Planets, SCAPE, and Beyond

My personal digital preservation journey

Dr. Ross King
AIT Austrian Institute of Technology GmbH

DL2014 Workshop:
“Digital Preservation Sustainability on the EU Policy Level”
London, September 8, 2014
SCAPE Project Data

- Project instrument: FP7 Collaborative Project
- 20 Partners from 11 countries
- 6. Call
  - Objective ICT-2009.4.1: Digital Libraries and Digital Preservation
  - Target outcome (a) Scalable systems and services for preserving digital content
- 10. Call
  - Objective ICT-2013.11.4: Supplements to Strengthen Cooperation in ICT R&D in an Enlarged European Union
- Duration: 44 months
  - February 2011 – September 2014
- Budget: 12.0 Million Euro
  - Funded: 9.2 Million Euro
SCAPE Consortium
SCAPE – what is it about?

• Planning and executing computing-intensive digital preservation processes such as the large-scale ingestion, characterisation or migration of large (multi-Terabyte) and complex data sets

• SCAPE results include
  • Preservation scenarios
  • Preservation tools
  • Preservation workflows
  • Preservation infrastructures
  • Preservation best-practices

SCAPE is a follow-up to the highly successful FP6 IP Planets.
The Planets Suite

Tools and Services

- Characterisation Tools
- Preservation Action Tools
- Emulation Tools
- Preservation Services

Registries

- Characterisation Registry
- Preservation Action Registry
- Core Registry

Applications

- Preservation Planning Application (PLATO)
- Testbed Application
- Service Registry
- Orchestration Tools
- Experiments Database
- Data Repository

Infrastructure

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).
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).
# Planets and SCAPE: A comparison

<table>
<thead>
<tr>
<th>Component</th>
<th>Planets</th>
<th>SCAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tools</strong></td>
<td>Characterisation (DROID) Action (migration)</td>
<td>Characterisation (TIKA/Nanite, FITS)</td>
</tr>
<tr>
<td></td>
<td>Validation (JHove) Emulation</td>
<td>Action (migration) Validation (Jpylyzer)</td>
</tr>
<tr>
<td><strong>Services</strong></td>
<td>Web Services</td>
<td>Web Services</td>
</tr>
<tr>
<td></td>
<td></td>
<td>APTs with Hadoop</td>
</tr>
<tr>
<td><strong>Registries</strong></td>
<td>Service Registry (custom)</td>
<td>Component Registry (myExperiment)</td>
</tr>
<tr>
<td></td>
<td>Format Registry (Tessella)</td>
<td></td>
</tr>
<tr>
<td><strong>Preservation Planning</strong></td>
<td>PLATO</td>
<td>PLATO 4.4</td>
</tr>
<tr>
<td><strong>Workflow</strong></td>
<td>Custom engine JBoss BPEL</td>
<td>Taverna</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Apache Pig, ToMaR</td>
</tr>
<tr>
<td><strong>Data Management</strong></td>
<td>JCR 170 repository (custom)</td>
<td>Data Connector API (RODA, Rosetta, Fedora4)</td>
</tr>
<tr>
<td><strong>Other stuff</strong></td>
<td>Testbed Application</td>
<td>Automated Watch (SCOUT, C3PO)</td>
</tr>
</tbody>
</table>

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).
Solving Preservation Problems the SCAPE Way

- Open Source Development
  - And/or implementation of open APIs

- Uniform Deployment
  - Use the SCAPE Toolspec+Toolwrapper to publish tools
    - As Advanced Packaging Toolkit (APT) packages
    - As SCAPE Components

- Preservation Planning
  - Use PLATO to test tools (as SCAPE Components) and make policy-based plans

- Process Modelling
  - Use Taverna to model preservation workflows
    - Taverna works directly with SCAPE components for experimental workflows
    - Taverna workflows can be converted to Hadoop/Pig workflows in some cases

- Hadoop Deployment
  - Use APT packages to deploy to a Hadoop environment

- Scalable Execution
  - SCAPE ToMaR can directly access tools through the toolspec

---

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).
Lessons Learned
Lessons Learned

- Lesson 1: Research projects often face structural impediments
The Wall

Production

• Practitioners

• Focus on daily business needs

• Service availability is a priority
  • Services are stable
  • Enjoy a large maintenance pool

Research and Development

• Developers

• Focus on innovation

• Services are prototypes
  • Unstable
  • Buggy
  • Maintenance pool limited to a few (or one) expert(s)

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).
Production

Research and Development

Digital Objects

SCAPE Platform

- ToMaR
- Pig
- Hadoop
- HDFS
- 1
- 2
- \( n \)

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).
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).
The Goal

Research and Development in Production

Digital Objects

Data Connector API

Repository

SCAPE Platform

ToMaR  Pig

Hadoop

HDFS

1  2  ...  n

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).
Lessons Learned

• Lesson 2: The time horizon for preservation problems is an order of magnitude too far for most decision makers.
Lessons Learned

- Lesson 3: Format Migration may not be as important as we thought it was
Lessons Learned

The DCC Curation Lifecycle Model

H2020

FP6, FP7 Projects

Most of the world?
H2020 Thoughts

- Research versus Production
- Storage versus Computation
- Access and Re-use versus Preservation

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).
SCAPE Additional Information
Additional Resources of Interest

• Development Infrastructure
  • Code repository hosted by the Open Planets Foundation and GitHub
    • https://github.com/openplanets/scape/
  • Development Wiki
    • http://wiki.opf-labs.org/display/SP/Home

• Experimental Workflows
  • http://www.myexperiment.org/search?query=SCAPE&type=all&commit=Search

• Publications
  • http://www.scape-project.eu/category/publication

• Public Deliverables
  • http://www.scape-project.eu/category/deliverable

• Tools
  • http://www.scape-project.eu/tools
SCAPE Contact Information

- http://www.scape-project.eu/
- Twitter: #scapeproject
- office@list.scape-project.eu

- Dr. Ross King
  AIT Austrian Institute of Technology GmbH
  Donau-City-Strasse 1
  A-1220 Wien
Thank you for your attention!