A Framework for Distributed Preservation Workflows

Rainer Schmidt
AIT Austrian Institute of Technology
rainer.schmidt@ait.ac.at
Outline

- Overview of the Integrated Environment
  - Main Objectives and Architecture
- Planets Preservation Services
- Digital Objects and Metadata
- Integrating Repositories
- The Workflow Execution Engine (WEE)
  - Workflow Templates
  - WEE Services and the GUI
Planets Project

- “Permanent Long-term Access through NETworked Services”
- Addresses the problem of digital preservation
  - driven by National Libraries and Archives
- Project instrument: FP6 Integrated Project
- 5. IST Call
- Consortium: 16 organisations from 7 countries
- Duration: 48 months, June 2006 – May 2010
- Budget: 14 Million Euro
- [http://www.planets-project.eu/](http://www.planets-project.eu/)
The Planets Environment

- An integrated System for the development and evaluation of preservation strategies.
- Uniform access mechanisms to a broad range of “commodity” tools, e.g. for characterization, migration, emulation.
- Integration of existing repositories, data/metadata formats.
- Specification, execution, recording of preservation workflows.
- Integration with end-user applications for preservation planning and the evaluation of tools стратегии.
  - PLANETS Preservation Planning Tool and Testbed
Service-Orientated Architecture

- XML Web Services (SOAP, WSDL, WS-*)
  - Platform, Language, and Location Independence
- Homogeneous interfaces for preservation activities, data management, workflow execution.
  - Remotely access repositories and data.
  - Discover and dynamically utilize tools in a workflow.
- Supports distributed and cross-organizational deployments
  - Shared hardware, software, maintenance
  - Browser-based access to large number of resources
Preservation Interfaces (the Verbs)

- Define atomic preservation activities (level-one)
  - Concentrates on low-level concepts and actions
    - Bit-stream operations, no data management
  - Designed to be light-weight and easy to implement
- Independent from a specific tool, language, or content type
  - *E.g.* Characterize, Migrate, Compare, CreateView
  - >50 Tools wrapped/provided as Planets Services
- Provides the basic abstractions for assembling workflows.
## Preservation Interfaces (the Verbs)

- **Define atomic preservation activities (level-one)**
  - Concentrates on low-level concepts and actions
  - Bit-stream operations, no data management
  - Designed to be light-weight and easy to implement
  - Independent from a specific tool, language, or content type
  - E.g. Characterize, Migrate, Compare, CreateView
  - > 50 Tools wrapped/provided as Planets Services
  - Provides the basic abstractions for assembling workflows.

### Planets IF Service Registry

<table>
<thead>
<tr>
<th>Name</th>
<th>Registered</th>
<th>Service Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>FloppyImageHelperUnix</td>
<td></td>
<td>Droid Service Endpoint Information</td>
</tr>
<tr>
<td>FloppyImageHelperWin</td>
<td></td>
<td>VSDL: <a href="http://ubuntu.planets-project.arcs.ac.at:80/">http://ubuntu.planets-project.arcs.ac.at:80/</a></td>
</tr>
<tr>
<td>FloppyImageModifyWin</td>
<td></td>
<td>Status: OK</td>
</tr>
<tr>
<td>AvidemuxMigration</td>
<td></td>
<td>DROID Identification Service</td>
</tr>
<tr>
<td>JhoveValidation</td>
<td></td>
<td>Type</td>
</tr>
<tr>
<td>JavaImageIO Migrate</td>
<td></td>
<td>eu.planets_project.services.identify.Identify</td>
</tr>
<tr>
<td>DioscuriArjMigration</td>
<td></td>
<td>Author</td>
</tr>
<tr>
<td>SanselanMigrate</td>
<td></td>
<td>CarlVilson, Fabian Steeg</td>
</tr>
<tr>
<td>PdfBoxMigration</td>
<td></td>
<td>Provider</td>
</tr>
<tr>
<td>InkscapeMigration</td>
<td></td>
<td>The Planets Consortium.</td>
</tr>
<tr>
<td>Droid Identity</td>
<td></td>
<td>Description</td>
</tr>
<tr>
<td>OpenJpegMigration</td>
<td></td>
<td>Identification service based on Droid (DROID 3.0, Signature File 15).</td>
</tr>
<tr>
<td>DioscuriPnmToPngMigration</td>
<td></td>
<td>You are not Authorised to Register Services</td>
</tr>
</tbody>
</table>

Your user account lacks the privileges required to register services. You should contact the IF Administrator and request that they register this service for you.

If you believe that this is an oversight, or that you have previously been able to register services then contact the IF Administrator.
Digital Objects (the Nouns)

- Generic data abstraction for modeling digital entities.
  - Encapsulates content and metadata
  - Consumed and/or produced by Planets preservation services
- Provides minimal and generic model for data management
  - Stored in Metadata Repository
- Does not prescribe serialization schema
  - May be created from DC/ORE RDF record and be serialized using METS/PREMIS schemas.
Digital Objects (the Nouns)

- **Events**: Type, Time, Agent, Service, Result, …
- **Properties**: Creator, Title, Description, Format, …
- **Metadata**: Tagged Uninterpreted Metadata Chunks
- **Content**: Embedded Data or Repository URL
  - **Digital Object**: fragment contains_object
  - **Relationships (possibly associated with event)**
Digital Object Managers

- Individual adapters for retrieving (& storing) Planets DOs
  - Provide access to existing repositories.
  - Map metadata records to Planets DOs
  - Ingest digital objects to Planets data repositories
- Current implementation for
  - retrieving OAI-PMH records, BL digitized newspaper, Web resources, Amazon S3 buckets, …
- Planets Data Registry services (ingesting DOs) based on Apache Jackrabbit and Fedora Commons.
Digital Object Managers

Data Registry Service

Different data sources, repositories, interfaces, and protocols.
You may upload data to the PLANETS FTP area using your wiki username and password and the following host name and port number:

Host name: ftp://www.planets-project.eu

The option to restrict the data based on supplied criteria will be added in a later version.
Workflow Definitions / Templates

- Separation of concerns:
  - Fragments of complex workflow logic are implemented by <<workflow developers>>
  - <<Experimenters>> selected from predefined templates, configure them, and execute individual processes.

- Templates implement abstract and reusable processes definitions based on level-on operations (API) and decision logic.

- Execute in trusted environment (level-two)
  - handle digital objects in metadata repository and
  - basis for recording provenance and preservation information
Workflow Execution Engine (WEE) Service

1. **register**
   - Workflow Developer
   - WEE Template Rep. Service

2. **select**
   - Experimenter
   - Workflow Client Application

3. **configure**
   - XML
   - WEE Execution Service

4. **execute**
   - Template
   - WEE Execution Service
The image shows a workflow control panel with various tabs and options for managing templates, loading workflows, editing workflows, selecting digital objects, and checking workflow status.

### Edit an XML Workflow Description

<table>
<thead>
<tr>
<th>Service ID</th>
<th>Type</th>
<th>Endpoint</th>
<th>Name/Value Pairs</th>
<th>Add Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>identify</td>
<td>Identify</td>
<td>DROID Identification Service</td>
<td></td>
<td>Edit Parameters</td>
</tr>
<tr>
<td>migrate</td>
<td>Migrate</td>
<td>GraphicsMagickMigrate</td>
<td></td>
<td>Edit Parameters</td>
</tr>
</tbody>
</table>

#### Parameters

<table>
<thead>
<tr>
<th>Name</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>compressionType</td>
<td>None, BZip, Fax, Group4, JPEG, Lossless, LZW, RLE, Zip</td>
</tr>
<tr>
<td>imageQuality</td>
<td>0-100</td>
</tr>
</tbody>
</table>

### Active Template

- `eu.planets_project.itr.core.wae.impl.templates.IdentifyMigrateTemplate`

### Active Workflow

- `none (in memory)`

### Input Objects Selected

- `false`

### Submit Workflow

The workflow can be submitted for execution.
Summary

- Research infrastructure for
  - integrating variety of tools and repositories
  - executing defined preservation operations
  - recording provenance and preservation metadata

- Not necessary an “out-of-the-box” solution
  - Extensible network of services,
  - Public deployment,
  - Allows sharing of resources and results.

- Downloadable package available for local installation of selected preservation tools/services.
Fin