Workflows for Digital Preservation and Curation Workshop

Open Repositories 2012

Stacy Kowalczyk
Beth Plale
Kavitha Chandrasekar
Yiming Sun
Agenda

• Introduction to Digital Curation
• Workflow Systems Overview
• Workflows for Digital Curation
• Break
• Implementing Workflows in Trident
• Modifying a Workflow
• Create a new Workflow
• Creating Components
• Wrap up
Acknowledgements

• This workshop was made possible through a generous grant by Microsoft Research

• And by the Data to Insight Center of Indiana University’s Pervasive Technology Institute

• Quan Zhou, Ph.D. student and developer, for his help with developing components, workflows, and documentation
Introduction to Digital Curation

• Defining curation
• Infrastructure for curation
• Curating the files
• Curating the object
Defining Curation

Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle.

The active management of research data reduces threats to their long-term research value and mitigates the risk of digital obsolescence. Meanwhile, curated data in trusted digital repositories may be shared among the wider ... research community.

As well as reducing duplication of effort in research data creation, curation enhances the long-term value of existing data by making it available for further high quality research.
Curation Infrastructure

- Repository
- Public access
- Policies
- Processes
- Institutional support
Curating the Files

• Bitstream Integrity
  – Fixity
  – Duplicate copies
• File integrity
  – Format verification
  – Format validation
File Formats

• Durability
  – Transparency
  – Documentation
  – Ubiquity
  – Renderability
  – Longevity
Format Choices

• Master files for preservation
  – Highest quality
  – Highest fidelity
  – Lossless

• Derivative files for active use and delivery
  – Smallest possible for user needs
  – Fast delivery
  – Easy to use format
Curating the Object

• Context
  – Relationships between files
  – Technical metadata
  – Intellectual metadata

• To Metadata
  – Implicit/explicit context
Curation Activities

• Ongoing verification
  – File integrity
  – Object integrity

• Metadata management

• Management of obsolescence
  – Hardware
  – Software
  – Formats
  – Documentation
Workflow Systems

• Purpose of workflow systems
• Types of workflow systems
• Trident Workflow Workbench
Why Workflow Systems

- Repetitive and mundane activities simplified
- Facilitates and enforces best practices
- Enables efficient scheduling
- Machinery for coordinating the execution of services and linking together resources
- Facilitates outreach to researchers for direct deposit and automatic curation
Types of Workflow Systems

Kepler

Taverna

Ptolemy II

BPEL
Trident

• Open source project
• Based on Microsoft Workflow Foundation classes
• Supported by Microsoft Research and academic researchers
• Integrates with myExperiment
• Well accepted in the research community
  – well over 100 peer-reviewed and white papers were discovered from one scholarly aggregation service
Trident Components

• Trident Management Studio
• Trident Workflow Composer
• Trident Workflow Application
• Microsoft SQL Server
• Trident Silverlight client for web execution of workflows
• Microsoft Visual Studio
  – C# development environment
Design

Visual Workflow Composer

Trident Registry

Workflow Packages
(domain specific)

Windows Workflow Foundation
.NET 4.0

Trident Runtime Services

Provenance

Monitoring

Workflow Scheduling Service

Admin

Admin Console

Workflow Monitor

Workflow Launcher

Community

Web Portal
Search
Launch Monitor

Results Repository

Workflow Repository (myExperiment)

Data Access Layer

Data Object Model (data source abstraction layer)

Data Storage Providers: SQL Server, Local XML store, ...
Workflows for Curation

• Goals
  – Systematic and repeatable processes
  – Helps remove human errors

• Data Ingest
  – Integrity checks
  – Format normalization/derivative generation
  – Metadata creations

• Curation activities
  – Integrity checks
  – Format migration
  – Media migration
Data Ingest Workflows

• Scenarios
  – Single part objects (individual images)
  – Multi-part objects (a book)
  – Multiple instantiations of a logical object (word, pdf and ppt of a research paper)
  – Multiple multi-part objects (a group of letters)
  – Research data products (multiple files of various types)
  – Scientific workflow process
Single Part Objects Workflow

- Magic Lantern Slides
  - Individual files
  - Spreadsheet

- Fixity Check
- Format Validation and Verification
- Image Quality Checks
- Derivative Generation
- Persistent Identification
- Create Tech Metadata
- Create Intellectual Metadata
- Create Object Metadata
- Deposit in Repository
Multi-part Object Workflow

- Comic Book
  - RIS
  - Set of .tif files

Object Integrity

Fixity Check

Format Validation and Verification

Image Quality Checks

Derivative Generation

Create Tech Metadata

Create Intellectual Metadata

Persistent Identification

Create Object Metadata

Deposit in Repository
Multiple Instantiations of a Logical Object Workflow

• Papers
  – Each logical object per subdirectory
  – RIS, word file and (perhaps) supplemental file

Fixity Check
Format Validation and Verification
Format Normalization
Derivative Generation
Create Tech Metadata
Create Intellectual Metadata
Persistent Identification
Create Object Metadata
Deposit in Repository
Multiple Multi-part Object Workflow

- Ball collection
  - RIS for collection and Inventory spreadsheet
  - Each logical object in separate subdirectory

- Object Integrity
- Collection Integrity
- Fixity Check
- Format Validation and Verification
- Image Quality Checks
- Derivative Generation
- Create Tech Metadata
- Create Intellectual Metadata
- Persistent Identification
- Create Object Metadata
- Create Collection Metadata
- Deposit in Repository
Research Data Products

• Vortex
  – Each subdirectory is an experiment with FGDC metadata

- Compress Data
- Fixity Check
- Persistent Identification
- Create Intellectual Metadata
- Create Object Metadata
- Deposit in Repository
Workflow Components

• Format Conversions (for normalization and derivative generation)
  – .xlsx to .csv
  – .docx to .pdf
  – .ppt to .pdf
  – .tif to .jpg
  – Zipping on demand
  – Image (.tif or .jpg) to .pdf
Workflow Components 2

• Context creation
  – MIX data generator and validator
  – METS data generator and validator

• Data Integrity
  – MD5 checksum generator
  – MD5 checksum validator
  – JHOVE for format verification and validation
  – Group validation (for object integrity)
Post Deposit Curation Workflow

• Scenarios
  – Fixity verification
  – Format normalization
  – New or additional derivative generation
  – Media migration
  – Persistent identifier updates
  – Metadata updates
Workflows in Trident
Executing Workflows

- Individual object ingest
- Multipart object ingest
- Multiple multipart object ingest
- Multiple instantiations of a single logical object
- Research data ingest
- Scientific workflow
- Fixity check curation workflow
Implementing Workflows in Trident

• Launch the Remote Desktop application
• User: AMAZONA-JJOAL14\oruser
• PWD: TridentOR12!!
• Computer ip addresses on slip of paper being passed out now.
Trident Workflow Composer
Participant Exercises
Modifying Workflows

• Add components to existing workflows
• Select the **Individual Ingest Workflow**
  – Add DOI component
    • Before the METS generator component
    • Make the connections
• Select the **Group Ingest Workflow Comic**
  – Add the METS generation component
    • After the last component in the main line
    • Make the connections
Simple Curation Workflow Creation

• Create a Workflow for a simple curation process – validate MD5 checksums
  – Define a directory of image files
  – Define a METS file
  – Define an output location
  – Link the MD5 checksum validation component
  – Link the MD5 checksum report component
  – Save and execute the workflow
Creating Components

• Exercise:
  – Create a new Trident workflow component
  – Implement the MARCXML to MODS Stylesheet
    http://www.loc.gov/standards/mods/v3/MARC21slim2MODS3-4.xsl
  – Kavitha Chandrasekar will demonstrate the process
Wrap Up

• Thumb drives
• Trident codeplex site
• Trident listserv
• Contributing to Trident
• Workshop Evaluation Form
• Ongoing conversation
Contacts for Further Discussion

• Trident CodePlex site: http://tridentworkflow.codeplex.com/
• Trident Listserv: trident-wf-l@iulist.indiana.edu
• Stacy Kowalczyk: skowalcz@indiana.edu
• Kavitha Chandrasekar: kavchand@imail.iu.edu
• Yiming Sun: yimsun@umail.iu.edu
• Quan Zhou: quzhou@indiana.edu