GUIMARÃES 2012
CAPITAL EUROPEIA DA CULTURA

SCAPE

# SCAPE Training event
# Keeping Control – Scalable Preservation Environments for Identification and Characterisation

**6-7 December 2012**
**Guimarães, Portugal**

## Learning outcomes

1. Distinguish between different file types and identify the requirements for characterising each of them.
2. Carry out a number of identification, characterisation, and duplication detection experiments on example files.
3. Critically evaluate characterisation and identification tools and assess their advantages and disadvantages when used in different scenarios.
4. Compare and contrast the differences in running characterisation and identification tools both stand-alone and within workflows.
5. Envisage a system that combines workflows with identification, characterisation and validation tools to suit a variety of scenarios.
6. Conduct an in-depth analysis of large volumes of identification and characterisation data and find representative sample records suitable for preservation planning experiments.

## AGENDA (draft)

**Thursday 6 December**

| Time | Session | Facilitators | Learning outcomes |
|---|---|---|---|
| 09.30 - 10.00 | Registration | | |
| 10.00 - 10.15 | Welcome and housekeeping | Miguel Ferreira, KEEPS | |
| 10.15 - 11.15 | **Introduction to file formats**<br>Understanding the different requirements for identification and characterisation experiments<br><br>**File format identification and characterisation tools: file, droid, tika, exiftool**<br>What can they do?<br>File Format Identification, File Format Characterisation,<br>File Format Validation, File Format Signature Files | Carl Wilson, OPF<br>Dave Tarrant, OPF | 1. |
| 11.15 - 11.30 | Coffee | | |
| 11.30 - 12.45 | **Applying file format tools to different scenarios** (demonstrations)<br>How do they compare? | Carl Wilson, OPF<br>Dave Tarrant, OPF | 1. |
| 12.45 - 13.45 | Lunch | | |
| 13.45 - 15.15 | **Break out groups: practical exercises**<br>Creating file format profiles with an example dataset<br>Command line processing<br><br>Evaluation of the results | Carl Wilson, OPF<br>Dave Tarrant, OPF | 2. |

| | | | |
|---|---|---|---|
| 15:15 - 15.30 | Coffee | | |
| 15.30 - 16:30 | **Wrapping tools for identification and characterisation**<br>FITS (File Information Tool Set)<br><br>**Panel session: advantages and disadvantages of wrapping tools**<br>Q&A | Petar Petrov, TUWIEN<br><br>All | 3. |
| 16.30 - 17.00 | Wrap up | Dave Tarrant, OPF | |
| 17.00 | Close | | |
| 20.00 | Event dinner | | |

## Friday 7 December

| Time | Session | Facilitators | Learning outcomes |
|---|---|---|---|
| 09.00 - 09.10 | Welcome back, overview of agenda for the day | Dave Tarrant, OPF | |
| 09.10 - 10.15 | **Using file format identification tools as part of a workflow**<br>Introduction to Taverna workflows<br>Demonstration: Web archive content identification over ARC files using tika in a Taverna workflow | Sven Schlarb, ONB | 4. |
| 10.15 - 10.30 | Coffee | | |
| 10.30 - 11.45 | **Comparing the Taverna workflow with a DROID version of the workflow**<br>Introduction to file format identification using a Hadoop cluster (demonstration)<br>Understanding the implementation differences | Sven Schlarb, ONB | 4. |
| 11.45 - 12.15 | Comparison of results | Sven Schlarb, ONB | |
| 12.15 - 13.15 | Lunch | | |
| 13.15 - 13.45 | **Content profiling and planning**<br>Introduction and motivation of large-scale content profiling for preservation analysis | Petar Petrov, TUWIEN | 5. |
| 13.45 - 14.15 | **Practical exercise:** analysing an example scenario file set without a content profiler<br>Discussion of results | Petar Petrov, TUWIEN | 6. |
| 14.15 - 14.45 | **c3po** (A content profiling prototype) demonstration of the tool and its capabilities | | |
| 14.45 - 15.30 | **Practical exercise:** analysing the scenario file set using c3po<br>Comparing the results and lessons learned | Petar Petrov, TUWIEN | 6. |
| 15.30 - 15.45 | Coffee | | |
| 15.45 - 16.30 | **Quality control for digital collections: the matchbox tool**<br>Identifying duplicate images in digital collections | Roman Graf, AIT | 4. |
| 16.30 - 17.00 | Wrap up discussion and event evaluation | Dave Tarrant, OPF | |
| 17.00 | Close | | |