Overview of Workflows and Microservices

Kari Smith & Courtney Mumma
Workflow Concepts

• Definition of workflow:
  – Description of practice and procedures
  – Automation of repetitive tasks
  – Graphic representation of flow of work

• Workflow engine concepts:
  – Orchestrations: composition and execution of new services (definition)
  – Choreography: interaction/coordinated action between services (description)
Workflow Influences

• Critical path method (project management)
  1. List all activities
  2. Determine time (duration) for completion
  3. Identify dependencies between activities

• Process Improvement examples:
  – Six Sigma
  – Total Quality Management (TQM)
  – Business Process Reengineering
Considerations

• Perspective
  – As is: document what is happening now
  – To be: document what should happen

• Right-sized
  – Appropriate granularity for problem, setting
  – Extent and type of documentation

• Maintenance
  – changes in staff, roles
  – New or changed functions
Benefits of Workflows

- Efficiency
- Shareable
- Scalable
- Consistency
- Common Outcomes
- Progress
Role of Workflows

• Manage risks and expectations
• Enable handshakes (people, technology)
• Identify bottlenecks
• Determine gaps and fill (as is/to be)
• Define and illustrate language (terms)
• Capture and apply decisions
• Document for transparency
• Support packaging (TIPR, Archivematica)
Impact of Workflows

Like policies, producing a workflow:

• Raises awareness
• Improves understanding
• Enables communication
• Captures commitment
• Supports planning (tools, investments)
Workflows:
An ICPSR Example
Step 2: Deposit Study Material

The Depositor -- that is, the Investigator or someone depositing material on behalf of the Investigator -- prepares and submits study materials to ICPSR for dissemination and preservation. The Investigator signs the deposit, which transfers custody over it to ICPSR.

Typically, the Depositor performs the deposit directly into the ICPSR web site’s Deposit Form.

Details
- Submitting a Study
- Submission Criteria
- Data Collection Preparation Tips
- Mode of Transmission
- Contact ICPSR Acquisitions Staff
- Obtaining a Study
- Depositing Data and Documentation

Resources
- ICPSR Guide to Social Science Data Preparation and Archiving, 4th Edition
- Data Deposit Form
- Handling Removable Media
- Receiving Physical Materials
ICPSR Pipeline: Step 2 Detail
ICPSR Pipeline: Step 4 Detail

Check for Issues:
- Check data, codebook, instruments and any other related files for inconsistencies or other issues such as those listed below:

<table>
<thead>
<tr>
<th>Issue Description</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Missing Data&quot; Values</td>
<td>Check data for &quot;missing data&quot; values provided by the PI's stat package (like SAS's .a, .b, ...) or other explicit values such as &quot;inapplicable&quot;, &quot;don't know&quot;, and &quot;no answer&quot;.</td>
</tr>
<tr>
<td>Undocumented Codes</td>
<td>Check the frequency output file for coded values that are invalid or insufficiently documented.</td>
</tr>
<tr>
<td>Inadequate Formats</td>
<td>Check data for variables with formats that fail to accommodate all values seen.</td>
</tr>
<tr>
<td>Nonportable Data Types</td>
<td>Check the work file for potential issues of portability across stat packages, like: data variables, variable names with special characters, variable names longer than 16 chars, values longer than max string length, missing values</td>
</tr>
</tbody>
</table>

Resolve Issues:
- Note any mismatches in a spreadsheet of issues to be resolved with the PI. For each issue, record the dataset name, variable name, and a description.
- Recode any nonstandard missing data values (like SAS's) to a standard numbering scheme (like 9-fills), using the PI's stat package (i.e., before converting data to SPSS). Recode to ensure consistency across datasets and variables.
- Label the undocumented-code data values as such, or assign appropriate labels, or recode.
- Recode the data or revise the variable format.
- Convert nonportable variable formats to portable ones. Example: replace stat package-specific date variables with numeric variables with formats like "yyyyymmdd".
MIT Example
MIT Example: A1 ver. 1

Diagram:
- Decide to Digitize (A1)
  - Identify physical/analog content to digitize
  - Determine whether to add digital object to collection
    - NO: Ad-hoc request that is returned to physical collection
    - YES: Determine scope of project, Identify funding sources, Establish priorities and schedule, List roles and responsibilities
  - Define Digitization Requirements (A2)
MIT Example: A1 ver. 2
MICROSERVICES
CDL Microservices

Object Modeling
- Merritt Object Modeling

Identity Services
- EZID
- ARK: Archival Resource Key
- NOID: Nice Opaque Identifier (Minter and Name Resolver)
- N2T: Name-to-Thing
- UC3 Identifier Conventions

Storage Services
- Storage
- CAN: A Simple File System-Based Object Store
- D-flat: A Simple File System Convention for Digital Object
- PairTree: Object Storage
- ReDD: Reverse Directory Deltas

Ingest Services
- Ingest
- BagIt

Fixity Service
- Fixity

Characterization Services
- JHOVE2
- Unified Digital Format Registry (UDFR)

Access Services
- Access
- Merritt Repository LDAP Access Control
- Merritt Data User Agreements (DUA)

Common Tools
- ANVL: A Simple Record Syntax
- Checkm: A Checksum-based Manifest Format
- Datenorm: Date normalization
- ERC: Electronic Resource Citation and Kernel Metadata
- Locklt: A Simple File-based Convention for Resource Locking
- Namaste: Directory Description with Namaste Tags
- RUU: (Are You You?): User account authentication
- THUMP: The HTTP URL Mapping Protocol

https://wiki.ucop.edu/display/Curation/Microservices
### Archivematica 1.0

<table>
<thead>
<tr>
<th>Micro-service</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Verify SIP compliance</strong></td>
<td>Verifies that the SIP conforms to the folder structure required for processing in Archivematica. The structure is as follows: logs, metadata, metadata/submissionDocumentation, objects.</td>
</tr>
<tr>
<td><strong>Verify transfer compliance</strong></td>
<td>Verifies the METS from the transfer.</td>
</tr>
<tr>
<td><strong>Rename SIP directory with SIP UUID</strong></td>
<td>Directly associates the SIP with its metadata by appending the SIP UUID to the SIP directory name and checks if SIP is from Maildir transfer type to determine workflow.</td>
</tr>
<tr>
<td><strong>Include default SIP processingMCP.xml</strong></td>
<td>Copies the processing configuration file added to the transfer in Include default Transfer processingMCP.xml, above, to the SIP.</td>
</tr>
<tr>
<td><strong>Remove cache files</strong></td>
<td>Removes any thumbs.db files.</td>
</tr>
<tr>
<td><strong>Clean up names</strong></td>
<td>Some file systems do not support unicode or other special characters in filenames. This micro-service removes prohibited characters and replaces them with dashes. Original filenames are preserved in the PREMIS metadata.</td>
</tr>
<tr>
<td><strong>Normalize</strong></td>
<td>Determines which normalization options are available for the SIP and presents them to the user as choices. Normalizes (i.e. generates preservation and/or access copies) based on selection. Thumbnail files are also generated during this micro-service.</td>
</tr>
<tr>
<td><strong>Process submission documentation</strong></td>
<td>Processes any submission documentation included in the SIP and adds it to the /objects/directory.</td>
</tr>
<tr>
<td><strong>Process metadata directory</strong></td>
<td>Processes metadata.</td>
</tr>
<tr>
<td><strong>Prepare DIP</strong></td>
<td>Creates a DIP containing access copies of the objects, thumbnails and a copy of the METS file.</td>
</tr>
<tr>
<td><strong>Upload DIP</strong></td>
<td>Allows the user to choose to upload the DIP to either ICA-AtoM or CONTENTdm.</td>
</tr>
<tr>
<td><strong>Upload DIP to ICA-AtoM</strong></td>
<td>The user uploads the DIP to a selected description in ICA-AtoM.</td>
</tr>
<tr>
<td><strong>Upload DIP to CONTENTdm</strong></td>
<td>The user uploads the DIP to a selected description in CONTENTdm.</td>
</tr>
<tr>
<td><strong>Prepare AIP</strong></td>
<td>Creates an AIP in BagIt format. Creates the AIP pointer file. Indexes the AIP, then losslessly compresses it.</td>
</tr>
<tr>
<td><strong>Store AIP</strong></td>
<td>Moves the AIP to /sharedDirectoryStructure/www/AIPsStore/ or another specified directory. Once the AIP has been stored, a copy of it is extracted from storage to a local temp directory, where it is subjected to standard BagIt checks: verifyvalid, checkpayloadoxum, verifycomplete, verifypayloadmanifests, verifytagmanifests.</td>
</tr>
</tbody>
</table>
Workflow Development

• Working with variety of organizational contexts
  – Looking at use cases
  – Compliance with OAIS while making workflow decisions

• Integration Issues and Considerations
  – Interacting with legacy systems
  – Technological suitability