SCAPE

Scalable Preservation Environments

Second EUDAT Conference, October 2013
Workshop: Digital Preservation of Cultural Data

Scalability in preservation of cultural heritage data

Simon Lambert

Scientific Computing Department
STFC
UK

Types of scalability

Thinking about a repository/archive …

- Total number of digital objects
- Individual size of digital objects
- (Rate of ingest of digital objects)
- Complexity of digital objects
- Heterogeneity of collections

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).

Aspects of cultural data

Who are the designated communities?

How is ingest done?

What types of data? What variety? How packaged?

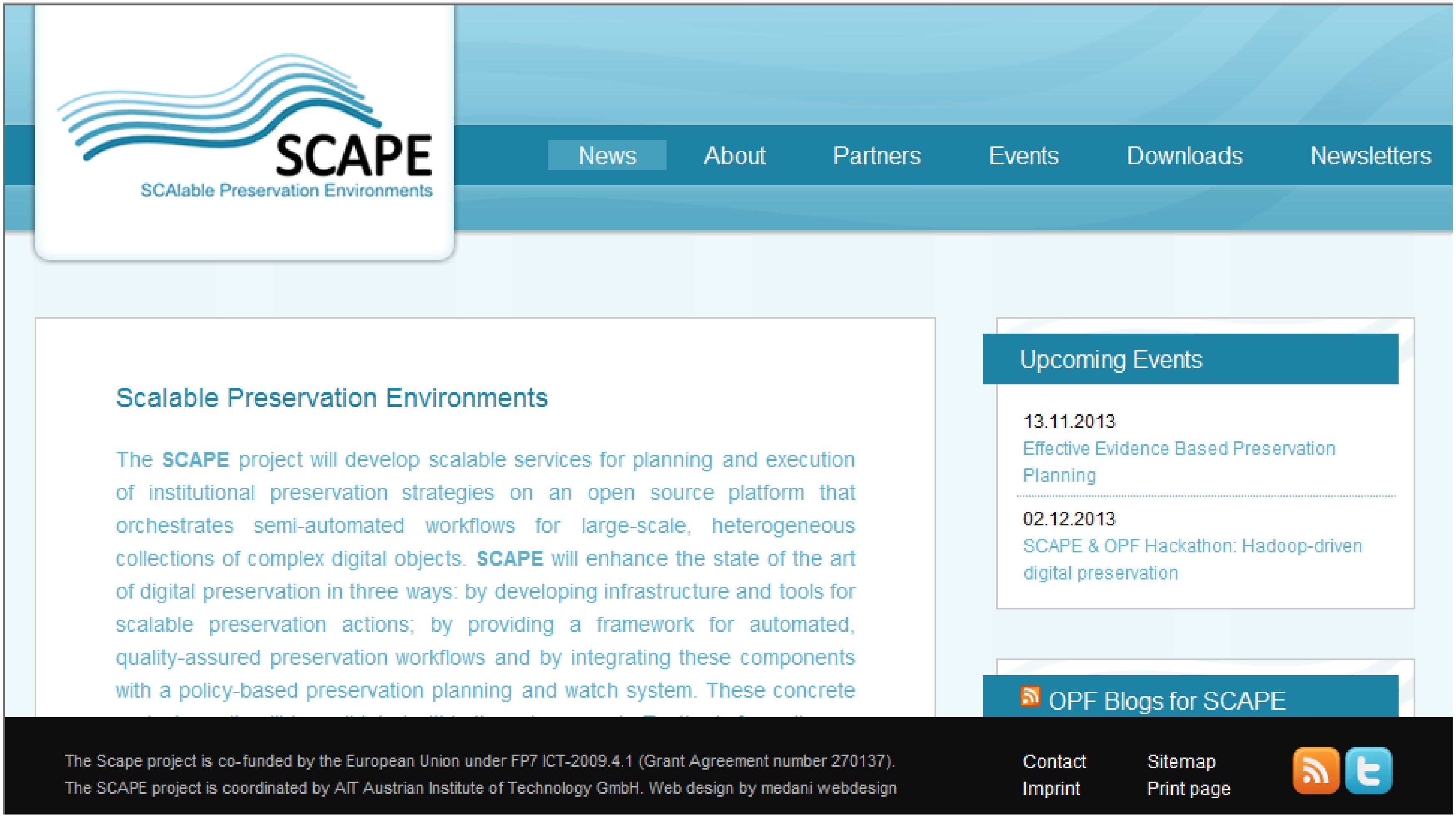
Validation of preservation actions

Importance of provenance

Access restrictions and DRM

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 project



Scalable Preservation Environments

The **SCAPE** project will develop scalable services for planning and execution of institutional preservation strategies on an open source platform that orchestrates semi-automated workflows for large-scale, heterogeneous collections of complex digital objects. **SCAPE** will enhance the state of the art of digital preservation in three ways: by developing infrastructure and tools for scalable preservation actions; by providing a framework for automated, quality-assured preservation workflows and by integrating these components with a policy-based preservation planning and watch system. These concrete

Upcoming Events

13.11.2013
Effective Evidence Based Preservation Planning

02.12.2013
SCAPE & OPF Hackathon: Hadoop-driven digital preservation

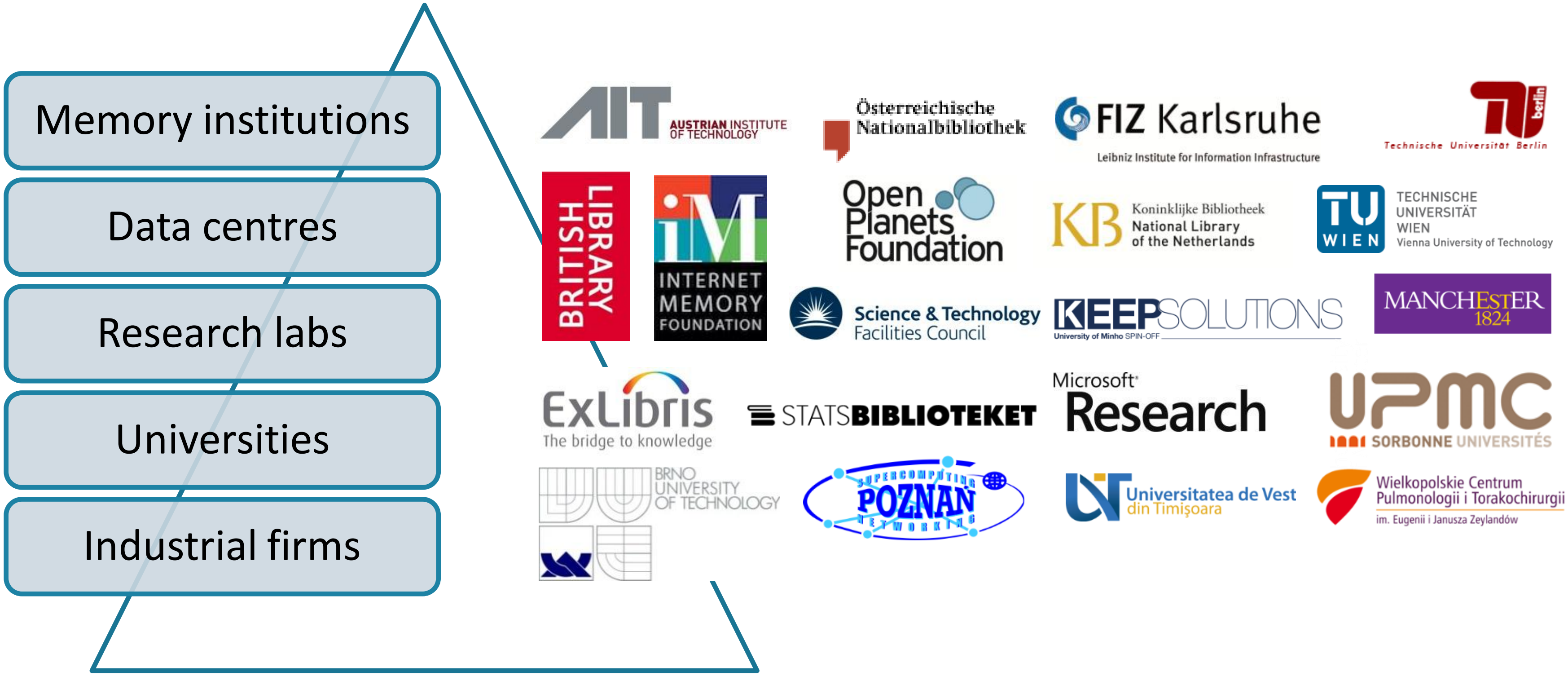
OPF Blogs for SCAPE

The Scape project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
The SCAPE project is coordinated by AIT Austrian Institute of Technology GmbH. Web design by medani webdesign

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).

Introducing SCAPE

The SCAPE Consortium brings together a broad spectrum of expertise from

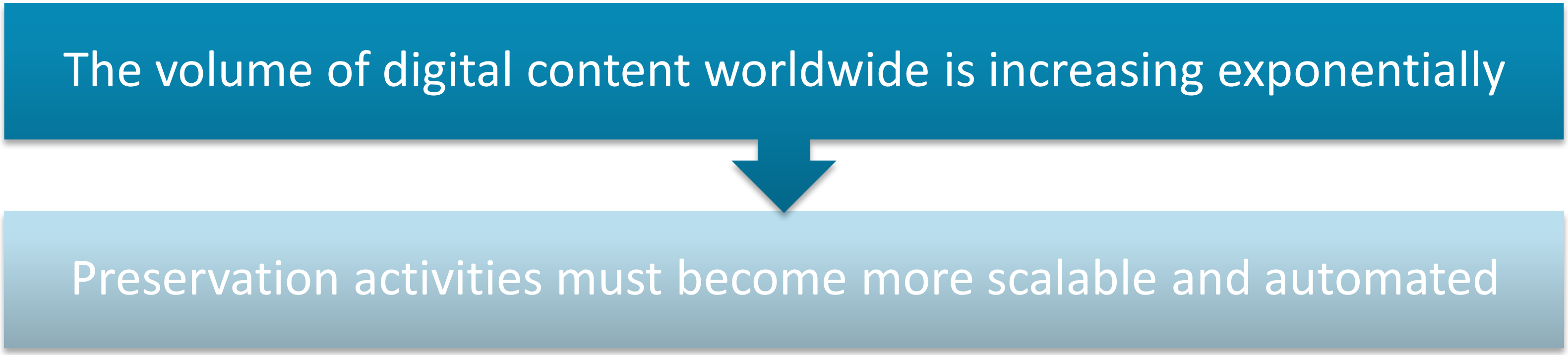


- Memory institutions
- Data centres
- Research labs
- Universities
- Industrial firms

AIT AUSTRIAN INSTITUTE OF TECHNOLOGY
Österreichische Nationalbibliothek
FIZ Karlsruhe
Leibniz Institute for Information Infrastructure
TU Berlin Technische Universität Berlin
BRITISH LIBRARY
iM INTERNET MEMORY FOUNDATION
Open Planets Foundation
KB Koninklijke Bibliotheek National Library of the Netherlands
TU WIEN TECHNISCHE UNIVERSITÄT WIEN Vienna University of Technology
Science & Technology Facilities Council
KEEPSOLUTIONS University of Minho SPIN-OFF
MANCHESTER 1824
ExLibris The bridge to knowledge
STATSBIBLIOTEKET
Microsoft Research
UPMC SORBONNE UNIVERSITÉS
BRNO UNIVERSITY OF TECHNOLOGY
SUPERCOMPUTING POZNAN NETWORKING
Universitatea de Vest din Timişoara
Wielkopolskie Centrum Pulmonologii i Torakochirurgii im. Eugenii i Janusza Zeylandów

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's contribution to digital preservation



The volume of digital content worldwide is increasing exponentially

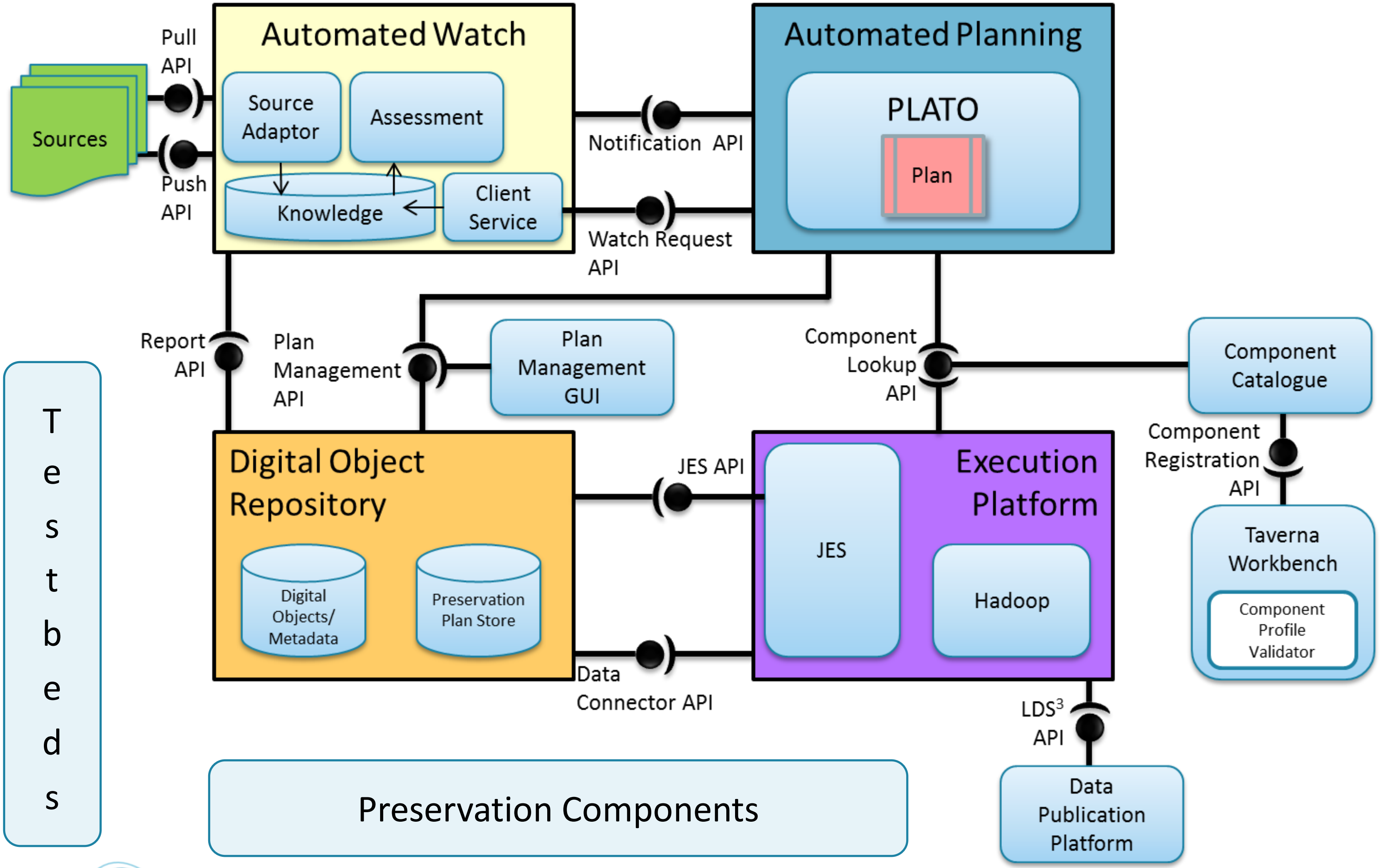
Preservation activities must become more scalable and automated

SCAPE is enhancing the state-of-the-art of long-term digital preservation in terms of

- **Scalability** of preservation actions
- **Automation** and **Quality Assurance** of scalable preservation workflows
- **Preservation Planning** driven by institutional policies

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).

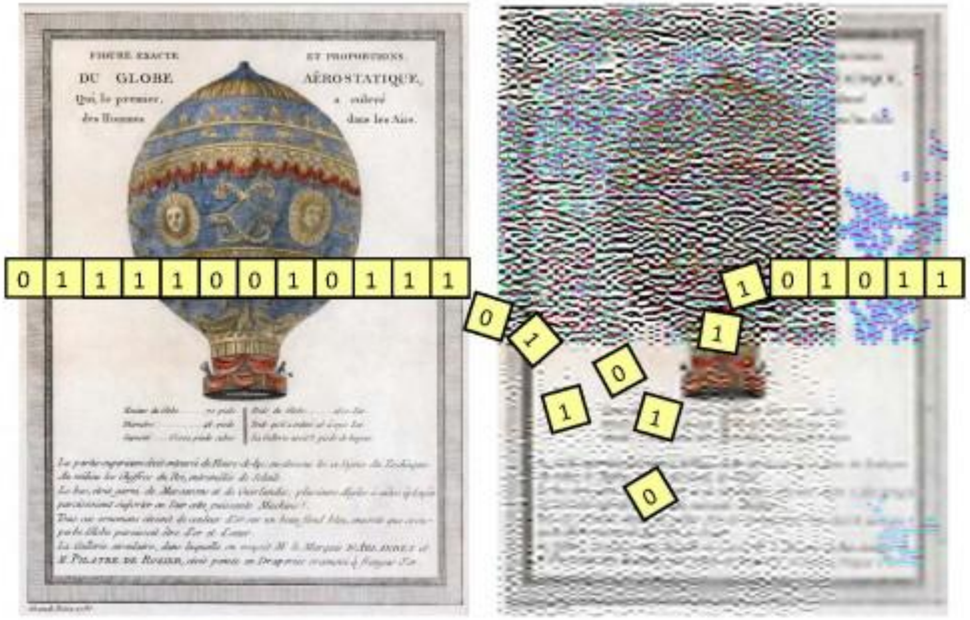
Overview: SCAPE Components



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

- Jpylyzer: quality assurance tool for validation of images in JPEG 2000 format (JP2)
  - Validation against the JP2 format specifications, which ensures that images are standards compliant
  - Extracted image and encoding properties can be validated against an institute-specific profile



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).

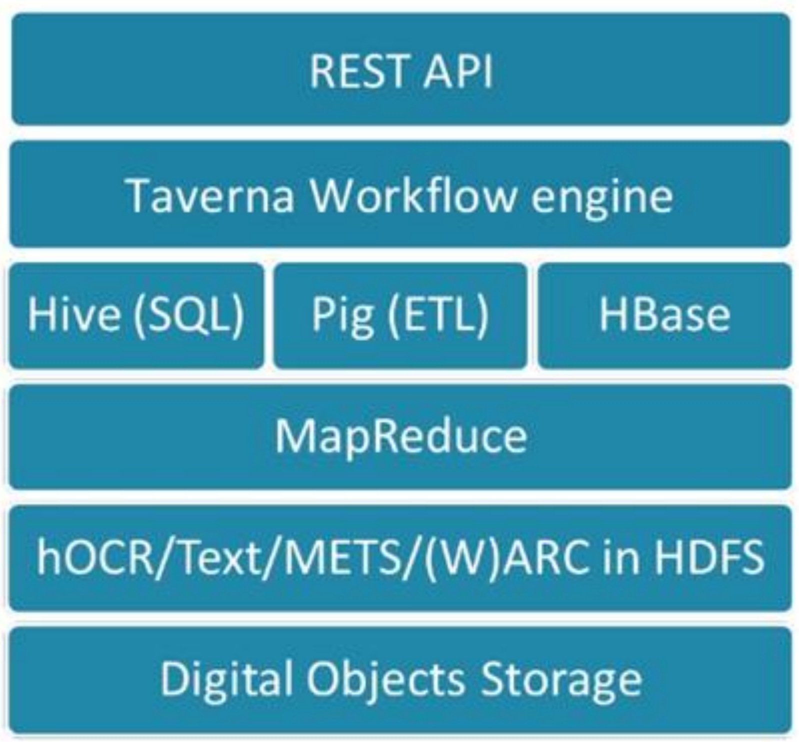
Preservation components

- xcorrSound package
  - QA tools for comparison of audio files
  - Overlap-analysis: detects overlaps between two audio files
  - Sound-match: finds occurrences of shorter WAV files within larger ones;
  - Waveform-compare: analyses audio files for similarity
- C3PO
  - Content profiling tool for preservation analysis
  - Processes FITS (or TIKA) meta data files and generates a profile of the content set in an automated fashion
  - With the Web App you can visualise, filter, and export the data

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 platform

- Reliable storage of voluminous data objects and records
- Parallel execution of preservation tools and workflows close to the data
- Scalable backend which can be attached to different data management systems



REST API

Taverna Workflow engine

Hive (SQL) | Pig (ETL) | HBase

MapReduce

hOCR/Text/METS/(W)ARC in HDFS

Digital Objects Storage

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).

APARSEN and scalability

- APARSEN has a work package on scalability
- Deliverable D27.1 “Recommendations about scalability”
  - Understand what the important scalability parameters are in preservation systems
  - Understand the scalability requirements of the preservation systems for the next few years
  - Identify gaps in technology that prevent us from getting to the right level of scalability
  - Summarize challenges and recommend areas that need to be addressed

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).

APARSEN and scalability

- Survey of repositories
  - Growth in volume and complexity
  - Majority use home-grown solutions
  - “Creating the next level of scalable systems cannot be achieved by point improvements to non scalable systems”
  - Most use and maintain own storage

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).

Services for assisting with preservation activities

APARSEN deliverable D21.1 “Overview of preservation services”

OAIS functional entity	High-level services
Ingest	Characterization Quality assurance Policy-based assessment Automated metadata creation
Preservation planning	Environment monitoring (preservation watch) Knowledge model comparison Preservation plan formulation Obsolescence substitution Dependency management
Data management	—
Archival storage	Long-term archiving service
Administration	Preservation actions Transformation Metadata migration
Access	Finding Federated search

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

Scalable Preservation Environments

APARSEN

Alliance Permanent Access to the Records of Science in Europe Network

www.scape-project.eu

www.aparsen.eu