SCAPE

Scalable Preservation Environments

Final version of action components

Authors

Hélder Silva (KEEP Solutions)

September 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

This report provides an overview of the suite of software components including existing tools modified to run on the SCAPE platform and new tools developed by the Action Services Components workpackage. Detailed information about how to install and use them is also provided, making it easy to deploy into the SCAPE platform.

Table of Contents

1 Introduction

When talking about Digital Preservation, i.e. the several processes that need to be put in action to ensure the continuous access to information and cultural heritage present in the digital form, we need to talk about concrete actions that must be performed to achieve this purpose. Things like technology update, medium refreshment, emulation or format migration [1]. The Action Services workpackage concerned itself with migration tools described in this report.
Several tools are available, for file format migration, but usually they present us with several challenges, for example the tools can be:

- difficult to install;
- difficult to use, with several options and documentation that is not clear as how to combine them;
- difficult to combine with other tools (e.g. when there isn't a tool to perform a migration from format A to B, and this can only be achieved via a two-step migration from A to X and then from X to B).

If this wasn't enough, as a great number of them are Open Source, there are no guarantees about the quality of the conversions made. This is why one also needs tools for assessing the quality of the conversions, usually by extracting properties/features of both original and migrated version to ensure that no information was lost, or by ensuring the compliance against standards that they should conform to.

While having tools that perform file format conversion (Migration), feature extraction (Characterisation) and quality assurance (Quality Assurance) is helpful, it's not enough when we consider large scale Digital Preservation and/or integrating this process with several software components. This brings us yet more challenges, for example:

- installing tools in a large scale environment;
- configuring them to operate in an efficient manner;
- discovering tools available for particular tasks, and integrating them with automated planning software (e.g. Plato).

The Digital Preservation Toolkit was created to address these challenges. It combines tools for file format migration, characterisation and quality assurance helping to solve the following challenges:

- **Easy install & deployment at scale through packaging:** providing both individual Debian packages for each SCAPE Component and Debian meta-packages [4] for installing groups of related components;
- **Easy use through normalised command-line invocation syntax:** each SCAPE Component is wrapped, using the toolspec (tool description) and the SCAPE Toolwrapper, which produces a bash wrapper;
- **Easy tool composition of SCAPE Components:** using the toolspec (tool description), componentspec (SCAPE component description) and the SCAPE Toolwrapper, a Taverna [2] workflow can be produced. This supports combining tools into components to perform a particular task, and provides special semantic annotations for easy discovery by query through SCAPE Component Catalogue API's, the myExperiment site [3].

1.1 Goals

The goals of this report are:

- List the tools used, adapted, and created for deployment on the SCAPE platform;
- Provide information as to how to use and install the tools, source code locations, and the like.

1.2 Components

This section lists the different software components, which are outputs of this deliverable.

1.2.1 SCAPE Toolwrapper

Under the motto "*Wrap preservation tools once, deploy them everywhere",* the SCAPE Toolwrapper is a Java tool that allows a user with knowledge of a preservation tool to describe it. This expert user provides its name, the URL of the website, any dependencies it relies on (for each operation system), and, most importantly of all, one or more operations, where an operation is a concrete action that the tool performs, for example, migration from format X to Z.

This tool description can be used to generate several outputs that can be shared with the others in the digital preservation community:

- **Bash wrapper:** a bash script that standardizes the invocation of preservation tools making them simpler to invoke. It also emulates command-line piping through STDIN and STDOUT if the tool doesn't natively support it;
- **SCAPE Component**: a Taverna workflow adhering to a Component Profile and used as a building block in a Preservation Action Plan. These workflows are stored on the myExperiment website allowing easy discovery and download. These workflows are semantically annotated with special tool information such as the supported input and output formats of file format migration tools or, for the characterisation tools, the characteristics the tool can extract, etc. enabling detailed search and discovery;
- **Debian package**: a Debian package which contains both the bash wrapper and the SCAPE Component.

The Toolwrapper also supports the upload of the generated workflow to the myExperiment website which in the SCAPE context it's called SCAPE Component Catalogue.

1.2.2 SCAPE Components

SCAPE components are Taverna workflows identified by the SCAPE Preservation Components sub-project, that conform to the SCAPE requirements for annotation of their behaviour, inputs and outputs. SCAPE components may be stored in and retrieved from the SCAPE Component Catalogue.

1.2.3 Digital Preservation Toolkit

The chosen supported OS for the SCAPE Platform is Ubuntu (a Debian derivate). The Digital Preservation Toolkit is a Debian metapackage [4] called digital-preservation-tools, which references three other metapackages for:

- Migration (digital-preservation-tools-migration),
- Characterisation (digital-preservation-tools-characterization), and
- Quality Assurance (digital-preservation-tools-quality-assurance).

Using these metapackages one may install all the tools (by installing digital-preservation-tools) or a specific set of related tools (by installing the one of the digital-preservation-tools-* packages).

Each of the metapackages (Migration, Characterisation and Quality Assurance) references tools wrapped with SCAPE Toolwrapper, in the form of Debian packages.

2 Software suite

This section provides more detailed information about the components released with this deliverable.

2.1 SCAPE Toolwrapper

Source code available at GitHub: https://github.com/openplanets/scape-toolwrapper

2.2 SCAPE Components

Migration Components:

- Audio Video Migration - set of digital preservation components that can be used to migrate audio and video between various formats: http://www.myexperiment.org/packs/595
- Document Migration - set of digital preservation components that can be used to migrate documents between various formats: http://www.myexperiment.org/packs/601
- Image Migration - set of digital preservation components that can be used to migrate images between various formats: http://www.myexperiment.org/packs/592
- Scientific Data Migration - set of digital preservation components that can be used to migrate scientific data between various formats: http://www.myexperiment.org/packs/598
- Webpage Migration – set of digital preservation components that can be used to migrate webpages and their archives between various formats: http://www.myexperiment.org/packs/604

Characterisation Components:

- Audio Video Characterisation - set of digital preservation components that can be used to characterise audio and video: http://www.myexperiment.org/packs/596
- Document Characterisation - set of digital preservation components that can be used to characterise documents: http://www.myexperiment.org/packs/602
- Image Characterisation - set of digital preservation components that can be used to characterise images: http://www.myexperiment.org/packs/593
- Scientific Data Characterisation - set of digital preservation components that can be used to characterise scientific data files: http://www.myexperiment.org/packs/599
- Webpage Characterisation - set of digital preservation components that can be used to characterise webpages and their archives: http://www.myexperiment.org/packs/605

Quality Assurance Components:

- Audio Video Quality Assurance - set of digital preservation components that can be used to compare audio or video: http://www.myexperiment.org/packs/597
- Document Quality Assurance - set of digital preservation components that can be used to compare documents: http://www.myexperiment.org/packs/603
- Image Quality Assurance - set of digital preservation components that can be used to compare images: http://www.myexperiment.org/packs/594
- Scientific Data Quality Assurance - set of digital preservation components that can be used to compare scientific data files: http://www.myexperiment.org/packs/600
- Webpage Quality Assurance - set of digital preservation components that can be used to compare webpages and their archives: http://www.myexperiment.org/packs/606

Source available at GitHub: https://github.com/openplanets/scape-toolspecs

2.3 Digital Preservation Toolkit

The source code is available at GitHub: https://github.com/openplanets/digital-preservation-toolkit, while the packages are held in the OPF Debian repository: http://ubapt.opf-labs.org.

Each wrapped tool has its own documentation, available at the tools website and usually as an accompanying man page [5].

3 Software deployment

This section describes how one can install and use the components released with this deliverable. Some technical knowledge about source compilation, software installation and configuration is recommended.

3.1 SCAPE Toolwrapper

These instructions describe how someone with an understanding of using ImageMagick[1] for migrating an image file (with text) to a text file could:

- wrap the command-line invocation (bash wrapper);
- generate a SCAPE Component (Taverna workflow);
- generate a package (Debian package); and
- upload the SCAPE Component to myExperiment site.

In order to achieve this one needs the right software tools installed (pre-requisites), a toolspec (see appendix 5.1), a componentspec (see appendix 5.2) and a Debian changelog (see appendix 5.3).

Pre-requisites

- Debian/Ubuntu operating system or derivative;
- Git;
- Java Development Kit (JDK) version >= 1.6;
- Build tools (Maven, build-essential, dh-make, devscripts, debhelper and lintian);

From a command line shell:

Clone and compile the code

1. Clone the repository
   git clone https://github.com/openplanets/scape-toolwrapper.git
2. Compile code with Maven
   cd scape-toolwrapper
   mvn package

Generate, for a given toolspec and componentspec (see examples in the Appendix), bash wrapper and SCAPE Component

- Execute
  ./toolwrapper-bash-generator/bin/generate.sh -t digital-preservation-migration-image-imagemagick-image2txt.xml -c digital-preservation-migration-image-imagemagick-image2txt.component -o outputDirectory

Generate, for a given toolspec and changelog (see examples in the Appendix), a Debian package

- Execute

[1] http://www.imagemagick.org

```
./toolwrapper-bash-debian-generator/bin/generate.sh -t digital-preservation-migration-image-imagemagick-image2txt.xml -ch digital-preservation-migration-image-imagemagick-image2txt.changelog -i outputDirectory -o outputDirectory -e email@domain.com
```

Upload a SCAPE Component to the myExperiment site (592 is the Image Migration family ID; see section 2.2 to find out the id for other Component families)

- Execute
  ```
  ./toolwrapper-component-uploader/bin/upload.sh -s digital-preservation-migration-image-imagemagick-image2txt.component –c outputDirectory/workflow/digital-preservation-migration-image-imagemagick-image2txt.t2flow -d “Converts any ImageMagick supported image format to Text” -i 592 -l Apache -u username -p password -t digital-preservation-migration-image-imagemagick-image2txt.xml
  ```

3.2 Digital Preservation Toolkit

In order to Install the digital preservation toolkit on a Debian/Ubuntu derivative, we recommend apt (Advanced Packaging Tool), which is the standard software management tool for these distributions.

Pre-requisites

- Debian/Ubuntu system.

How to add OPF Debian repository

- Add repository key
  ```
  wget -O - http://ubapt.opf-labs.org/scape-components.gpg.key | sudo apt-key add -
  ```
- Add repository information, by editing file /etc/apt/source.list and adding the following line to the end of the file
  ```
  deb http://ubapt.opf-labs.org precise main
  ```

Install Digital Preservation Toolkit

- Execute
  ```
  sudo apt-get install digital-preservation-tools
  ```

Install only tools related to Migration

- Execute
  ```
  sudo apt-get install digital-preservation-tools-migration
  ```

Install only tools related to Characterisation

- Execute
  ```
  sudo apt-get install digital-preservation-tools-characterisation
  ```

Install only tools related to Quality Assurance

- Execute
  ```
  sudo apt-get install digital-preservation-tools-quality-assurance
  ```

List all Digital Preservation Toolkit packages

- Execute
  ```
  sudo apt-cache search --names-only '^digital-preservation-'
  ```

Execute installed wrapped tool (for example for package *digital-preservation-migration-image-imagemagick-image2txt*). Providing no parameters cause the bash wrapper to show the usage message

- Execute *digital-preservation-migration-image-imagemagick-image2txt*

3.3 SCAPE Components

Pre-requisites:

- Taverna workbench;
- A SCAPE Component (see section 2.2);
- Debian package associated with the previous mentioned SCAPE Component (as the workflow executes a specific bash wrapper).

Execute SCAPE Component in Taverna workbench[2]

1. Open workflow;
2. Run workflow;
3. Provide input parameters, for both input ports and parameters (if any) and click run.

4 Report benchmarking

Large scale tests and benchmarking have been performed, using tools created and/or adapted in this workpackage, by the Testbeds (TB) workpackage on the corpora provided by the content holders. Because these results are reported in deliverable D18.2 and that deliverable is lengthy, interested readers should refer to that document. The results are left out here for both readability and to avoid repetition.

5 Appendix

5.1 Example of a toolspec

digital-preservation-migration-image-imagemagick-image2txt.xml – Toolspec[3] describing ImageMagick (homepage, software license, how one can install it on Ubuntu, etc.) and a concrete ImageMagick operation: migration of an image file to a text file (extraction would be more suitable but a broader term is used instead). The operation is called “digital-preservation-migration-image-imagemagick-image2txt” and has a description, a command-line (that performs the operation) and a set of inputs/outputs.

```
<?xml version="1.0" encoding="utf-8"?>
<tool xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://scape-
project.eu/tool" xmlns:xlink="http://www.w3.org/1999/xlink"
xsi:schemaLocation="http://scape-project.eu/tool toolwrapper-
data/src/main/resources/tool-1.1_draft.xsd" schemaVersion="1.1" name="ImageMagick"
version="2.0.0" homepage="http://www.imagemagick.org/script/convert.php">
```

[2] Detailed information about using Taverna workbench can be found in D7.2

[3] https://github.com/openplanets/scape-toolwrapper#tool-specification-toolspec

```
	<license name="Apache-2.0" type="FLOSS"
uri="http://opensource.org/licenses/Apache-2.0"/>
	<installation>
		<operatingSystem operatingSystemName="Debian">
			<packageManager type="Dpkg">
				<config>imagemagick</config>
				<source>deb http://scape.keep.pt/apt stable main</source>
			</packageManager>
			<dependency name="imagemagick">
				<license name="Apache-2.0" type="FLOSS"
uri="http://opensource.org/licenses/Apache-2.0"/>
			</dependency>
		</operatingSystem>
		<operatingSystem operatingSystemName="Ubuntu">
			<packageManager type="Dpkg">
				<config>imagemagick</config>
				<source>deb http://scape.keep.pt/apt stable main</source>
			</packageManager>
			<dependency name="imagemagick">
				<license name="Apache-2.0" type="FLOSS"
uri="http://opensource.org/licenses/Apache-2.0"/>
			</dependency>
		</operatingSystem>
	</installation>
	<operations>
		<operation name="digital-preservation-migration-image-imagemagick-
image2txt">
			<description>Converts any ImageMagick supported image format to
Text</description>
			<command>/usr/bin/convert ${input} txt:${output}</command>
			<inputs>
				<input name="input" required="true">
					<description>Reference to input file</description>
				</input>
				<parameter name="params" required="false">
					<description>Additional conversion
parameters</description>
				</parameter>
			</inputs>
			<outputs>
				<output name="output" required="true">
					<description>Reference to output
file</description>
					<extension>txt</extension>
				</output>
			</outputs>
		</operation>
	</operations>
</tool>
```

5.2 Example of a componentspec

digital-preservation-migration-image-imagemagick-image2txt.component – Componentspec[4] describing which input and output formats ImageMagick and in particular this operation (digital-preservation-migration-image-imagemagick-image2txt) allows. This information is necessary if one wants to generate a SCAPE Component as the information hereby available is used to enrich the Taverna workflow.

[4] https://github.com/openplanets/scape-toolwrapper#component-specification-componentspec

```
<?xml version="1.0" encoding="UTF-8"?>
<components xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://scape-project.eu/component"
xmlns:xlink="http://www.w3.org/1999/xlink" xsi:schemaLocation="http://scape-
project.eu/component toolwrapper-data/src/main/resources/component-1.1_draft.xsd"
schemaVersion="1.1">
       <component xsi:type="MigrationAction"
profile="http://purl.org/DP/components#MigrationAction" profileVersion="1.0"
name="digital-preservation-migration-image-imagemagick-image2txt" author="Hélder
Silva">
             <license name="Apache-2.0" type="FLOSS"
uri="http://opensource.org/licenses/Apache-2.0"/>
             <migrationPath>
                    <fromMimetype>image/bmp</fromMimetype>
                    <toMimetype>text/plain</toMimetype>
             </migrationPath>
             <migrationPath>
                    <fromMimetype>image/gif</fromMimetype>
                    <toMimetype>text/plain</toMimetype>
             </migrationPath>
             <migrationPath>
                    <fromMimetype>image/vnd.microsoft.icon</fromMimetype>
                    <toMimetype>text/plain</toMimetype>
             </migrationPath>
             <migrationPath>
                    <fromMimetype>image/jpeg</fromMimetype>
                    <toMimetype>text/plain</toMimetype>
             </migrationPath>
             <migrationPath>
                    <fromMimetype>image/png</fromMimetype>
                    <toMimetype>text/plain</toMimetype>
             </migrationPath>
             <migrationPath>
                    <fromMimetype>image/tiff</fromMimetype>
                    <toMimetype>text/plain</toMimetype>
             </migrationPath>
             <migrationPath>
                    <fromMimetype>image/jp2</fromMimetype>
                    <toMimetype>text/plain</toMimetype>
             </migrationPath>
       </component>
</components>
```

5.3 Example of a changelog

digital-preservation-migration-image-imagemagick-image2txt.changelog – Changelog[5] used for generating the Debian package, which keeps track of the changes performed in the toolspec and/or componentspec. This has direct implications into the generated Debian package (e.g. the version of the package is obtained from the changelog information).

```
digital-preservation-migration-image-imagemagick-image2txt (2.0.0) unstable;
urgency=low

  * Added SCAPE Component annotations to Taverna Workflow.

 -- SCAPE project <hsilva@keep.pt>  Thu, 05 Dec 2013 12:02:10 +0000
```

[5] https://www.debian.org/doc/debian-policy/ch-source.html#s-dpkgchangelog

```
digital-preservation-migration-image-imagemagick-image2txt (1.0.2) unstable;
urgency=low

  * Initial Release.

 -- SCAPE project <hsilva@keep.pt>  Thu, 01 Jan 2013 12:02:10 +0000
```

6 References

1. Format migration in digital preservation is described in Section 5 of “Reference Model for and Open Archival Information System”: http://public.ccsds.org/publications/archive/650x0m2.pdf
2. Taverna Workflow Software: http://www.taverna.org.uk
3. MyExperiment Taverna workflow hosting: http://www.myexperiment.org
4. Ubuntu documentation on metapackages: https://help.ubuntu.com/community/MetaPackages
5. Linux man pages: http://en.wikipedia.org/wiki/Man_page