SCAPE
Scalable Preservation Environments

Peter May
Slides courtesy of Per Mølrup-Dalum (State and University Library, Denmark)

SCAPE Information Day
• What is SCAPE?
• What can it offer?
• Who are involved in SCAPE?
Digital Preservation – What do I need?

- Your collection of digital data is **growing** rapidly.

- Your preservation activities must become more **efficient** and more **scalable**.

- The SCAPE project has developed scalable solutions for long-term preservation of large-scale and heterogeneous data sets.

*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).*
Scalability — 4 dimensions

- Heterogeneity of collections
- Size of objects
- Number of objects
- Complexity of objects

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).
What is SCAPE?

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).
Solutions Tested in Real Life

Web Content

Digital Repositories

Data Centres

Research Data Sets

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).
Scalability
In four dimensions: Heterogeneity of collections as well as number, size and complexity of objects

Automation
Through scalable, automated and simple to design preservation workflows

Planning
Answering core preservation planning questions

Integration
Through a robust, integrated, open source preservation system

Centralised services
Allowing content holders to share information globalising the basis for decision-making

Solutions for Content Holders

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).
Overview: SCAPE Platform

The SCAPE Platform is a reference architecture for scalable preservation environments.

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 SCAPE Preservation Components are tools which enhance the functionality of a digital preservation system in:

- Scalability
- Functional coverage
- Quality
Overview: Planning & Watch Services

The SCAPE Planning and Watch components address the bottleneck of decision processes and processing information required for decision making.
Examples of tools and services
• Desktop application
• GUI
• Plug-in Framework
• Themed editions
• Intermediate results views
• Search for Web Services in catalogues
• Search and publish to myExperiment
• Taverna Server
FLint – Validation of files against institution policies

- Detect presence of DRM or encryption in PDF and EPUB files
- Identify files that fall outside of your own institutional policy, for example PDFs with missing embedded fonts
- Aggregates information from several sources, for example: Apache Tika, Apache PDFBox, IDPF EpubCheck, Jhove, Calibre and iText
Preservation Components

Jpylyzer – Easy Validation of JPEG 2000

• Automated JP2 validation and feature extraction
• Enables you to confirm whether an image is a valid, intact JP2 file
• Reports the key technical properties of the image

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).
Preservation Components

C3PO – Clever, Crafty, Content Profiling of Objects

- Analyses characterisation metadata for digital collections
  - Aggregates and combines metadata information across collections
- Gives you in depth knowledge of your files
  - Generates a profile of the content set
- Integrates with Scout (Watch) and Plato (Planning) services

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).
Nanite – Scalable format identification and characterisation system

- Combines core identification code in a way to make it easier to run identification and characterisation at scale
- Built around DROID, Apache Tika™ and Unix file
- Designed with large scale characterisation of WARC files in mind
  - Over Hadoop
- Can output C3PO compatible results
Preservation Components

Matchbox – Duplicate Image Detection Tool

- Identifies duplicate images even where files are different
  - Format, size, rotation, cropping, colour-enhancement etc.
- Works where OCR will not
  - Handwriting, music scores
- Useful for automating QA processes, reducing manual effort and error rates

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 tools are published as open source software.

Tools and services from SCAPE are sustained by

- Open Planets Foundation - address core digital preservation challenges and engage with the community

- COPTR - Community Owned digital Preservation Tool Registry
Ultimate Sustainability goal:

- Supporting communities of practice by enabling efficient collaboration during the project and beyond.

Open Planets Foundation will take post-project ownership of the outputs, supported by other partners providing specific capabilities.

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).
About SCAPE

- EU-funded project under FP7 (Research and Technological Development)
- Project runtime: February 2011 to September 2014
- 20 partners from 10 countries - from memory institutions, data centers, research labs, universities, and industrial firms
- Public Project materials are licensed under a CC-BY-SA International License

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).
Additional Sources 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
- Tools
  - http://www.scape-project.eu/tools
- 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
More Information

- SCAPE website: [www.scape-project.eu](http://www.scape-project.eu)
- Tools and Services: [https://github.com/openplanets/scape](https://github.com/openplanets/scape)
- SCAPE Twitter: @SCAPEProject, #SCAPEProject
- SCAPE Newsletter: Sign up via [www.scape-project.eu](http://www.scape-project.eu)

All images © the SCAPE Project or its partners, except images on slides 4, 6 and 24 © [www.digitalbevaring.dk](http://www.digitalbevaring.dk)
Thank you!