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/
The Planets Software 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/strategies.
  - 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.
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

Properties

Metadata

Content

Digital Object

Type, Time, Agent, Service, Result, …

Creator, Title, Description, Format, …

Contains_object

Fragment

Embedded Data or Repository URL

Tagged Uninterpreted Metadata Chunks

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

GUI / App.

dm1

dm2

dm3

retrieve dig. obj.

Different data sources, repositories, interfaces, and protocols.

XML

SOAP

HTML

PMH/ORE

Fedora

WEB
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**
   - Experimenter
   - WEE Template Rep. Service

2. **Select**
   - WEE Template Rep. Service
   - Workflow Client Application

3. **Configure**
   - Workflow Client Application
   - XML

4. **Execute**
   - Workflow Client Application
   - WEE Execution Service

**Template**
- Cmp.
- Cmp.
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