SERI
State Electronic Records Initiative
COUNCIL OF STATE ARCHIVISTS

Advanced Electronic Records Institute

Overview of Workflows and Microservices

Kari Smith & Courtney Mumma

INSTITUTE of Museum and Library SERVICES

This project made possible by a grant from the U.S. Institute of Museum and Library Services

Workflow Concepts

- Definition of workflow:
  - Description of practice and procedures
  - Automation of repetitive tasks
  - Graphic representation of flow of work
- Workflow engine concepts:
  - Orchestration: 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

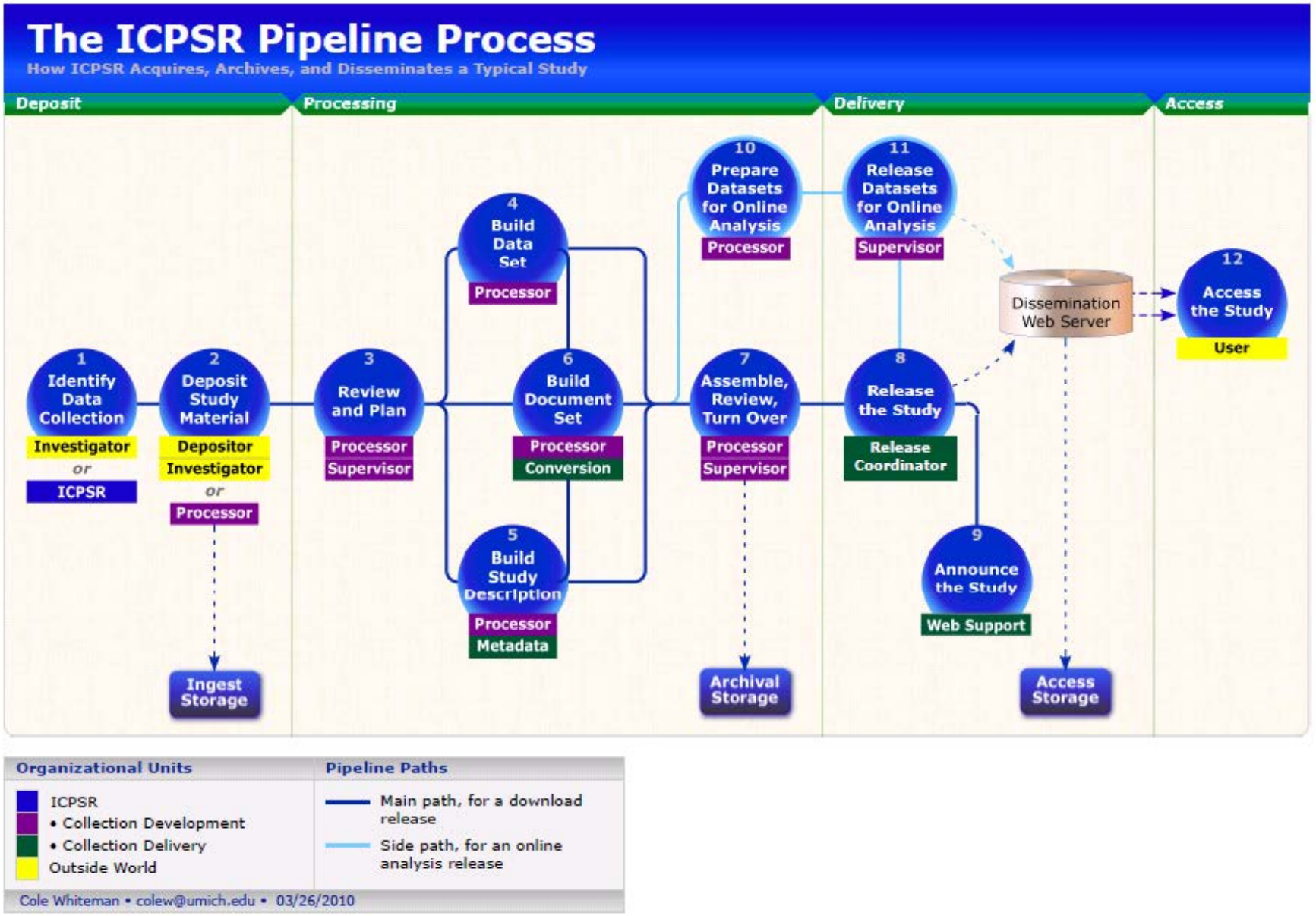
- Manage risks and expectations
- Enable handshakes (people, technology)
- Identify bottlenecks
- Determine gaps and fill (as is/to be)
- Define and illustrate language (terms)
- Capture and apply decisions
- Document for transparency
- Support packaging (TIPR, Archivematica)

Impact of Workflows

Like policies, producing a workflow:

- Raises awareness
- Improves understanding
- Enables communication
- Captures commitment
- Supports planning (tools, investments)

Workflows:
An ICPSR Example

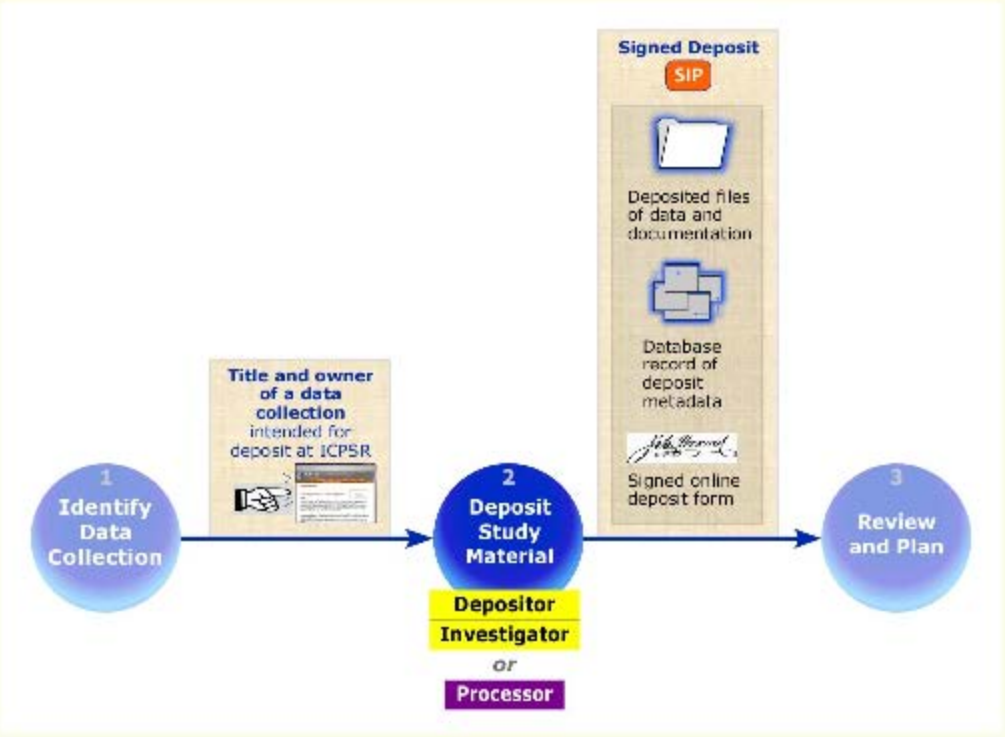

The ICPSR Pipeline Process
How ICPSR Acquires, Archives, and Disseminates a Typical Study
Deposit
Processing
Delivery
Access
1 Identify Data Collection
Investigator or ICPSR
2 Deposit Study Material
Depositor Investigator or Processor
3 Review and Plan
Processor Supervisor
4 Build Data Set
Processor
5 Build Study Description
Processor Metadata
6 Build Document Set
Processor Conversion
7 Assemble, Review, Turn Over
Processor Supervisor
8 Release the Study
Release Coordinator
9 Announce the Study
Web Support
10 Prepare Datasets for Online Analysis
Processor
11 Release Datasets for Online Analysis
Supervisor
12 Access the Study
User
Dissemination Web Server
Ingest Storage
Archival Storage
Access Storage
Organizational Units
ICPSR
• Collection Development
• Collection Delivery
Outside World
Pipeline Paths
Main path, for a download release
Side path, for an online analysis release
Cole Whiteman • colew@umich.edu • 03/26/2010

ICPSR Pipeline: Step 2

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.



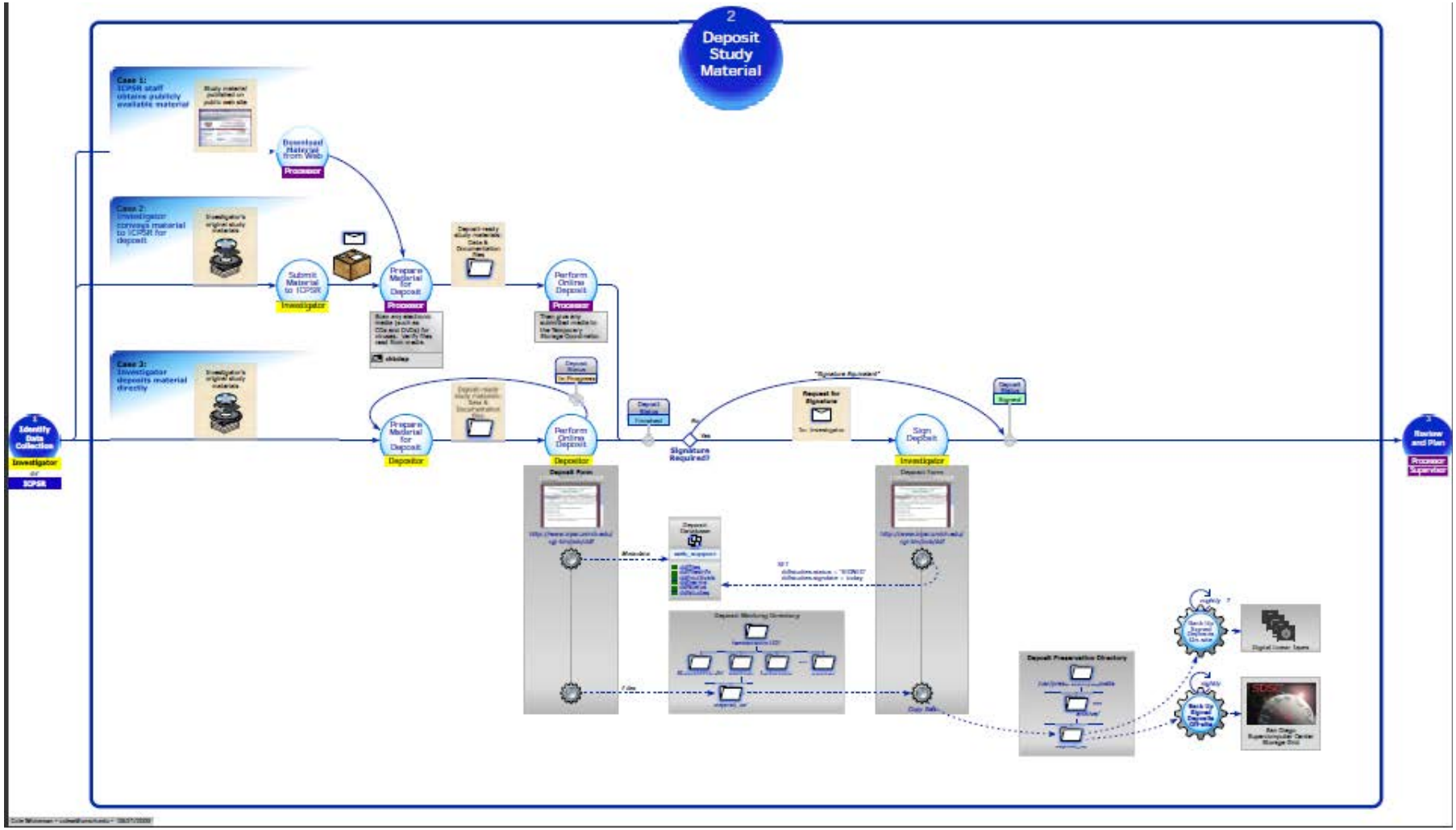
Details

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

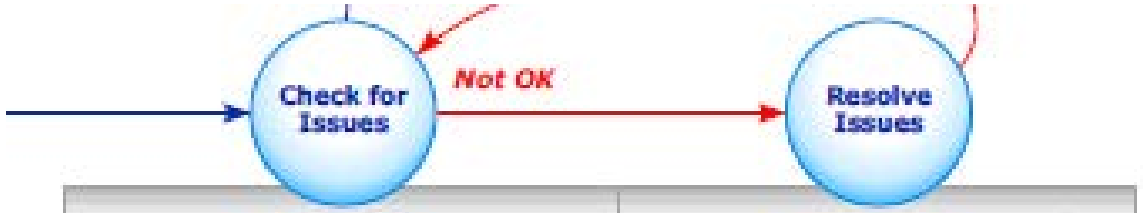
Resources

- 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

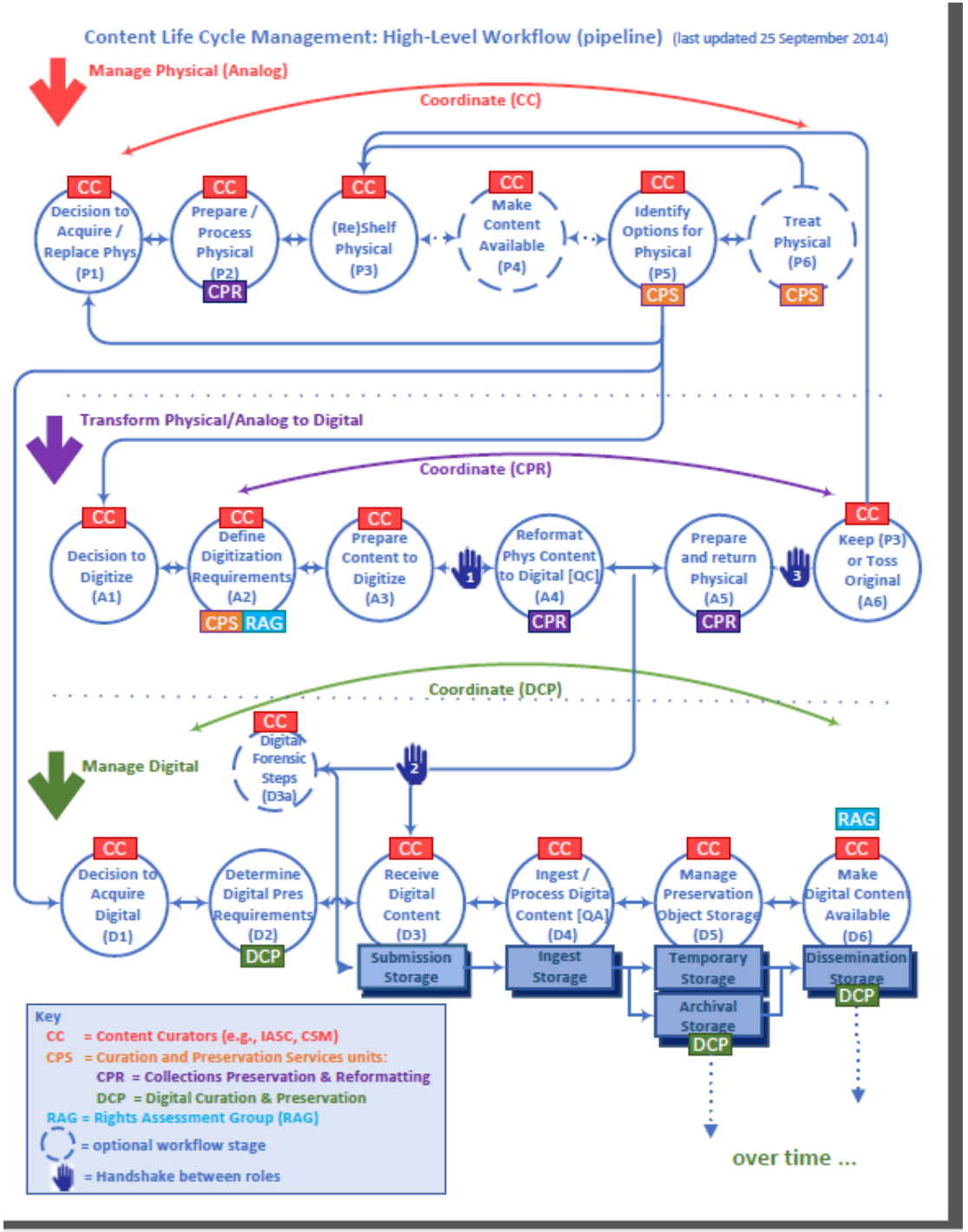


ICPSR Pipeline: Step 4 Detail

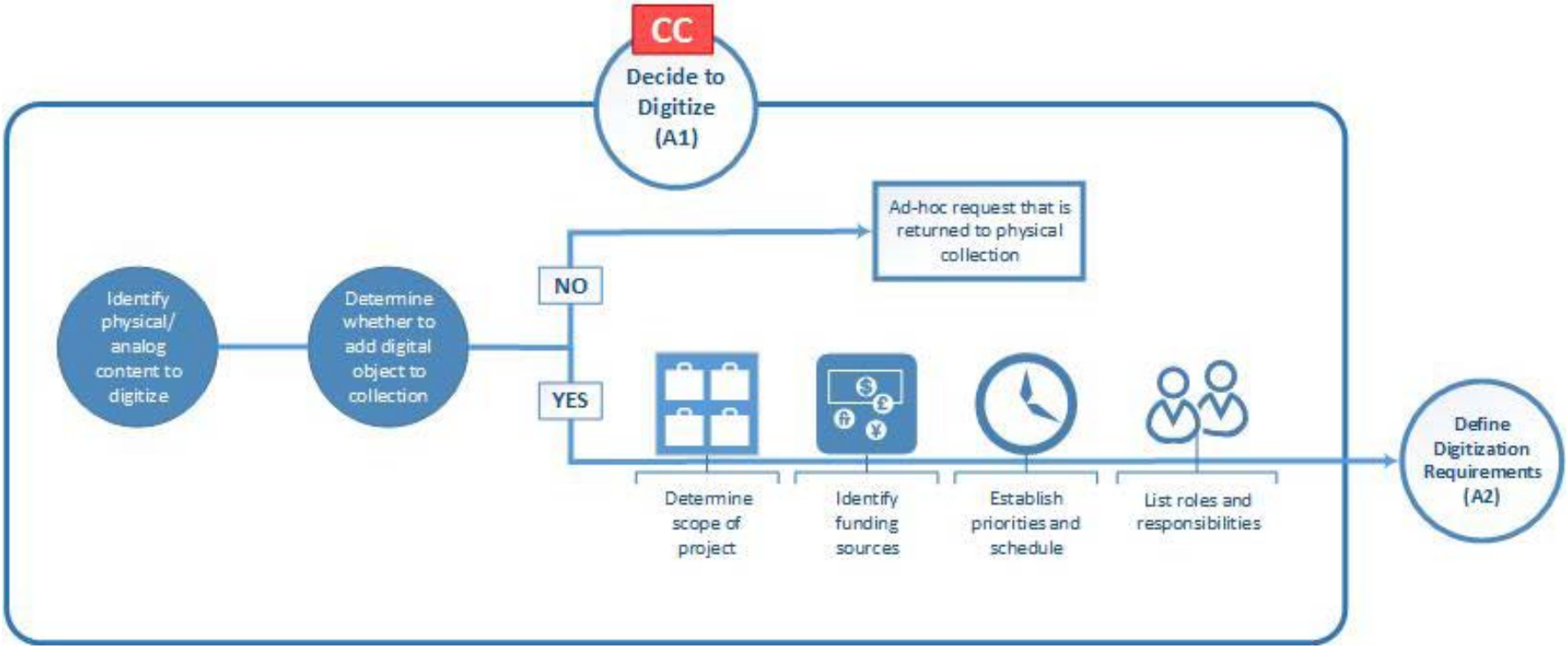


Check data, codebook, instruments and any other related files for inconsistencies or other issues such as those listed below:	Note any mismatches in a spreadsheet of issues to be resolved with the PI. For each issue, record the dataset name, variable name, and a description.
"Missing Data" Values Check data for "missing data" values provided by the PI's stat package (like SAS's .a, .b, ...) or other explicit values such as "inapplicable", "don't know", and "no answer".	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.
Undocumented Codes Check the frequency output file for coded values that are invalid or insufficiently documented.	Label the undocumented-code data values as such, or assign appropriate labels, or recode.
Inadequate Formats Check data for variables with formats that fail to accommodate all values seen.	Recode the data or revise the variable format.
Nonportable Data Types Check the work file for potential issues of portability across stat packages, like: • date variables • variable names with special characters • variable names longer than 16 chars • values longer than max string length • missing values	Convert nonportable variable formats to portable ones. Example: replace stat package-specific date variables with numeric variables with formats like "yyyymmdd".

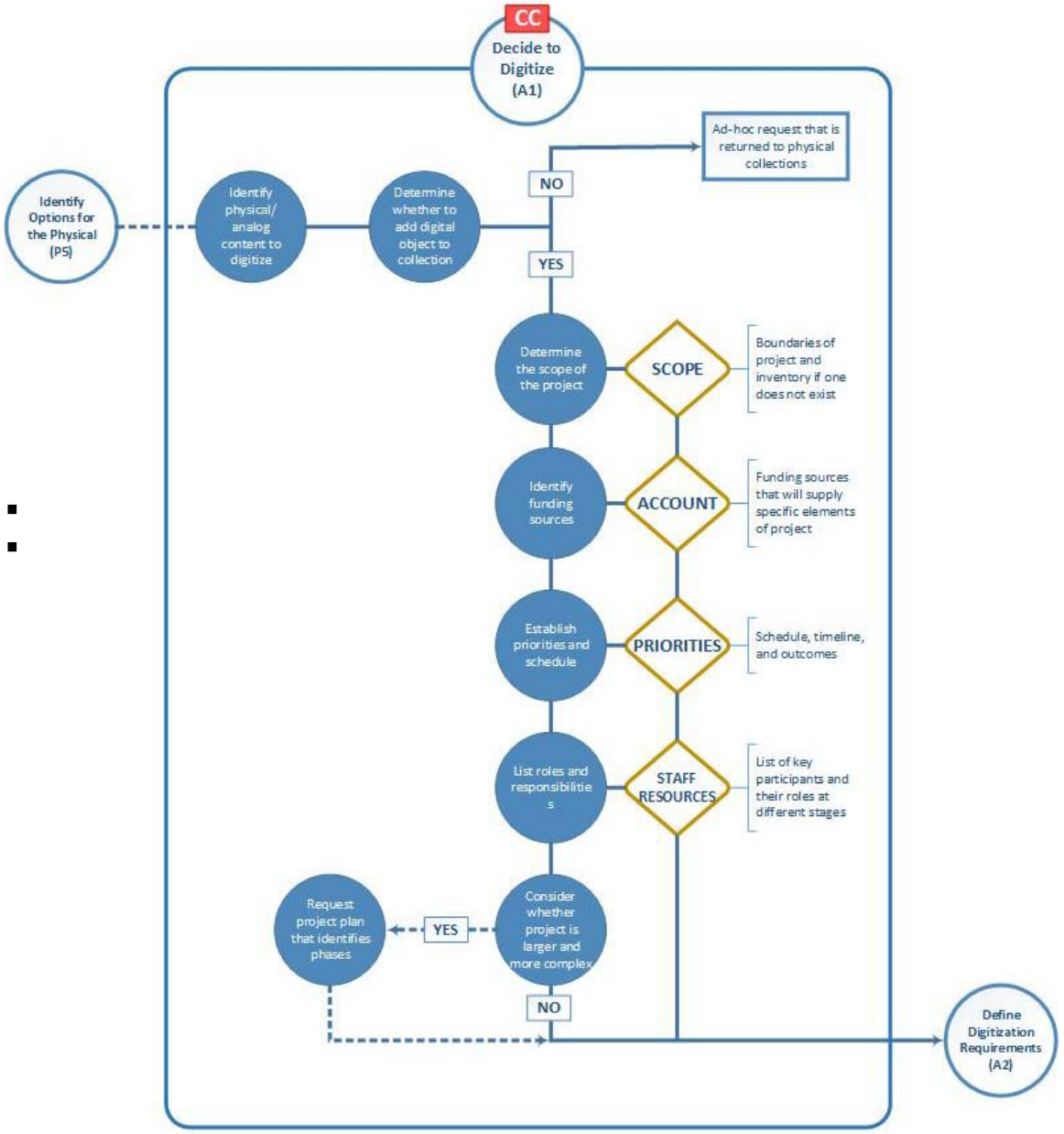
MIT Example


Content Life Cycle Management: High-Level Workflow (pipeline) (last updated 25 September 2014)
Manage Physical (Analog)
Coordinate (CC)
CC
Decision to Acquire / Replace Phys (P1)
Prepare / Process Physical (P2)
CPR
(Re)Shelf Physical (P3)
Make Content Available (P4)
Identify Options for Physical (P5)
CPS
Treat Physical (P6)
CPS
Transform Physical/Analog to Digital
Coordinate (CPR)
Decision to Digitize (A1)
Define Digitization Requirements (A2)
CPS RAG
Prepare Content to Digitize (A3)
1
Reformat Phys Content to Digital [QC] (A4)
CPR
Prepare and return Physical (A5)
CPR
3
Keep (P3) or Toss Original (A6)
Coordinate (DCP)
Manage Digital
Digital Forensic Steps (D3a)
2
Decision to Acquire Digital (D1)
Determine Digital Pres Requirements (D2)
DCP
Receive Digital Content (D3)
Submission Storage
Ingest / Process Digital Content [QA] (D4)
Ingest Storage
Manage Preservation Object Storage (D5)
Temporary Storage
Archival Storage
DCP
RAG
Make Digital Content Available (D6)
Dissemination Storage
DCP
over time ...
Key
CC = Content Curators (e.g., IASC, CSM)
CPS = Curation and Preservation Services units:
CPR = Collections Preservation & Reformatting
DCP = Digital Curation & Preservation
RAG = Rights Assessment Group (RAG)
= optional workflow stage
= Handshake between roles

MIT Example: A1 ver. 1

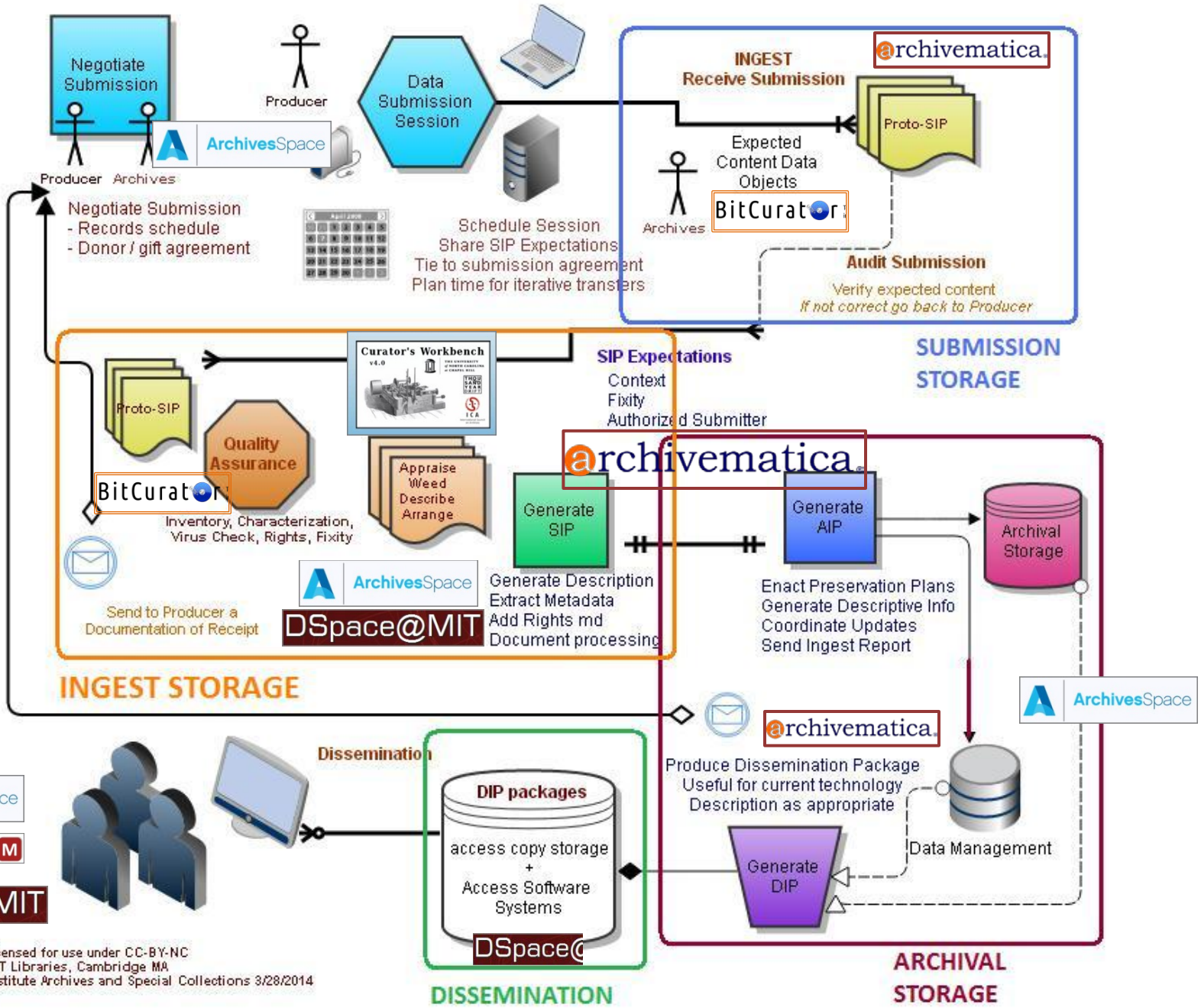


MIT Example: A1 ver. 2



PAIMAS and OAIS diagram showing Producer and Archive roles with tools and storage spaces, version 2.0

Kari R. Smith, Digital Archivist at MIT Libraries



MICROSERVICES

2003-2004

DSTC

PANIC: Preservation & Archival of New media & Interactive Collections

Dr. Jane Hunter, Sharmin Choudhury (DSTC)

Project Objectives

- Compare emulation, migration and metadata approaches to multimedia preservation;
- Determine the optimum media formats, authoring tools, metadata & preservation processes to maximize longevity, accessibility and preservation of multimedia objects;
- Develop recommendations and guidelines for multimedia content creators and collecting agencies;
- Implement metadata schemas, metadata capture tools, workflows for capturing essential metadata and automating preservation actions;
- Implement a Preservation Web Services Architecture that enables:
  - Automatic notification that a preservation action is required – monitor latest software version registries;
  - OWL-S ontologies for describing, discovering and matching preservation services;
  - Automatic discovery of appropriate preservation service(s);
  - Semi-automated invocation of preservation services;
  - Choreographing simple preservation services into more complex services.

Preservation Issues

- Composite – digital (images, video, audio, text) + physical
- Highly proprietary – software & hardware dependent
- Dynamic, variable, chaotic, random – computer-generated, high interactivity, live, situation-dependent
- Ephemeral – high obsolescence rate, high fragility
- Difficult to capture – boundary problem
- High mobility – travelling exhibitions
- No standards or best practice guides
- Obsolescence of physical storage device
- Obsolescence of file formats
- Translation to new display device, playback software – may alter the "look", "feel", "meaning"
- Creator's input is essential – need to capture/recreate key ideas, intentions, aspects, components

Example

Nam June Paik
The elements (1989)

- Lacquer cabinet
- 6 TV monitors with Laser discs
- Roots, 3 silk flowers
- 3 glass beakers
- 195x126x65 cm
- The Kenneth and Yasuko Myer Collection

Copyright: Queensland Art Gallery

Metadata Schema

Extended METS Schema

Metadata Encoding and Transmission Standard

Composite Multimedia Object

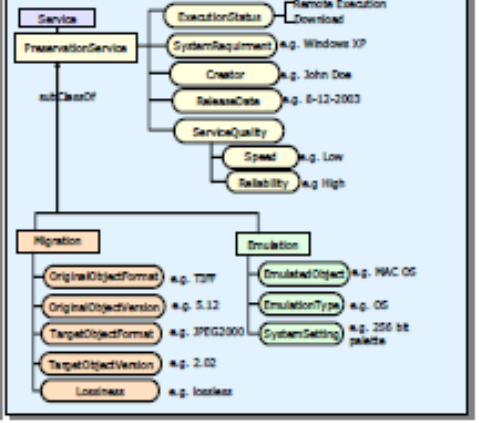
- Descriptive metadata
- Administrative
  - Technical metadata
  - Rights metadata
  - Source metadata
  - DigiProv metadata
- File Groups
- Structural Map

Extensions

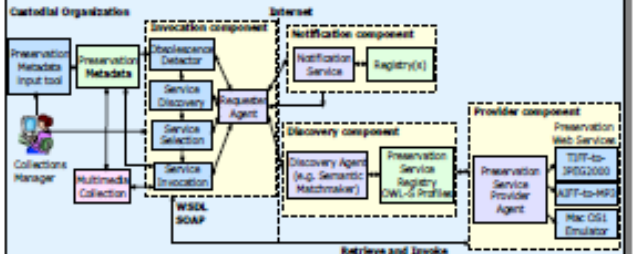
- Presentation Metadata
- Intention Metadata

- Images – TIFF, PNG, JPEG2000
- Audio – AIFF, WAV, MP3
- Video – MPEG-2, MPEG-4
- Composite object – SMIL
  - Synchronized Multimedia Integration Language

Preservation Services Ontology



Web Services Architecture



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
- LockIt: 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

Ingest

Micro-service	Description
Verify SIP compliance [Expand]	Verifies that the SIP conforms to the folder structure required for processing in Archivematica. The structure is as follows: */logs/*, */metadata/*, */metadata/submissionDocumentation/*, */objects/*.
Verify transfer compliance [Expand]	Verifies the METS from the transfer.
Rename SIP directory with SIP UUID [Expand]	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.
Include default SIP processingMCP.xml [Expand]	Copies the processing configuration file added to the transfer in **Include default Transfer processingMCP.xml**, above, to the SIP.
Remove cache files [Expand]	Removes any thumbs.db files.
Clean up names [Expand]	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.
Normalize [Expand]	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.
Process submission documentation [Expand]	Processes any submission documentation included in the SIP and adds it to the */objects/*directory.
Process metadata directory [Expand]	Processes metadata.
Prepare DIP [Expand]	Creates a DIP containing access copies of the objects, thumbnails and a copy of the METS file.
Upload DIP [Expand]	Allows the user to choose to upload the DIP to either ICA-AtoM or CONTENTdm.
Upload DIP to ICA-AtoM [Expand]	The user uploads the DIP to a selected description in ICA-AtoM.
Upload DIP to CONTENTdm [Expand]	The user uploads the DIP to a selected description in CONTENTdm.
Prepare AIP [Expand]	Creates an AIP in Bagit format. Creates the AIP pointer file. Indexes the AIP, then losslessly compresses it.
Store AIP [Expand]	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.

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