SCAPE Information Day

The information day was a full day event, held in the British Library Conference Centre, London, on 14th July 2014.

The event was heavily advertised on Twitter, as well as via the OPF[34] mailing lists and normal SCAPE dissemination channels. Some attendees were personally invited. In total seven different institutions/organisations were represented by nine attendees.

The presentations were primarily given by William Palmer, with additional presentations by Peter May and Alecs Geuder of the British Library, and by Carl Wilson of Open Planets Foundation. Some live demonstrations were on the agenda and other technologies/outputs were shown via diagrams and pictures.

After a short introduction and formalities, Peter May began by giving an overview of the SCAPE project and architecture, along with a high level introduction to some of the SCAPE outputs that would be presented in subsequent presentations.

William Palmer then gave an introductory overview of *Apache Hadoop*, followed by describing the Testbeds workflows/experiments and results of Testbeds activity carried out at the British Library.

Alecs Geuder presented work on detecting Digital Rights Management (DRM) in PDFs and EPUBs with *Flint*, a modular and extendible file/format validation framework. *Flint* can assess whether or not a file contains features allowed within a particular policy or not, for example if a PDF contains JavaScript. He demonstrated how to use *Flint* through the GUI application, and how straight forward it was to enable/disable individual checks in profiles.

William then talked about JPEG2000 files, and the complexity of encoding parameters required. He demonstrated how *Jpylyzer* could be used to assess whether a JP2 file was valid in respect to the file format specifications. Additionally, William discussed how *Jpylyzer*'s output could be used in conjunction with *Schematron* to verify that encoded files match an organisation's JPEG2000 profile. It was noted that these checks do not verify whether or not the actual image data is readable or not.

Next, William presented information about *Nanite*, a tool for performing identification and characterisation on web archive data (see 3.2.2.3). *Nanite* is written in Java and its tight coupling to Hadoop gives it its speed. Details of its performance on the British Library cluster/data were given, along with information about its use at The State and University Library (SB), where it was demonstrably faster than previous tools used. The characterisation outputs from *Nanite* are created to be compatible with, and directly importable to, *C3PO*.

For the next presentation William talked through Roman Graf's latest slides about *Matchbox*'s finger detection and cropping error detection. The presentation detailed how *Matchbox* could be used for detecting duplicates within a set of data, or matching pairs of pages from two sets of scans of the same book. Also shown were slides about new work on finger detection in scans, and cropping error detections (where a scanned image of a book has been incorrectly cropped).

Carl Wilson then presented and gave a live demonstration of *C3PO*, a tool for visualising characterisation information.

Finally, a number of other SCAPE outputs were presented in brief, such as *RODA/Fedora 4*, *Plato*, *SCOUT*, *Taverna*, *Pagelyzer* and *xCorrSound*.

During the introduction participants were invited to ask questions throughout the day, to keep the event as interactive as possible. Indeed, the attendees did ask a lot of questions which helped them better understand and it demonstrated their interest in the SCAPE outputs.

---

[33] See section 3.2.3.3

[34] Open Planets Foundation - http://openplanetsfoundation.org

5.1.2 Event two details and agenda

- Date: 14 July 2014
- Time: 9:30 — 14:50
- Location: British Library, London, UK
- Agenda
    - 12:30-12:40 Welcome drinks (tea & coffee)
    - 10:00-10:10 Welcome and introductions (William Palmer)
    - 10:10-10:50 Introduction to the SCAPE project and overview of SCAPE project architecture (Peter May)
    - 10:50-11:10 Large scale processing with Hadoop (William Palmer)
    - 11:10-11:30 Detecting DRM in PDFs and EPUBs with Flint (Alecs Geuder)
    - 11:30-11:50 Using *Jpylyzer* and *Schematron*[35] for validating JPEG2000 files (William Palmer)
    - 11:50-12:00 Questions / open discussion (William Palmer)
    - 12:00-12:45 Lunch
    - 12:45-13:05 Characterising content in web archives with *Nanite* (William Palmer)
    - 13:05-13:25 Duplicate image detection with *Matchbox* (William Palmer)
    - 13:25-13:45 Visualising characterisation information with *C3PO* (Carl Wilson)
    - 13:45-14:30 Other SCAPE outputs available (William Palmer)
    - 14:30-14:45 Questions / open discussion (William Palmer)
    - 14:45 Close

5.1.3 Conclusion

The Information Day had attendees from several institutions/organisations, which in itself demonstrates the interest in the SCAPE Project. Throughout the day there were many questions about the SCAPE outputs that again showed the interest there is in the project. Overall the information day event was well received and productively disseminated knowledge of SCAPE to an external audience.

5.2 Internet Memory Foundation

The Internet Memory Foundation is involved in SCAPE as a content holder as well as to bring its expertise in terms of web archiving. The institution participated in the Web Content Testbed work package (TB.WP.1) as well as the Platform (PT.WP.1) and Quality assurance work packages (PC.WP.3).

As a content holding hosting the SCAPE central instance and holding a large amount of web content, the institution was responsible for the planning and the execution of large scale tests of tools enhanced or developed during the SCAPE project. The Internet Memory Foundation was also involved within the user group activities and participated to the dissemination of SCAPE outcomes in the preservation community.

In order to be confident that a website was preserved correctly, a digital preservation system is required that can automate the comparison of the two Web Snapshots - for example a harvested copy and a previous harvested copy that has been manually verified as an accurate representation of the site. This allows making sure that web content was successfully harvested.

[35] http://schematron.com

5.2.1 Planned Events and Visiting opportunities

The first event organised at IMF targeted a web archiving partner institution of the IMF. It was therefore not advertised and therefore took a rather informal format.

The main focus of the event was to offer an overview of the SCAPE project and to share information about ongoing tests at IMF and what we thought could be added to our web archiving production workflow.

The second event was a half a day event that took place on the 4th of July at the IMF Paris offices. It welcomed institutions that had requested to visit IMF following the SCAPE dissemination about content holders *Demo Day*s.

The main focus was again to introduce the project as well as to present the IMF work and processes and demonstrate how some of the tools developed or enhanced within SCAPE could integrate the IMF production workflows.

Most presentations were co-presented by Stanislav Barton and Leïla Medjkoune. The Foundation work was introduced by Julien Masanès and Leïla Medjkoune.

After an introduction of the Foundation, and the formal introductions, the first presentation (Presentation of scape project and outcomes) was made using the freely available SCAPE slides.

Through these slides, we offered an overview of the SCAPE projects as well as a summary of preservation challenges the project is willing to tackle. We insisted on the fact that SCAPE is based upon real use cases defined by content holders such as National Libraries. Once the project structure and goals were defined, we looked at the list of tools available and presented these one by one, asking each time if such or such tool/system would be of any use for each participants. This led to an interesting discussion concerning the various SCAPE components. The participants were especially interested in knowing more about content holders real use cases, such as for instance the use of *xCorrSound*, *Pagelyzer* or *Scout*.

The second presentation was made by Stanislav Barton. He provided an overview of the *Central Instance* of the *SCAPE Platform* as well as examples of potential preservation use cases. He also presented the tests in progress as part of the Web Content Testbed work package (TB.WP.1) and the evaluation of tools such as *Pagelyzer* or *Apache Tika*. This led us to a presentation of the *SCAPE Platform*. Overall, this presentation showed the benefits of using such a platform and shared performance figures. We confirmed that the scalability meant an issue for most institutions and that sharing infrastructure designs was of a great use to other institutions, as many questions and interest arose from this presentation.

The last presentation was about the *Pagelyzer* and was unfortunately quite brief due to time constraints. We first explained the issue we are trying to address and summarised work made so far in collaboration with the UPMC team. We then shared our use case and the implementation on our platform as well as performance figures and scoring quality.

5.2.2 Event details and agenda

- Date: 04 July 2014
- Time: 13:30 — 16:30
- Location: Internet Memory Foundation Paris offices, 45 Ter, rue de la Révolution, 93100 Montreuil, France
- Agenda
    - 12:30-12:40 Welcome and Introductions
    - 12:40-13:40 Lunch
    - 13:40-14:15 Presentation of scape project and outcomes
    - 14:15-15:15 Presentation of IM platform and the scape central instance
    - 15:15-15:40 Coffee break

- 15:40-16:10 Presentation of *Pagelyzer*
- 16:10-16:30 Questions / Wrap up
- 16:30 Close

5.2.3 Conclusion

The events held were very interesting and showed a real need and interest from preservation institutions. The institutions welcomed at IMF for both events had very different profiles, from the traditional archive to the broadcast archive. We also received a service provider specialised in preservation systems. This allowed a very diverse discussion about many of the tools and systems developed or enhanced during the SCAPE project. We hope for further exchange and discussions with all participants. We also advertised the London workshop that seemed of great interest to participants and whom said they would most probably attend the workshop. Finally, we provided contact point to participants interested in some of the tools implementation within SCAPE, such as the xcorrSound and SCOUT at SB.

5.3 Austrian National Library

As the work package lead of the Web Content Testbed (TB.WP.1) the role of the Austrian National Library was to coordinate the demonstration of web-archiving tools. Furthermore, as participant of the Large Scale Digital Repositories Testbed (TB.WP.2), the Austrian National Library also presented large scale digital repository tools. The focus was on applications for identification and characterisation of digital objects, information extraction and the general evaluation of the *SCAPE Platform*.

5.3.1 Planned Events and Visiting opportunities

ONB had one request for a visit with the purpose of general information exchange with the National Library of the Czech Republic on the 22nd of November 2013. Even though the meeting was a general purpose meeting related to long-term preservation and related to building digital repositories, it was a convenient opportunity to present outcomes of the SCAPE project in this context. At this time ONB's Local Instance of the SCAPE Platform was installed and initial versions of the large-scale workflows developed in the work packages TB.WP.1 and TB.WP.2 were available. The experiences and lessons learnt from these activities were communicated during this meeting.

Additionally, ONB organized a half-day demonstration event about the SCAPE Platform on the 5th of May 2014 in the Oratorium, a room used for presentations at the Austrian National Library. The event, held in German, was predominantly attended by members of the institution although a few members of other organisations were also present.

The information day was publicized through the normal SCAPE dissemination channels and on the SCAPE Confluence Wiki. Guests for this small event, however, were obtained through personal invitations; non-involved parties from the ONB as well from outside institutions were contacted via email. At the Austrian National Library, invitations were sent to colleagues in IT-Services, at the Web@rchive Austria, in Digital Services and in the Research and Development Department. External persons from institutions in Vienna known to be interested in the topic of long term digital preservation were directly contacted as well. These institutions included the University of Vienna, the Technical University of Vienna, the *Österreichisches Staatsarchiv* (Austrian National Archives) , the Federal Computing Centre (BRZ), and the *Bundeskanzleramt für Kunst und Kultur* (Office of the Federal Chancellor for Art and Culture). In the end, a total of fourteen participants took part in the demonstration event. All of the departments invited to the event from the ONB were represented. In addition, there were a total of three external participants, coming from the Austrian National Archives and the University of Vienna.

Introductory words were spoken by Max Kaiser, Head of the Research and Development Department at the Austrian National Library. In addition to Sven Schlarb from the SCAPE project at the Austrian National Library, three other SCAPE members were invited to talk: Roman Graf and Rainer Schmidt from the Austrian Institute of Technology (AIT) and Krešimir Đuretec from the Institute of Software Technology and Interactive Systems at the Technical University of Vienna. Each speaker gave a half-hour talk about one aspect of the project. The event agenda and slides from the presentations have been published on the SCAPE Confluence Wiki.

After Max Kaiser's welcoming words, Sven Schlarb provided an overview of the SCAPE-Project. This included informing the audience about SCAPE's "big data" context, about the structure of project (Dissemination, *Testbeds*, Platform, Planning and Components) and about central SCAPE topics such as data processing, the software tools developed in the project, the planning of long-term archiving, and the components of the SCAPE architecture.

This general introduction to the SCAPE project was followed by Rainer Schmidt's (AIT) overview of the *SCAPE Platform* Sub-project. Schmidt focused on three main areas: the Execution Platform, the MapReduce Tool Wrapper (*ToMaR*) and Workflow Support. In the first area of the presentation dedicated to the Execution Platform, Schmidt explained the reasons for wrapping sequential tools, writing a custom MapReduce application and using languages such as Hive and Pig. The second part of his presentation dealt specifically with topics related to *ToMaR* such as the Hadoop Streaming API and the SCAPE Tool Specification Language. Schmidt also provided examples and discussed specific use cases. The third area explained the workflow concept, talked about workflow engines (*Taverna workbench*, *Taverna server*, *Apache Oozie*) and discussed their compatibility with *ToMaR* and MapReduce workflows. To conclude his presentation, Schmidt referred to the source project and README that can be found on *Github* and talked about the next steps in the Platform Sub-project.

Sven Schlarb gave an overview of the different application scenarios at the Austrian National Library. Related to the Web Archiving area he presented digital object processing workflows to determine the characteristics of files stored as web archive container files. Related to the Austrian Books Online project, he explained how different outcomes of the SCAPE project can be used to support quality assurance in the context of large digitisation projects.

"Quality Assurance Tools for Document Image Collections" was the title of Roman Graf's presentation. After a short introduction about reasons for quality assurance in the context of the SCAPE project, the talk was divided into five main topic areas:

- Quality Assurance challenges of document image collections
- The *Matchbox* tool for duplicate page detection
- The blank page detection method
- The tool for finger detection on scans
- The tool for cropping error detection

The presentation highlighted, among other things, the benefits of the *Matchbox* tool and challenges and solutions in the case of finger detection workflow for improving the OCR of manually scanned texts. Examples of correct detection of fingers on the digital images manuscripts as well as of false positives and positive negatives were shown. An outcome of the finger detection *User Story* was that the classification accuracy of the finger detection algorithm was at 86 percent, making it "very promising for making the digitization process more reliable and for ensuring the quality of digital collections". The tool for cropping error detection was mentioned as a further tool for improving the long-term archiving of digital collections.

The last presentation of the event was held by Krešimir Đuretec and dealt with *SCAPE Preservation Planning and Watch*. He talked about measures in digital preservation such as the format, size,

nature, and cost of data collections which affect the long-term preservation of a collection. In addition, he explained how Preservation Planning and Watch makes use of such measures for collection profiling, which form the basis of preservation planning in general. The tools and methods discussed in this context include Crafty Content Profiling of Objects (*C3PO*), the Preservation Monitoring System Scout, the creation of controlled vocabularies and ontologies for preservation ecosystems, and the preservation planning tool Plato. It was discussed how SCAPE Preservation Planning and Watch brings these tools and methods together with the policies of libraries and cultural heritage institutions to aid in decisions concerning preservation strategies.

5.3.2 Event details and agenda

- Title: SCAPE Information Day and Demo Event
- Date: Monday, 5th of May, 2014
- Time: 13:30 to 16:30
- Location: Oratorium, Austrian National Library, Josefsplatz 1, 1015 Wien
- Agenda
    - 13:30 Max Kaiser, ÖNB: Greeting
    - 13:40 Sven Schlarb, ÖNB: Introduction to the SCAPE Project
    - 14:00 Rainer Schmidt, AIT: The *SCAPE Platform*
    - 14:30 Sven Schlarb: SCAPE at the Austrian National Library
    - 15:00 Coffee break
    - 15:20 Roman Graf, AIT: The *Matchbox* Tool
    - 15:50 Krešimir Đuretec, TU Wien: Preservation Planning and Watch

5.3.3 Conclusion

To conclude the report on the ONB's demonstration activities, it can be said that the successes of the project in the areas of tools and technology around the *SCAPE Platform* (PT subproject) and Preservation Watch technology (PW subproject, mainly the tools *C3PO* and Plato), as well as selected tools from the PC subproject such as *Matchbox* and Hawarp and the application of several characterisation and identification tools such as *Apache Tika*, *DROID*, and *FITS*, were demonstrated to other members of the hosting institution as well as to the external parties during the SCAPE information event on the 5th of May at the Austrian National Library.

These successes were shown in the context of the preservation *User Stories* developed in the *Testbeds*, namely in the areas of Web Archiving (TB.WP.1) and Large Scale Digital Repositories (TB.WP.2). All of the slides regarding the demonstrations can be viewed on the SCAPE Confluence Wiki.

5.4 State and University Library (Statsbiblioteket)

SB contributed to work packages TB1 and TB2 and defined and executed large scale experiments. As a content holder of large amounts of data, and also being able to utilise its development department, it is possible to implement experiments that reflect real life scenarios.

In this context SB used web-archiving tools and large scale digital repository tools for identification and characterisation of digital objects, information extraction and general evaluation of the *SCAPE Platform*.

The stories included in the demonstration event are listed below. They are further described in the presentation as well as the following section.

- 4.2.1 Large Scale Audio Migration - Migration of audio files based on xcorrSound
- 4.1.3 File Format Identification and Characterisation of Web Archives : Identification and feature extraction of web archive data based on *Nanite*
- 4.2.5 Validation of Archival Content Against an Institutional Policy : Policy driven validation of JPEG 2000 files based on Jpylyzer

5.4.1 Planned Events and Visiting opportunities

The SB held a half-day information and demonstration event about the SCAPE project on the 25th of June 2014 at the State and University Library, Aarhus.

Statsbiblioteket (SB) welcomed a group of people from The Royal Library, The National Archives, and Danish e-Infrastructure Cooperation on June 25, 2014. They were invited for our *SCAPE Demo day* where some of SCAPE's results and tools were presented. Bjarne S. Andersen, Director of Digital Preservation Technology, welcomed everybody and then our software engineers presented and demonstrated SB's SCAPE work.

The day started with an *introduction to the SCAPE project* by Per Møldrup-Dalum, including short presentations of some of the tools which would not be presented later in a demo. This triggered questions about how to log in to *Plato* (see 3.3.1).

Per Møldrup-Dalum continued with a presentation about Hadoop, its applications at the SB, and how it has proven really useful for large-scale digital preservation. In addition, the specifications of the Hadoop-Cluster were mentioned: the SB uses an Isilon Scale-Out NAS storage cluster, which enables different experiments on the four 96GB RAM CPU nodes each with a 2 Gbit Ethernet interface.

Bolette A. Jurik told the *User Story* about how the SB had wanted to migrate audio files using Hadoop. The files were supposed to be migrated from mp3 to wav. Checking this collection using Plato gave the SB the result, 'Do nothing', which meant that it would be best to leave the files in the mp3 format. The SB still wanted to perform the experiment, however, in order to test that it has the tools to migrate, extract and compare properties, validate the file format and compare the content of the mp3 and wav files, and a scalable workflow can be created for this. Since the SB did not have a tool for the content comparison, it developed one, *xcorrSound waveform-compare*. The audience was also informed about a competition Bolette started among her colleagues to create "migrated" sound files with problems that the tool would not find.

After this presentation, Per Møldrup-Dalum described the experiment: *Identification and feature extraction of web archive data based on Nanite* in his talk. The test was to extract different kinds of metadata (like authors, GPS coordinates for photographs etc.) using *Apache Tika*, *DROID*, and libmagic[36]. The experiment was run on the *Danish Netarchive* (archiving of the Danish web – a task undertaken by The Royal Library and SB together). For the live demo, a small job with only three ARC files was used – taking all of the 80,000 files in the original experiment would have lasted 30 hours.

For the demonstration, an analysis of the original *Nanite* experiment was done live in Mathematica[37]; some interesting facts and artefacts were presented such as the number of unique MIME types in the 80,000 ARC files, the 260,603,467 individual documents to 1384

[36] http://sourceforge.net/projects/libmagic
[37] http://www.wolfram.com/mathematica

different MIME types reported by the HTTP server at harvest time, the fact that *DROID* counted 319 MIME types, and Tika counted 342 MIME types.
It was shown how an extreme shift in performance is due to SB's involvement in the SCAPE project. Two years ago, the SB concluded a job that had run for 15 months - 15 months of *FITS* characterising, 12TB of web archive data - the experiment with Nanite characterised 8TB in 30 hours.
After sandwiches and a quick tour to the library tower, Asger Askov Blekinge took over to talk about *Integrating the Fedora based DOMS repository with Hadoop*. He described Bitmagasinet (SB's data repository) and DOMS (SB's Digital Object Management System based on Fedora) and how our repository is integrated with Hadoop.
The last demo was presented by Rune Bruun Ferneke-Nielsen. He described the policy driven validation of JPEG 2000 files based on Jpylyzer and performed on SB's Newspaper digitization project. One of the visitors at the demo is working at The Royal Library with the NetArchive and would like to make some crawl log analyses. These could perhaps be processed by using Hadoop and we would like to discuss to see if our two libraries can work together on this.

5.4.2 Event details and agenda

The agenda is publicly accessible[38], and the slides are all available on the wiki (linked by the corresponding agenda items) and on Slideshare.

- Title: SCAPE Information Day at the Danish State and University Library
- Date: 25 June 2014
- Time: 10:00 — 14:30
- Location: Meeting room 3.61, State and University Library, Victor Albecks Vej 1, 8000 Aarhus C
- Agenda
    - 10:00 Welcome by Bjarne Andersen, head of Digital Preservation Technology
    - 10:10 Introduction to the SCAPE project by Per Møldrup-Dalum
    - 10:30 Hadoop and its applications at the State and University Library by Per Møldrup-Dalum
    - 11:00 Demonstration 1: Migration of audio files based on xcorrSound by Bolette Ammitzbøll Jurik
    - 11:30 Demonstration 2: Identification and feature extraction of web archive data based on Nanite by Per Møldrup-Dalum
    - 12:00 Lunch
    - 13:00 Integrating the Fedora based DOMS repository with Hadoop by Asger Askov Blekinge
    - 13:30 Demonstration 3: Policy driven validation of JPEG 2000 files based on Jpylyzer by Rune Bruun Ferneke-Nielsen
    - 14:00 Questions and answers

[38] wiki.opf-labs.org/display/SP/Agenda+of+the+SCAPE+Information+Day+at+the+State+and+University+Library

- 14:30 Farewell

Slides

- *Introduction to the SCAPE project by Per Møldrup-Dalum*[39]
- *Hadoop and its applications at the State and University Library by Per Møldrup-Dalum*[40]
- Demonstration *1: Migration of audio files based on xcorrSound by Bolette Ammitzbøll Jurik*[41]
- Demonstration *2: Identification and feature extraction of web archive data based on Nanite by Per Møldrup-Dalum*[42]
- *Integrating the Fedora based DOMS repository with Hadoop by Asger Askov Blekinge*[43]
- Demonstration *3: Policy driven validation of JPEG 2000 files based on Jpylyzer by Rune Bruun Ferneke-Nielsen*[44]

5.4.3 Conclusion

The SCAPE project was presented with emphasis on live demos to guests from three Danish institutions with interest in digital preservation and scalable processing environments. There was a group of six guests from three institutions in total, The Royal Library, The National Archives, and Danish e-Infrastructure Cooperation. Four of the six guests were software engineers, so also it was possible go into some technical details to benefit this group. The event was held using slides in English, but the presentations were in Danish, as all the attendees were also Danish.

5.5 Science and Technology Facilities Council

The Science and Technology Facilities Council has both developed tools (in PC.AS) and tested them in the Research Data Testbed (TB.WP.3). It was the TB3 Workpackage Lead. STFC provides the UK with access to ISIS, a Neutron Spallation Source that enables researchers to probe the structure of matter. The resulting data has been kept by the ISIS department over the 30 years. A point to note for STFC is that preservation activities are viewed as part of a data management lifecycle and there are no parts of the organisation solely dedicated to preservation activities. One aim of the STFC's work has been to widen the types of material which the SCAPE has addressed.

5.5.1 Planned Events and Visiting opportunities

Due to the specialised nature of the first scenario, STFC decided to focus the activities at the different communities within STFC itself. For this reason we organised four demonstration/discussion events. In addition, Catherine Jones has spoken at STFC Staff Fora at

[39] www.slideshare.net/SCAPEproject/scape-scalable-preservation-environments-scape-information-day-25-june-2014scape-information-day-sb-scape-presentation
[40] www.slideshare.net/SCAPEproject/scape-information-day-sb-hadoop-presentation
[41] www.slideshare.net/SCAPEproject/scape-information-day-sb-migration-xcorr-sound
[42] www.slideshare.net/SCAPEproject/scape-information-day-sb-nanite-experiment
[43] www.slideshare.net/SCAPEproject/scape-information-day-sb-integrating-doms-with-hadoop
[44] www.slideshare.net/SCAPEproject/scape-information-day-sb-policy-drivenvalidation

two of the four STFC sites as the science speaker; her topic described the work on creation and preservation of context for scientific data.

5.5.1.1 Event One

The first event was a meeting with ISIS Data Manager responsible for the Computing Data Management infrastructure within ISIS. The aim of this meeting was to highlight the work on these *User stories* and the policy work from PW.WP2. This took place on 21st of March 2014

The informal agenda was as follows:

Catherine Jones: Basic introduction to SCAPE and the Policy Catalogue

Catherine Jones introduced the aims of the SCAPE project and then discussed the SCAPE work on policy creation. The three levels in the framework were described as well as the aim of having machine readable policy. Examples of the catalogue elements were shown to encourage use of this work.

Alastair Duncan: Using Hadoop to migrate raw files to NeXus format

Alastair outlined the approach to using Hadoop and SCAPE tools such as *ToMaR* to scale up the migration of files.

Antony Wilson: Investigation research Objects

Antony introduced the concept of Investigation Research Objects and the types of context that might be added to them. Examples at the time included work from the PANKOS ontology to add descriptions of the scientific technique. The idea of a "data journal" as the front end to the landing pages for the Digital Object Identifier was demonstrated together with the use of Fedora 4 for storing the METS packages for preservation.

The ISIS Data Manager felt that both the policy work and the Hadoop work were interesting but there was no immediate applicable use. The current stance of ISIS is to maintain both raw and NeXus files and to ensure that the analysis software provided is able to read both types of file and so there are no current plans to do large scale format migrations.

On the work of IRO's he felt that the developments we were demonstrating were along the right track. The discussion touched on where the authoritative source of data would be, what different interfaces would be needed for different end-users and whether there was programmatic access to the underlying data. There were further discussions about whether using a triple store caused performance issues.

5.5.1.2 Event Two: Meeting with ISIS Impact Manager

The aim of this meeting was to show the potential for the context and linking work. This took place on the 7th of May 2014.

This demonstration focussed on the Investigation Research Object work and was undertaken by Catherine Jones. The work was shown and a general discussion about whether researchers would find these types of links to be useful; the conclusion was that there needed to be some examples showing the benefit before researchers would be prepared to change their working practices.

5.5.1.3 Event Three: Meeting with the Head of the Centre for Environmental Data Archival

The Centre for Environmental Data Archival (CEDA) is based at STFC but responsible for a different type of data. CEDA was interested in our experience of using Hadoop rather than the specific workflow generated in SCAPE. This meeting was held on the 15th of July 2014. The same format for the talks was used:

Catherine Jones: Basic introduction to SCAPE and the Policy Catalogue

Catherine introduced the aims of the SCAPE project and then discussed the SCAPE work on policy creation. The three levels of the framework were described as well as the goal of having machine readable policy. Examples of the catalogue elements were shown to encourage use of this work.

Alastair Duncan: Using Hadoop to migrate raw files to NeXus format

Alastair outlined the approach to using Hadoop and SCAPE tools such as *ToMaR* to scale up the migration of files. He then showed the results from altering the number of maps and splits and discussed the scalability demonstrated.

Antony Wilson: Investigation research Objects

Antony introduced the concept of Investigation Research Objects and the types of context that might be added to them. Examples at the time included the work from the PANKOS ontology regarding adding descriptions of the scientific technique. The idea of a "data journal" as the front end to the landing pages for the Digital Object Identifier was demonstrated together with the use of Fedora 4 for storing the METS packages for preservation.

We discussed the issue of how to ensure that machine readable policy has actually implemented the written policy, how one can go from low level back to high level. The Head of CEDA was interested in this as the security model they have implemented is complicated and encoded in machine understandable way and so it is hard to check that it is actually implementing written policy. We agreed that policy was important.

Alastair's talk instigated a discussion about effective chunking of large data. The current CEDA data is held on Panasas kit (this is run by another part of Scientific Computing Department for CEDA. It is highly parallelised - http://www.panasas.com). CEDA staff have been running checksums across the 2PB of data - which takes about a week – and they have been experimenting with chunking and so he was interested in Alastair's findings of performance differences with changes in maps and splits. It was agreed that one of the issues of using either Panasas or HDFS is the time taken to get the data onto the file system.

In discussions about the Investigation Research Object work, the Head of CEDA was interested in the way the architecture separated the metadata from the data catalogue from the annotations which add context. This could be used to control who is authorised to make annotations or to be able to judge the provenance of the annotation. This work is building on concepts developed in the JISC project, CLADDIER (2005-2007), to which Catherine and the Head of CEDA contributed, which explored issues of linking data to publications and what publishing data means. So there was some general discussion about advances in technology enabling the implementation of links and the preservation of those links. There was further discussion about not only having the right technology, but also ensuring that understanding of how to will migrate data and metadata to the next system. The Head of CEDA'S feedback on this topic was interesting; he would be interested in seeing further developments.

5.5.1.4 Event Four: Department talk to the Scientific Computing Department.

The aim of this event was to widen the internal knowledge within the department the SCAPE team belong to of the work done by STFC on Hadoop in SCAPE. This meeting was held on the 18th of July 2014. Twenty nine people attended the meeting, twenty-eight were from the Rutherford Appleton Lab and one member from the Daresbury Lab attended using Access Grid Technology. The audience included those supporting the Large Hadron Collider's UK Tier 1 site and those who support high performance computing clusters in the Research Infrastructure Group along with colleagues from the Research Data Group and the Numerical Analysis Group.

Alastair's talk was entitled "The Trials and Tribulations and ultimate success of parallelisation using Hadoop within the SCAPE project"

The talk introduced the SCAPE project and the tools which form the *SCAPE Platform*. Alastair went on to discuss the data migration workflow and the three datasets used for evaluation. He then walked through the experiments he had done with changing the Maps and Splits to get the best results, showing Ganglia plots for the different set-ups and showing that, although the files can't be streamed to get the best out of Hadoop, there was scalability.
There was a technical discussion afterwards about the pros and cons of using Hadoop versus the other types of high performance clusters run within the Department which are designed for doing computations on scientific data. It was noted by Catherine that other parts of the SCAPE project had seen better results using Hadoop for file types that could be streamed.

5.5.1.5 Science Speaker at the STFC Staff Fora

This is a ten minute slot at the end of quarterly staff fora where the STFC Chief Executive and other members of the Executive Board update STFC employees on STFC internal activities. The science slots are designed for a staff member to describe their work to all their STFC colleagues. Catherine spoke at the Rutherford Appleton Laboratory (ca. 500 people) and the Royal Observatory Edinburgh: home of the Astronomy Technology Centre (ca. 20 people). Her talk was entitled "Beyond Bits to the preservation of meaning". It outlined bit preservation; discussed what metadata and context means and how it is dependent on the purpose; discussed the context and links for ISIS experiments by showing an example which has been made for SCAPE purposes and highlighted what the future might look like. SCAPE was identified as the project within which the work on creating and preserving context and links is being done; the STFC SCAPE team were acknowledged on the final slide. The type of questions asked revealed that there was an interest in being able to navigate between different outputs so that these links should be preserved.

5.5.2 Agenda

As these demonstration events were tailored to the specific people we were meeting, there were no formal agendas, see the event details for more information.

5.5.3 Conclusion

Four specialized events addressing four different stakeholder groups were held and two staff fora talks were given to a cross-section of STFC staff. The final two site talks will be given in the autumn. The identified stakeholder groups were as follows:

1. With the ISIS data manager responsible for the data management infrastructure
2. With the impact manager
3. With the head of a domain specific data centre with a preservation remit
4. With the computing colleagues who run high performance computing for scientific research data

All work that STFC has done in SCAPE was introduced and discussed: policy developments, migration workflows which use Taverna and Hadoop, and the creation and preservation of context and links as appropriate to the audience. The area which had the most interest outside the department was the creation and preservation of contextual information. This is not surprising as it is the most widely applicable part of our work and the creation of the links can contribute to information discovery whereas the preservation can keep those links for the long-term.

5.6 Ex Libris

Ex Libris's function in this part of the project was on demonstrating the integration of SCAPE tools and components in Rosetta as a commercial system, which is already in production in many large libraries worldwide. Ex Libris created a SCAPE-dedicated Rosetta instance for testing and demonstrating the functionality of these tools in various Rosetta workflows.
For above testing purposes, Ex Libris created two test workflows, integrating the following tools and components:

- Loader Application: A standalone, java-based tool for feeding SCAPE METS files[45] to the data connector APIs.
- Data Connector APIs: RESTful APIs for loading content into the repository. The API layer can be deployed on the Rosetta application server (JBoss) or an external application server (e.g. Jetty).
- Jpylyzer: JP2000 metadata extractor. Integrated as a metadata extractor plug-in implementation, Rosetta alongside other extractors (e.g. JHOVE, NLNZ, etc.).
- DRMLint: PDF digital rights identifier. Integrated in Rosetta as a Risk Extractor plug-in implementation, advising staff of associated long-term access risks and the necessary steps to address them.

Both workflows utilize both of the first two tools and one of the latter two. Due to administrative constraints concerning the deployment of a large-scale Rosetta environment, limited data sets were created based on original newspapers in TIFF format, provided by the British Library. The images were converted to JP2000 and PDF, and password protection was added to the PDFs. The data was loaded onto two Rosetta instances (a single server and a multi-server).

5.6.1 Planned Events and Visiting opportunities

A summary of Ex Libris's activity in the SCAPE project was presented during the annual Rosetta Advisory Group meeting, held on the 18th of June 2014 in Jerusalem. This event (by invitation only) was attended by representatives of almost all institutions using Rosetta (as well as several institutions interested in using Rosetta) - altogether roughly 50 people.
Rosetta, like all Ex Libris systems, relies on its Developer Network for disseminating tools that utilize APIs and other integration points. Information regarding the *Jpylyzer* and *Flint (*formerly named DRMLint*)* integrations has been published.[46] Information regarding the loader application and using the APIs will be similarly published by the end of 2014, at which time these components will be made available to all Rosetta users.

5.7 Microsoft Research

Microsoft Research focused their effort in the demonstration work package TB.WP.5 on building a demonstration environment which includes outcomes of the SCAPE project.
The SCAPE Azure Platform can demonstrate the execution of semi-automatic document preservation workflow as illustrated in Figure 12.

[45] The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.
[46] https://developers.exlibrisgroup.com/blog/Jpylyzer-Technical-Metadata-Extractor-Plugin; https://developers.exlibrisgroup.com/blog/drmlint-risk-extractor

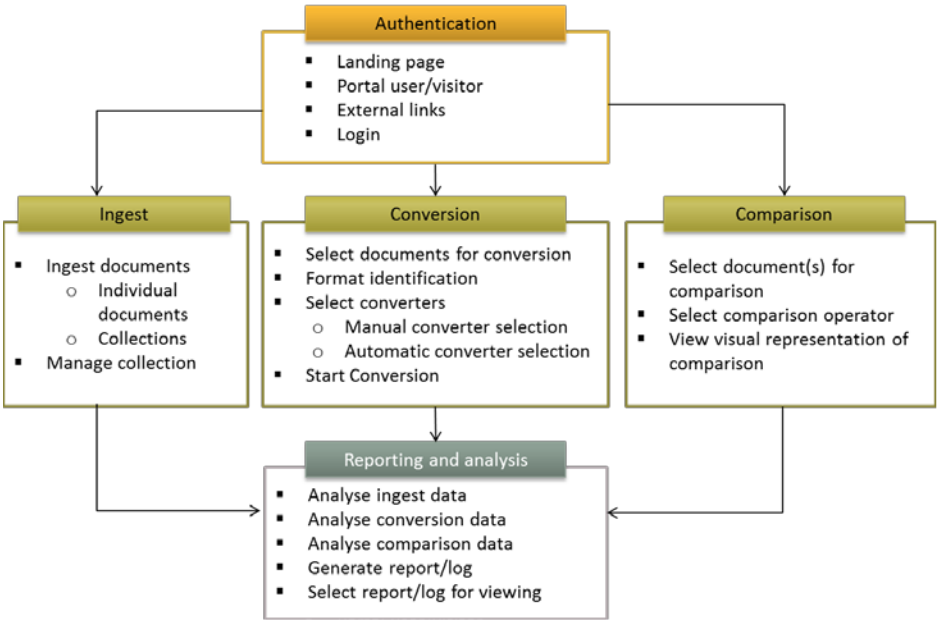

Figure 12 - Semi-automatic document preservation workflow

For demonstration purposes, an interactive workflow for conversion and comparison of office documents was prepared which can be demonstrated via a SCAPE demonstration web application.[47] These demonstrations include various user stories which have been developed together with Testbeds sub-project partners, namely the "Large scale document characterization and identification with *Apache Tika* and *DROID* on SCAPE Azure Platform" user story[48] and the "Characterization and identification on SCAPE Azure Platform" user story.[49]

5.7.1 Planned Events and Visiting opportunities

The following event was part of the demonstration activities of Microsoft Research in the context of TB.WP.5. iCertis is the leading provider of enterprise solutions in the Microsoft Cloud – Windows Azure environment. It offers products and services that address enterprise business needs by leveraging the cloud's elasticity, ubiquity, and availability. iCertis' products include comprehensive solutions for Contract Lifecycle Management and Partner Relationship Management. They are offered through flexible deployment models (on-cloud, on premise, and hybrid) to address diverse requirements for cost, availability, and security.

The objective of the Technology Briefing was to present to iCertis the technical and business opportunities within digital preservation domain and demonstrate the services and tools that have been developed as part of SCAPE.

5.7.2 Event details and agenda

- Title: Technology Briefing to Enterprise Solution Provider: iCertis (http://www.icertis.com/)
- Date: Friday, 7 March, 2014
- Time: 13:00-15:00 PST

[47] http://scapestaging.cloudapp.net:8080

[48] http://wiki.opf-labs.org/display/SP/Large+scale+document+characterization+and+identification+with+Tika+and+DRIOID+on+SCAPE+Azure+platform

[49] http://wiki.opf-labs.org/display/SP/Characterisation+and+Identification+on+SCAPE+Azure+Platform

- Location: Conf Room 121/3475,Microsoft, Redmond Campus, Redmond, WA
- Agenda
    - Introduction
    - Background on Digital Preservation issues and approaches
    - Demonstration of SCAPE Azure services and SCAP Format Migration Portal
    - Discussion of possible collaboration.

5.7.3 Conclusion

iCertis sees business potential in providing preservation services in the cloud but any commitment of resources to technology and business development would require a rigorous assessment of the market potential.

6 Communication channels used to announce demonstration events

In the context of the Demonstration workpackage (TB.WP.5), a page was created on the SCAPE Public Wiki to help promote the possibility to visit a SCAPE partner and to provide more information about both the demonstrations and the project.[50]

The wiki page includes information about the partners involved in the demos, their location, the tools and services they each specialize in, the agendas for the planned demonstration, and other information to help visitors decide where to go.

Mails with link to the wiki were sent out to a number of mailing lists encouraging people to go to an already scheduled event or contact the Dissemination Work Package Lead to set up a demonstration at their own request – e.g. to visit a partner near them or with focus on specific tools and/or services.

- The following mailing lists were used:
- DIGITAL-PRESERVATION@JISCMAIL.AC.UK
- diglib@infoserv.inist.fr
- Open Planets Foundation (OPF) Community Mailing list

An announcement was made for SCAPE's website[51],, and the larger, planned events were included on the website as 'Upcoming events'. The events were also announced in the SCAPE Newsletter[52] and in numerous tweets which were retweeted by others.

7 Conclusion

The goal of the SCAPE Demonstration work package TB.WP.5 was to demonstrate the outcomes of the SCAPE Project to stakeholders in the *Demonstrating Institutions* as well as to interested third parties.

Chapters 2 to 4 gave an overview about what could actually be demonstrated as the outcome of the SCAPE project by the different institutions.

A SCAPE wiki page for demonstration was created to enable the *Demonstrating Institutions* to organise and document demonstration events and to provide information to allow requests for *Scheduled Visit*s. Even though the final demonstration report marks the end of the demonstration

[50] http://wiki.opf-labs.org/display/SP/Demonstrations
[51] http://www.scape-project.eu/news/scape-demos-at-partners
[52] http://bit.ly/Scape8

activities of the Testbeds sub-project, the Wiki allows for the interested parties to gain insight into the SCAPE outcomes and to contact the *Demonstrating Institutions*.
All of the *Demonstrating Institutions* were able to give a demonstration of their demonstration assets in one form (demonstration event) or another (*Scheduled Visits*).
It must be pointed out that the response to the offer of *Scheduled Visits* was not satisfying at the end. There were only requests from two persons, one person from Tessella who ended up attending the BL *Demo Day* and the other person from VIAA in Ghent who visited IMF. As outlined in this report, some of the *Demonstrating Institutions* have organised a *Demo Day* where internal staff and/or external interested parties were invited and SCAPE outcomes were demonstrated in an institutional context. It should be highlighted that these events have been organised exactly to mitigate this risk that there might not be enough response to the offer of *Scheduled Visit*s which were announced via various communication channels. For this reason, a proactive strategy was suggested to the *Demonstrating Institutions* and it was recommended to get in contact directly with national institutions which have an interest in long-term preservation and large-scale data processing topics. According to the Description of Work, these events were not planned in the same way as the training events organised by TU.WP.2. Especially, no tasks related to event evaluation were planned in order to measure the success of these events by evaluating the feedback from visitors, for example. Nevertheless, as it was outlined in this report, internal staff and external visitors who participated in either *Scheduled Visit*s or *Demo Day*s were interested in the SCAPE outcomes and to see the various ways in which SCAPE technology was adopted by *Demonstrating Institutions*. However, as a lesson learnt from planning the demonstration activities, we would state that the proactive invitation to such events as the *Demo Day*s should be planned as targeted invitations right from the beginning and planned in this way in the Description of Work.

8 Glossary

Term	Abbreviation	Definition
Action Service	AS	An action service is a type of a digital preservation service that performs some kind of action on a digital object, e.g. migrating the object to a new file format.
Apache Hadoop		Framework for processing large data sets on a computer cluster. See http://hadoop.apache.org
Apache Oozie		Workflow scheduler for Apache Hadoop jobs. See http://oozie.apache.org
Apache Pig		A high-level language for creating workflows which run on top of Hadoop/MapReduce.
Apache Tika		File format identification tool (MIME type detection). See https://tika.apache.org
ARC		Web Archive Container Format. See also “WARC”. See http://archive.org/web/researcher/ArcFileFormat.php
Characterisation Service	CS	A characterisation service is a type of a digital preservation service that extracts any kind of information from a digital object, as an identifier or file related properties, for example.
Checksumming		In the context of the SCAPE project the term “checksumming” refers to calculating a “small-size string” (checksum) based on the file content which changes even if only minor changes are applied to the file content. If calculated hash value is equal for two files, it means that the content is bit-wise identical. The calculation is done by a checksum algorithm, such as MD5, SHA-1, SHA-256, etc.

Central Instance		The *Central Instance* of the SCAPE Project is an environment where the *SCAPE Platform* (see "SCAPE Platform") together with *SCAPE Components* (see "SCAPE Components") is installed and which is accessible for all project partners. The *Central Instance* was hosted at IMF. The *Demonstrating Institutions* (see "Demonstrating institution") of the SCAPE project decided to create various *Local Instance*s (see "Local Instance") at the *Demonstrating Institutions* in order to avoid the requirement to move large amounts of data to the *Central Instance*.
Component		See SCAPE Component
DataNode		In the context of an Apache Hadoop environment, the DataNode stores data in the HDFS. A functional filesystem has more than one DataNode, with data replicated across them. See "Apache Hadoop". See "HDFS". See https://wiki.apache.org/hadoop/DataNode
Demo Day		One-day presentation at one of the *Demonstrating Institutions* showing SCAPE outcomes in the institutional context. The target group of the Demo Day is internal staff of the institution as well as interested third parties.
Demonstration		In the context of the *SCAPE Testbeds* sub-project, the term "Demonstration" refers to the act of demonstrating the results of the large-scale *Testbeds* for web content, the large-scale digital repositories and the research datasets to third parties.
Demonstration Environment		In the context of the *SCAPE Testbeds*, the term "Demonstration environment" refers to the hardware which was chosen to set-up a computing cluster and the *SCAPE Platform* together with selected SCAPE components used to demonstrate SCAPE project results in the context of specific *user stories* (see "user story") which the institution was responsible for.
Demonstrating Institution		Institutions showing SCAPE outcomes in various institutional contexts. For this purpose, various SCAPE

		Local Instances (see “Local Instance”) have been created to demonstrate the SCAPE Platform together with SCAPE Components applied to various kinds of data sets.
DROID		Software developed by the National Archives (UK) to determine a unique file format identifier (*PUID*, see corresponding glossary entry). *DROID* is a software tool developed by The National Archives (UK) to perform identification of file formats. See http://digital-preservation.github.io/*DROID*/
Github		Default code versioning system used in the SCAPE project. See https://github.com
(SCAPE) Experiment Evaluation		
(SCAPE) Experiment		A unit of work that combines a dataset, one or more *Preservation Component*s, a workflow and a processing platform that can be used to evaluate SCAPE technology and provide evidence of scalable processing
SCAPE Platform		An extensible infrastructure for the execution of digital preservation processes on large volumes of data (using a combination of Apache Hadoop and Taverna).
File Format Characterisation		The process of determining the properties of a file format, for example, the bit depth, colour space, width of an image, the frames per second of a video, etc.
File Format Identification		The process of determining the identity of a file format instance, typically by assigning an identifier, as the *PUID* (see corresponding glossary entry) as a precise identifier or a MIME Type (see corresponding glossary entry) identifier as a vague file type identifier.
FITS		File format characterisation tool which executes various identification and information extraction tools. See http://projects.iq.harvard.edu/fits
Hadoop		See Apache Hadoop.

HBase	HBase	Distributed database on top of Hadoop/HDFS, see https://hbase.apache.org
HDFS	HDFS	Hadoop Distributed File System. This is Hadoop's file system which is designed to store files across machines in a large cluster.
Heritrix Web Crawler		Web crawler engine used to harvest content from the internet and store it in a web archive. The Heritrix Web Crawler was originally developed by the Internet Archive (see corresponding glossary entry). See https://webarchive.jira.com/wiki/display/Heritrix/Heritrix
Hive		Hive defines a simple SQL-like query language that enables users familiar with SQL to query data stored in Apache Hadoop's HDFS (see "HDFS"). It generates MapReduce jobs created based on the HiveQL (see "HiveQL" statements). See https://cwiki.apache.org/confluence/display/Hive/Home
HiveQL		Query language used in the Apache Hadoop extension Apache Hive which is similar to MySQL's SQL syntax. See https://cwiki.apache.org/confluence/display/Hive/LanguageManual
Internet Archive		The Internet Archive is a digital library which provides permanent storage of and free public access to collections of digitized materials, including websites, music, moving images, and nearly three million public-domain books. See https://archive.org
ISIS		ISIS is a centre for research in the physical and life sciences at the STFC Rutherford Appleton Laboratory near Oxford in the United Kingdom. Our suite of neutron and muon instruments gives unique insights into the

		properties of materials on the atomic scale. See http://www.isis.stfc.ac.uk
JobTracker		In the context of an Apache Hadoop environment, the JobTracker is the service within Hadoop that farms out MapReduce tasks to specific nodes in the cluster, ideally the nodes that have the data, or at least are in the same rack. See “Apache Hadoop”. See https://wiki.apache.org/hadoop/JobTracker
Local Instance		See “SCAPE Local Instance“
MapReduce	MR	A programming paradigm for processing large data sets using a parallel, distributed algorithm on a SCAPE cluster.
MapReduce Toolwrapper		A tool to run toolrapper wrapped tools/Taverna workflows using Hadoop
MIME Type		A standard identifier used on the Internet to indicate the type of data that a file contains.
NameNode		In the context of an Apache Hadoop environment, the NameNode is the centerpiece of an HDFS file system. It keeps the directory tree of all files in the file system, and tracks where across the cluster the file data is kept. It does not store the data of these files itself. See “Apache Hadoop”. See https://wiki.apache.org/hadoop/NameNode
NeXus		The NeXus format is a common data format for neutron, x-ray and muon science. It has been developed as an international standard by scientists and programmers representing major scientific facilities in Europe, Asia, Australia, and North America in order to facilitate greater cooperation in the analysis and visualization of neutron, x-ray, and muon data.
Panasas Kit		

Pig		See “Apache Pig”.
PigLatin		Expression language used by Apache Pig, see “Apache Pig”.
PRONOM Unique Identifier	PUID	The *PRONOM* Persistent Unique Identifier (*PUID*) is an extensible scheme for providing persistent, unique and unambiguous identifiers for records in the *PRONOM* registry. Such identifiers are fundamental to the exchange and management of digital objects, by allowing human or automated user agents to unambiguously identify, and share that identification of, the representation information required to support access to an object. This is a virtue both of the inherent uniqueness of the identifier, and of its binding to a definitive description of the representation information in a registry such as *PRONOM*. From: http://www.nationalarchives.gov.uk/aboutapps/*PRONOM*/*PUID*.htm
PRONOM		*PRONOM* is an information system about data file formats and their supporting software products. See https://www.nationalarchives.gov.uk/*PRONOM*
PRONOM Signature File		Signature files are generated by *PRONOM* (see corresponding glossary entry) and used by *DROID* (see corresponding glossary entry) for file format identification. The signature file contains a subset of the information from the *PRONOM* knowledge base required by the *DROID* software to perform the file format identification. See https://www.nationalarchives.gov.uk/aboutapps/*PRONOM*/*DROID*-signature-files.htm
Preservation asset		A preservation asset in the SCAPE project refers to a software artefact (SCAPE component/application) which can be demonstrated by showing it‘s functionality and capabilities to interested parties.

Preservation Component	PC	See „SCAPE Component“.
Preservation User Story		See “User story”.
Platform		The term “Platform” refers either to the SCAPE Platform sub-project or the SCAPE Platform (see “SCAPE Platform”).
Preservation Planning and Watch		See “SCAPE Preservation Planning and Watch“.
Preservation Story		See “User story”.
PRONOM		PRONOM is an information system about data file formats and their supporting software products. See https://www.nationalarchives.gov.uk/PRONOM
PUID		See „*PRONOM* Unique Identifier“.
Quality Assurance Component		A Quality Assurance Component is used to determine a quality measure related to the outcome of applying an Action Service (see corresponding glossary entry) to a digital object.
raw		In the context of the Research Data Sets Testbed (TB.WP.3), this concept refers to data produced by the ISIS instruments at the Rutherford Laboratory (STFC), i.e. data produced by the instruments which are stored together with and associated metadata and log files in a non-standardised way. Apart from this specific meaning in the context of the Research Data Sets Testbed (TB.WP.3), the concept refers in other contexts also more generically to data where the structure is not prescribed by a file format specification.
Scalable Preservation Environments	SCAPE	An EU funded project developing scalable services for the planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale,

		heterogeneous collections of complex digital objects.
SCAPE Application		Similar to a *SCAPE Component* (see "*SCAPE Component*"), the SCAPE Application is also a software artefact which has been developed in the SCAPE Project. The difference compared to the *SCAPE Component* is that the SCAPE Application is a complex software construct with layered architecture, various interfaces and APIs and usually it has its own web interface.
SCAPE Characterisation Component		Characterisation components are a family of *SCAPE Components* (defined to wrap tools produced in WP9) that compute one or more properties of a *single* instantiated digital object or file. The output ports that produce measures are always annotated with the metric (in the *SCAPE Ontology*) that describes what the component computes.
SCAPE Component		SCAPE components are *Taverna Components*, identified by the SCAPE *Preservation Component*s sub-project, that conform to the general SCAPE requirements for having annotation of their behaviour, inputs and outputs. SCAPE components may be stored in the *SCAPE Component Catalogue*.
SCAPE Local Instance		The local instance in the SCAPE project is an environment in a memory institution where the SCAPE Execution Platform together with SCAPE *Preservation Component*s is deployed.
SCAPE Migration Component		Migration components are a family of *SCAPE Components* (defined to wrap tools produced in WP10) that apply a transformation to an instantiated digital object or file to produce a new file. The input is annotated with a term (from the *SCAPE Ontology*) that says what sort of digital object/file is accepted, and the output is annotated with a term that says what sort of file is produced.
SCAPE Ontology		The SCAPE Ontology is an OWL ontology that formally defines the terms used by computing systems in SCAPE.
SCAPE Platform local		See "local instance".

instance		
SCAPE Preservation Planning and Watch		SCAPE sub-project SCAPE Preservation Planning and Watch. The SCAPE Planning and Watch software suite makes preservation planning and monitoring context-aware through a semantic representation of key organizational factors, and it collects and reasons on preservation-relevant information.
SCAPE QA Component		QA components are a family of *SCAPE Components* (defined to wrap tools produced in WP11) that compute a comparison between *two* instantiated digital objects or two files. They produce at least one output that has a measure of similarity between the inputs, and that output is annotated with the metric (in the *SCAPE Ontology*) that describes the nature of the similarity metric.
SCAPE Testbeds		See “Testbeds”.
SCAPE Tool		See “SCAPE Component“.
Scheduled visit		Possibility for third party to ask for a visit at one of the *Demonstrating Institutions* to see outcomes of the SCAPE project in the institutional context.
Taverna Components		Taverna components are *Taverna workflow* fragments that are stored independently of the workflows that they are used in, and that are semantically annotated with information about what the behaviour of the workflow fragment is. They are logically related to a programming language shared library, though the mechanisms involved differ. Taverna components are stored in a component repository. This can either be a local directory, or a remote service that supports the Taverna Component API (e.g., the *SCAPE Component Catalogue*). Only components that are stored in a publicly accessible service can be used by a *Taverna workflow* that has been sent to a system that was not originally used to create it.

TaskTracker		In the context of an Apache Hadoop environment, a TaskTracker is a node in the cluster that accepts tasks - Map, Reduce and Shuffle operations - from a JobTracker. See “Apache Hadoop”. See https://wiki.apache.org/hadoop/TaskTracker
Taverna		Taverna is an XML-based language for describing data-flow oriented workflows and an execution engine that interprets this workflow description language. Taverna is the basis for both, *Taverna Workbench* or *Taverna Server*.
Taverna Workbench		The Taverna Workbench is a desktop application for creating, editing and executing *Taverna workflows*.
Taverna Workflow		A Taverna workflow is a parallel data-processing program that can be executed by *Taverna Workbench* or *Taverna Server*. It is stored as an XML file, and has a graphical rendering.
Taverna Server		Server version of the Taverna workflow software suite which allows executing data-flow oriented workflows described in an XML-based workflow description language using a REST API. See http://www.taverna.org.uk/download/server/
Testbeds		The SCAPE project has four application areas for which scenarios are developed to derive requirements, solutions are developed, evaluated, and demonstrated in the areas of web content, large-scale digital repositories, research datasets, and data centres. These four application areas correspond to the four *Testbeds* (Web Content Testbed, Large-scale Digital Repositories Testbed, Research Data Sets Testbed, and the Data Centres Testbed).
Tika		
Toolspec		An XML file written to a standard API that contains details of how to execute a tool for a particular purpose; for example txt2pdf might define how to use a command line tool to convert text to pdf. Toolspecs can have different types such as migration or QA.

Toolwrapper		The toolwrapper is a Java tool developed in the SCAPE Project to simplify the execution of the following tasks: Tool description (through the toolspec); Tool invocation (simplified) through command-line wrapping; Artifacts generation (associated to a tool invocation, e.g., Taverna workflow); and Packaging of all the generated artifacts for easier distribution and installation
User story		A user story, in the context of the *SCAPE Testbeds*, is a description of a digital preservation issue related to a concrete data set that is used to derive requirements for tool development.
WARC		Web archive container format. See also “ARC”. See http://bibnum.bnf.fr/WARC/
Wayback Machine		In the context of this deliverable, the term Wayback Machine refers to a software used to render archived web pages, originally developed by the Internet Archive (see corresponding glossary entry).
Web Crawler		Software used to capture and store web pages used by Web Archiving institutions to build their archives.
Web ARChive file format	WARC	The WARC (Web ARChive) file format offers a convention for concatenating multiple resource records (data objects), each consisting of a set of simple text headers and an arbitrary data block into one long file. See http://bibnum.bnf.fr/WARC/.
Web Content Testbeds	WCT	The Web Content Testbed is one of the *Testbeds* of the SCAPE project. The *Testbeds* are represented by memory institutions holding large data sets that are used to test the applicability of tools, workflows, and solutions developed in the SCAPE project.
Web Snapshot		The image capture of a web page that is taken when a web page is rendered in a web browser.