Abstract

The PC subproject develops and packages a variety of tools for preservation actions, characterisation and quality assurance. Additionally it packages existing tools to ensure that all services are compliant with and can be deployed in the SCAPE platform. Existing tools usually come in the form of command line programs that provide a variety of output formats. To guarantee interoperability between services and enable automation of preservation processes, PC creates preservation components that provided a standardised interface. These preservation components are based on Taverna workflows as a common language. They either wrap command line tools with necessary post-processing, are directly implemented as workflows or are composed of other components.

This document provides a high-level specification of the Profiles to which the Preservation Components must comply.

Keyword list

Preservation Components, Profiles, Actions, Characterisation, Quality Assurance, Tools, Planning, Taverna
### Authors

<table>
<thead>
<tr>
<th>Person</th>
<th>Role</th>
<th>Partner</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miguel Ferreira</td>
<td>Author</td>
<td>KEEPS</td>
<td>Main contributor</td>
</tr>
<tr>
<td>Carl Wilson</td>
<td>Author</td>
<td>OPF</td>
<td>Contributor</td>
</tr>
<tr>
<td>Donal Fellows</td>
<td>Author</td>
<td>UNIMAN</td>
<td>Contributor</td>
</tr>
<tr>
<td>Hélder Silva</td>
<td>Author</td>
<td>KEEPS</td>
<td>Contributor</td>
</tr>
<tr>
<td>Zeynep Pehlivan</td>
<td>Comments</td>
<td>UPMC</td>
<td>Reviewer</td>
</tr>
<tr>
<td>Per Mølrup-Dalum</td>
<td>Comments</td>
<td>SB</td>
<td>Reviewer</td>
</tr>
</tbody>
</table>

### Document Approval

<table>
<thead>
<tr>
<th>Person</th>
<th>Role</th>
<th>Partner</th>
</tr>
</thead>
</table>

### Distribution

<table>
<thead>
<tr>
<th>Person</th>
<th>Role</th>
<th>Partner</th>
</tr>
</thead>
</table>

### Revision History

<table>
<thead>
<tr>
<th>Version</th>
<th>Status</th>
<th>Author</th>
<th>Date</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>Draft</td>
<td>Miguel Ferreira</td>
<td>2013-06-20</td>
<td>Original document</td>
</tr>
</tbody>
</table>

### References

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Document</th>
<th>Date</th>
<th>Details and Version</th>
</tr>
</thead>
</table>
Table of Contents

Whitepaper Preservation Component Profiles i

1 Introduction 1

2 Data types 2
   2.1 Representation 2
   2.2 Format 3
   2.3 Custom parameters 3
   2.4 Custom Report 3
   2.5 Similarity 3
   2.6 Valid 3

3 Media types 4

4 Preservation Component Profiles 5
   4.1 Migration Component 5
   4.2 Characterisation Component 5
      4.2.1 Identification Component 6
      4.2.2 Format Validation Component 6
      4.2.3 Profile Assessment Component 6
      4.2.4 Image Feature Extraction Component 7
   4.3 Quality Assurance Component 7
1 Introduction

The PC subproject develops and packages a variety of tools for preservation actions, characterisation and quality assurance. Additionally it packages existing tools to ensure that all services are compliant with and can be deployed in the SCAPE platform. Existing tools usually come in the form of command line programs that provide a variety of output formats. To guarantee interoperability between services and enable automation of preservation processes, PC creates preservation components that provided a standardised interface. These preservation components are based on Taverna workflows as a common language. They either wrap command line tools with necessary post-processing, are directly implemented as workflows or are composed of other components.

To allow discoverability (e.g. for preservation planning, watch or other users), the components must be annotated with relevant information using a common vocabulary. Standardised interfaces add the possibility of automating parts of a preservation process. Additionally, dependencies allow the execution platform to run the components. All these aspects are defined as preservation component profiles.

Using Taverna workflows allows domain experts from the testbed as well as the wider community to engage in the development of preservation components. The Taverna workbench will implement a mechanism to easily validate workflows against profiles.

To increase the level of automation and thus scalability in preservation planning, the planning component queries the component repository (myexperiment.org) for migration components based on the migration paths and other criteria specified in the policy of the organisation (e.g. by license).

The preservation planner selects fitting alternatives and all representative samples are migrated to the respective target format. Based on the objective tree for this plan, a template of a QA workflow is generated that contains input ports for the path to the files as well as output ports for measurable properties for this file.

A preservation planner can then open this template in Taverna and add characterisation and quality assurance components from myexperiment.org that provided the needed properties. The QA components are then used to evaluate the possible alternatives. The preservation planner can select the best alternative and an executable plan is generated by combining the migration component with quality assurance.

This document provides a high-level specification of the Profiles to which the Preservation Components must comply. The following profiles have been agreed by all the WP leads in the Preservation Components Subproject, the TCC (on the person of Carl Wilson) and the main Taverna developer (Donal Fellows).
2 Data types

This section describes the basic data types that components should be able to support.

2.1 Representation

A representation is a set of files that together embody a digital Intellectual Entity. A Representation is the set of Files needed to provide a complete and reasonable rendition of an Intellectual Entity.

A Representation is one of the most relevant inputs to Preservation Components. For example, a Migration Component will expect a Representation as input, and will output a new Representation.

In a Preservation Component a Representation is defined as a list of URIs. These URIs SHOULD NOT be used to reference a set of independent intellectual entities, even if these share the same format. Instead, all the URIs in a Representation must be processed in one single step by the Preservation Component. If one wants to parallelise processing of the URIs within a Representation, then these should be broken apart and be fed to a Preservation Component as individual Intellectual Entities.

Preservation Components that receive a Representation as input should be able to cope with the whole complexity of the compound object.

2.1.1 Example: Single Tiff image

A single TIFF image should be included in a Representation as a list with a single URI.

2.1.2 Example: A research dataset

A research dataset in the NeXus format should include all the files that compose that format in the list of URIs, i.e. the instrument parameters, actual data, instrument metadata, general metadata, etc.

2.1.3 Example: Website

A Website is usually composed of several files. All of these files, although in different formats, are part of the same intellectual entity. Therefore, all of these files should be included in the list of URIs.

2.1.4 Example: A series of thesis in PDF format

A series of PDF files representing various thesis should not be bound together on the same list of URIs as they are not part of the same intellectual entity.

Instead, create various lists, each of which with a single PDF inside.
2.2 Format

Formats should be described by MIME types. Good practice is to include format name and version and use MIME type only if no other data is available, e.g. “application/pdf; version 1.4”.

2.3 Custom parameters

A String that is passed along to the tool or algorithm running within the Preservation Component.

2.4 Custom Report

A String that the Component may choose to output. It could be the standard output, the standard error, it could be XML. Its therefore Component specific and allows the user of the Component to know more about the results of the component (e.g. understand why a Representation is not similar to another on a QA process).

2.5 Similarity

The output of a QA component. Its values are numeric and range from 0 to 1.

2.6 Valid

It’s an output of a Validation process. The type is Boolean: True if the file is valid according the rules being tested or False otherwise.
3 Media types

The following media types are currently supported in the project and will have Preservation Components:

1. Audio
2. Video
3. Image
4. Documents
5. Research Datasets
4 Preservation Component Profiles

This section presents a high-level view of the interfaces of Preservation Components. The PC subproject will produce tools that comply with the following profiles and sub-profiles:

- **Migration Component**
- **Characterisation Component**
  - Identification Component
  - Format Validation Component
  - Profile Assessment Component
  - Feature Extraction Component
    - Image Feature Extraction Component
    - Video Feature Extraction Component
    - Audio Feature Extraction Component
    - Document Feature Extraction Component
    - Web Feature Extraction Component
    - Research Dataset Feature Extraction Component
- **Quality Assurance Component**

4.1 Migration Component

4.2 Characterisation Component

In order to avoid any confusion regarding terminology, this report uses the same terminology as the Evaluation of Characterisation Tools report (Wilson, 2011) that states the following:

*In practice the term “characterisation” is often used to indicate quite different things. To avoid any confusion, this document follows the terminology used in the JHOVE2 project. Here, characterisation is loosely defined as the process of deriving information about a digital object that describes its character or significant nature. This process is subdivided into four aspects:*

- **Identification** - the process of determining the presumptive format of a digital object;
- **Feature extraction** - the process of reporting the intrinsic properties of a digital object that are significant to preservation planning and action;
• **Format validation** - the process of determining the level of conformance of a digital object to the normative syntactic and semantic rules defined by the authoritative specification of the object’s format;

• **Profile assessment** - the process of determining the level of acceptability of a digital object for a specific purpose based on locally defined policy rules.

These four aspects of a Characterisation component are distinct in nature, so different interfaces must be created to cope with their own particular complexity. Furthermore, the interfaces of the Feature Extraction components vary depending on the media type.

Due to these reasons, the definition of interfaces for the characterisation components depend on the type of characterisation and the type of media (in what regards Feature extraction)

### 4.2.1 Identification Component

![Identification Component Diagram]

### 4.2.2 Format Validation Component

![Format Validation Component Diagram]

### 4.2.3 Profile Assessment Component

![Profile Assessment Component Diagram]
4.2.4 Image Feature Extraction Component

This component interface is just an example. There should be a profile per media type where the expected outputs are clearly identified. The strategy to pursue is to start with a small manageable number of features that the existing characterisation tools are able to extract, and then expand those features as the project progresses.

4.3 Quality Assurance Component