Supporting the preservation lifecycle in repositories

Luis Faria lfaria@keep.pt
KEEP SOLUTIONS www.keep-solutions.com

Open Repositories 2013
Charlottetown, PEI, Canada, 2013-07-09

http://goo.gl/V6142
KEEP SOLUTIONS

- Company specialized in information management
- Digital preservation experts
- Open source: RODA, KOHA, DSpace, Moodle, etc.
- Scientific research
  - SCAPE: large-scale digital preservation environments
  - 4C: digital preservation cost modeling

http://www.keep-solutions.com
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 past: RODA 1.0.0

- Presented in Open Repositories 2009
- Open source digital repository
- Based on Fedora Commons
- Modern web interface
- For archives
- For digital preservation
Disseminations of 20 - "National Geographic"

- JPEG (original)
- TIFF (normalized)
- Photo preview
- Book preview
The present: RODA Community

- Adapted to be a true open-source project
- For users
  - Easy to install
  - Easy to test (virtual machine)
  - Support mailing lists and documentation
  - Free or paid support
- For developers
  - Development and translation guidelines
  - Easy build (maven)
  - Available on GitHub
  - Support mailing lists
  - Plenty more documentation
- More info: http://www.roda-community.org

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).
RODA
The world’s most advanced open-source digital repository

An open-source digital repository designed for preservation

RODA is a complete digital repository that delivers functionality for all the main units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to the various types of digital content produced by large corporations or public bodies. RODA is based on open-source technologies and is supported by existing standards such as the OAIS, METS, EAD and PREMIS.

Conforms to open standards
RODA follows open standards using EAD for description metadata, PREMIS for preservation metadata, METS for structural metadata, several standards for technical metadata (e.g. NISO Z39.87 for digital still images).

Vendor independent
RODA is 100% built on top of open-source technologies. The entire infrastructure required to support RODA is vendor independent. This means that you may use the hardware and Linux distributions that best fit your institutional needs.

Scalable
The service-oriented nature of RODA’s architecture allows the system to be highly scalable.

Embedded preservation actions
Preservation actions and management within RODA is handled by a task scheduler. The task scheduler is designed to allow users to run preservation actions, check preservation state, and log preservation actions.
Current practice problems

- Repository has content
- Organization has policies in place (e.g. no compression allowed)

P1: Does the content conform to policies? Are there any risks? Even on a changing content, policies and environment?

- Found a preservation risk!

P2: How to easily and trustworthily decide which action to take?
Current practice problems

• Know what action to take

P3: How to ensure and monitor the quality of chosen action and that the decision assumptions remain valid?

• Content grows exponentially in volume, heterogeneity and complexity

P4: How to do digital preservation in large-scale environments?
Preservation lifecycle

Environment and users

access, ingest, harvest

Repository
Preservation lifecycle

Environment and users

Watch

monitored environment and users

monitored content and events

access, ingest, harvest

Repository
Preservation lifecycle

Environment and users

Watch

Planning

Repository

monitored environment and users

create/re-evaluate plans

access, ingest, harvest

monitored content and events
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 lifecycle

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 lifecycle

Environment and users

Repository

access, ingest, harvest

monitored environment and users

Watch

monitored content and events

Policies

monitored actions

Operations

School action plan

Planning

create/re-evaluate plans

deploy plan

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 lifecycle (in practice)

Environment and users

Repository

Operations

Watch

Policies

Planning

monitored environment and users

monitored content and events

monitored actions

create/re-evaluate plans

deploy plan

access, ingest, harvest

execute action plan

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 lifecycle (in practice)

Environment and users

Repository

Operations

Workflow engine

Scout

Watch

Policies

Planning

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 Preservation Suite

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 Preservation Suite

Environment and users
- access, ingest, harvest

Repository
- execute plan

Workflow engine
- Workflow engine API

Data Connector API
- monitored events and actions

Plan management API
- monitored content

Scout Adaptors
- Scout Web UI & Email notification

Scout
- Planner
  - create/re-evaluate plans

Plato
- Plato Web UI
  - deploy plan

Small scale: Taverna
Large scale: SCAPE platform

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).

http://www.taverna.org.uk
http://wiki.opf-labs.org/display/SP/SCAPE+Platform
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 Preservation Suite
Tools and APIs
SCAPE Digital object model

• Standard model for representing digital objects
• Based on METS and PREMIS
• Specifies intellectual entity (SIP, AIP and DIP)
• Specification:
  https://github.com/openplanets/scape-platform-api
Data Connector API

- Access and modify content on the repository
- HTTP REST API
- Methods:
  - **Retrieve** intellectual entity, metadata, representation, file or named bit stream
  - **Ingest** intellectual entity (sync or async)
  - **Update** intellectual entity, representation or file
  - **Search** intellectual entities, representations or files (SRU)
- API specification: [https://github.com/openplanets/scape-platform-api](https://github.com/openplanets/scape-platform-api)
- Ref. implementation: Fall 2013 in Fedora 4 and RODA
Report API

- Provides access to repository events

- Events:
  - **Ingest** started and finished
  - **Viewed** or **downloaded** descriptive metadata or representation
  - Preservation **plan executed**

- OAI-PMH data provider

- PREMIS events metadata
  - Agent: **who** triggered the event
  - Date/time: **when** did the event occur
  - Details: **what** happened

- API specification: [https://github.com/openplanets/scape-platform-api](https://github.com/openplanets/scape-platform-api)

- Ref. implementation: [https://github.com/openplanets/roda](https://github.com/openplanets/roda)
Scout: a preservation watch system

- Monitors aspects of the world to detect preservation risks and opportunities
- Triple store
- Adaptors
  - Data Connector & Report API
  - SCAPE Policy model
  - PRONOM
  - Web semantic extraction
  - Renderability experiments
- Web interface
- Triggers: templates and SPARQL
- Email notifications
Dashboard

All about your own information.

My triggers

You have no triggers defined, create one now!

+ Create trigger

My policies

<table>
<thead>
<tr>
<th>Objective</th>
<th>Measure</th>
<th>Description</th>
<th>Modality</th>
<th>Qualifier</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Running costs per object</td>
<td>Running operational costs of an action in € per object.</td>
<td>MUST</td>
<td>LT</td>
<td>0.24</td>
</tr>
<tr>
<td>1</td>
<td>elapsed time per MB</td>
<td>elapsed processing time per Megabyte of input data, measured in milliseconds</td>
<td>MUST</td>
<td>LT</td>
<td>2000</td>
</tr>
<tr>
<td>2</td>
<td>stability judgement</td>
<td>Judgement of the stability of an action</td>
<td>SHOULD</td>
<td>stable</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ease of integration</td>
<td>Assessment of how easy it is to integrate an action into a particular server environment.</td>
<td>SHOULD</td>
<td>good</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>software licence source code</td>
<td>Indicates if and in which way the source code of the software is accessible.</td>
<td>MUST</td>
<td>openSource</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ease of use</td>
<td>Assessment of how easy it is to use an action in operations</td>
<td>SHOULD</td>
<td>openSource</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>image width equal</td>
<td>true if image width has been preserved.</td>
<td>MUST</td>
<td>true</td>
<td></td>
</tr>
</tbody>
</table>
Collection size  The overall size

43.97 GB

Data type: Very big integer number (File or storage size).

Property history: There are 8 different values of this property, this is number 7 (starts at 0).

Value provenance: Current value was measured 1 times by 1 different sources.

Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.
Format distribution

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tagged Image File Format</td>
<td>160</td>
</tr>
<tr>
<td>Hypertext Markup Language</td>
<td>23</td>
</tr>
<tr>
<td>Portable Document Format</td>
<td>17</td>
</tr>
<tr>
<td>Plain text</td>
<td>16</td>
</tr>
<tr>
<td>XLS</td>
<td>16</td>
</tr>
<tr>
<td>FPX</td>
<td>9</td>
</tr>
<tr>
<td>Microsoft Word</td>
<td>7</td>
</tr>
<tr>
<td>Extensible Markup Language</td>
<td>2</td>
</tr>
<tr>
<td>Extensible Hypertext Markup Language</td>
<td>2</td>
</tr>
<tr>
<td>Postscript</td>
<td>2</td>
</tr>
<tr>
<td>Macromedia Flash data (compressed), version 6</td>
<td>1</td>
</tr>
<tr>
<td>Macromedia Flash data, version 5</td>
<td>1</td>
</tr>
<tr>
<td>PPT</td>
<td>1</td>
</tr>
<tr>
<td>news or mail, ASCII text</td>
<td>1</td>
</tr>
</tbody>
</table>
Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.
Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.
### Category

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>format</td>
<td>Represents a file format</td>
</tr>
</tbody>
</table>

#### Entities

1-20 of 843

<table>
<thead>
<tr>
<th>Name</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcast WAVE[audio/x-wav; version=0]</td>
<td></td>
</tr>
<tr>
<td>Broadcast WAVE[audio/x-wav; version=1]</td>
<td></td>
</tr>
<tr>
<td>Graphics Interchange Format[image/gif; version=1987a]</td>
<td></td>
</tr>
<tr>
<td>Graphics Interchange Format[image/gif; version=1989a]</td>
<td></td>
</tr>
<tr>
<td>Audio/Video Interleaved Format[video/x-msvideo]</td>
<td></td>
</tr>
<tr>
<td>Waveform Audio[audio/x-wav]</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Value</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Minimum preservation action execution time</td>
<td>1.5002512</td>
</tr>
<tr>
<td>Average preservation action execution time</td>
<td>1.8746954</td>
</tr>
<tr>
<td>Maximum preservation action execution time</td>
<td>2.3340003</td>
</tr>
<tr>
<td>Ingest average time (ms)</td>
<td>1092798.0</td>
</tr>
</tbody>
</table>
Advanced query

Use SPARQL to make your own query

Target
- Category
- Property
- Entity
- Value
- Measurement

Snippets
- Relations
- Resources

SPARQL

```
SELECT ?s WHERE { ?s rdf:type watch:EntityType . }
```

[+ Create trigger] [Search]
Query

Select a pre-made question template or go to **advanced query**.

**Query templates**

- Check collection policy conformance
- Collection size limit

**Check collection policy conformance**

Check if selected collection conforms to the defined policy (only compression scheme policy is checked right now)

**Collection**

The ID from the URL

Your collection profile already inserted into scout

[Search]  [Create trigger]
Plan management API

- Deploy and management preservation plans in the repository
- HTTP REST API

Methods:
- **Search** and **retrieve** plans
- **Deploy** a new plan
- **Retrieve** or **add** a plan execution state (in progress, success or fail)
- **Update** plan lifecycle status (enabled or disabled)

Implementation can use:
- Workflow engine: Taverna or SCAPE platform
- Data connector API

API specification: [https://github.com/openplanets/scape-platform-api](https://github.com/openplanets/scape-platform-api)

Ref. implementation: Fall 2013 for Fedora 4 and RODA
Plato: a preservation planning tool

http://ifs.tuwien.ac.at/dp/plato

- Systematic planning
- Traceable, documented, trustworthy
- Integrated:
  - Data Connector API (Content)
  - Scout (Watch, Content profile, sampling)
  - SCAPE Policy model
  - Plan management API (Operations)
  - Taverna compatible workflows

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 Plato?

Digital content is short-lived, yet may prove to have value in the future. How can we keep it alive? Finding the right action to enable future access to digital content in a transparent way is the task of Plato.

The mission of digital preservation is to ensure continued, authentic long-term access to digital objects in a usable form for specific user communities. This requires preservation actions to be carried out when the original environment of digital objects is unavailable. A variety of preservation actions exist, but each shows specific peculiarities, and a variety of factors influence the decision.

The **mission of preservation planning** is to ensure authentic future access for a specific set of objects and designated communities by defining the actions needed to preserve it.

The planning tool **Plato** is a decision support tool that implements a solid preservation planning process and integrates services for content characterisation, preservation action and automatic object comparison in a service-oriented architecture to provide maximum support for preservation planning endeavours.

What's new?

**June 2013: Plato 4.2**

We are pleased to announce the release of Plato 4.2, including:

- Better policy integration
- Improved selection of alternatives
- Improved tree views
- Initial MyExperiment migration service allows passing parameters

For further details check the [Github milestone](https://github.com/ifs-tuwien/plato/releases).
Preservation lifecycle (in practice)

- Environment and users
  - monitored environment and users
- Repository
  - access, ingest, harvest
  - monitored content and events
  - execute action plan
- Operations
  - Workflow engine
  - monitored actions
- Scout
- Policies
  - create/re-evaluate plans
- Planning
  - deploy plan
- Plato
Conclusions

P1: Does the content **conform to policies**? Are there any **risks**? Even on a **changing** content, policies and environment?

S1: Use Scout: preservation watch system

P2: How to easily and **trustworthily decide** which action to take?

S2: Use Plato: preservation planning tool
Conclusions

P3: How to ensure and monitor the **quality** of chosen action and that the **decision assumptions** remain valid?

S3: Q&A in preservation plans (Plato), monitoring of Q&A (Report API & Scout), automatic Scout triggers created by Plato

P4: How to do digital preservation in large-scale environments?

S4: Automation and end-to-end integration of preservation processes.
Roadmap

• Scout:
  • User support
  • More adaptors
  • More trigger templates

• Plato:
  • Automatic create Scout triggers
  • Automatic deploy using plan management API

• Repository reference implementations: RODA and Fedora 4
Conclusions

- All APIs published
- Ref. implementations in RODA and Fedora 4 in Fall 2013
- All tools available in Github

Add preservation to your repository now!
Supporting the preservation lifecycle in repositories

Luis Faria  lfaria@keep.pt
KEEP SOLUTIONS  www.keep-solutions.com

Open Repositories 2013
Charlottetown, PEI, Canada, 2013-07-09

http://goo.gl/V6142