

# Digital Preservation in the IU Libraries

---

Jon Dunn  
Stacy Kowalczyk

*IU Digital Library Brown Bag Series  
September 17, 2008*



**LIBRARIES**

---

INDIANA UNIVERSITY

Bloomington

# Agenda

- Overview of Digital Preservation Basics
  - Preservation Strategies
  - The OAIS Model
  - The 4 Goals of Digital Preservation
- Digital Preservation in the IU Libraries
  - Projects
  - Infrastructure: local and collaborative
- Questions



# Framing the Problem

- Scholarly dissemination
- Cultural history
- Design
- Commerce
- Current knowledge is produced, disseminated and stored in digital format.



# Framing the Problem

- Digital objects are more numerous and mutable than their predecessors.
- Digital objects are more expensive to store than their predecessors.
- Digital objects depend on and are bound to a technical environment/infrastructure—as the environment changes, so might the objects.

***Data will not be preserved by benign neglect.***



# Digital Preservation

Defined as

“the managed activities necessary: 1) For the long-term maintenance of a byte stream (including metadata) sufficient to reproduce a suitable facsimile of the original document and 2) For the continued accessibility of the document contents through time and changing technology” (RLG & OCLC, 2002).

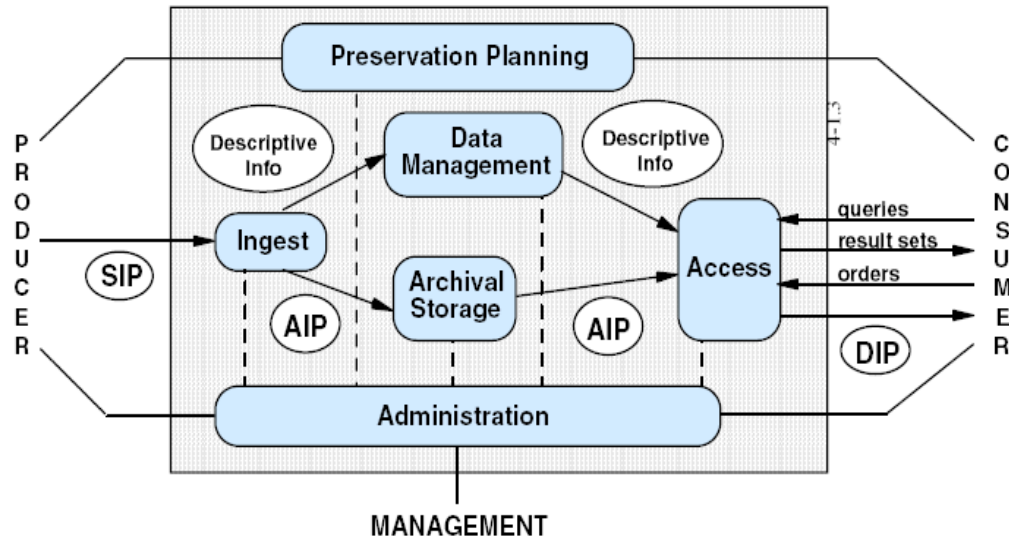


# Preservation Strategies

- Technology preservation
  - Keep the hardware alive
- Technology emulation
  - Create an environment to be able to run the existing software
- Data migration
  - Convert data to new formats to run in new applications



# Open Archival Information System



(CCSDS, 2002)

- SIP – submission information package
- AIP – archival information package
- DIP – dissemination information package



INDIANA UNIVERSITY

# OAIS Impact

- Provided a common language for describing the functions of digital preservation
- Since the first draft of the OAIS in late 1999, most of the research in digital preservation has focused on defining the functions of a digital repository - a system to manage digital objects.



# Four Goals of Preservation

## Preservation Goals

- Keep the bits safe
- Keep the files useable
- Keep the integrity of the object
- Keep the context of the object

Requires an active, systematic program (Waters & Garrett, 1996.)



# Bit Level Integrity

- Keeping the bits safe
- Multiple copies in multiple locations
  - Monitored for obsolescence
  - Monitored for degradation
- ***Repositories should follow Data Center best practice***



# Bit Level Integrity

- Fixity
  - Digital files are easily changed
  - Technology solution is simple
  - Insuring the fixity of each digital file is essential to bit level integrity
- ***To insure fixity, a digital repository should implement a checksum or digital signature on archived files and validate them on a regular schedule.***

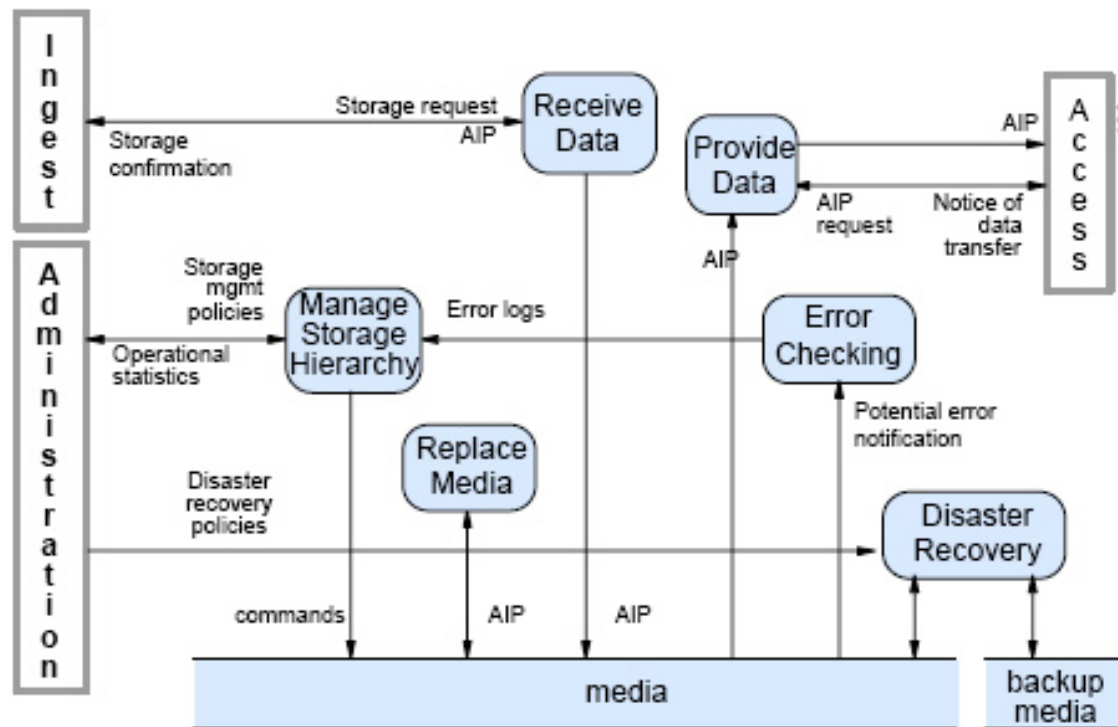


## Bit Level Integrity (2)

- Contingency Planning
  - Develop the contingency planning policy statement
  - Conduct the business impact analysis (BIA)
  - Identify preventive controls
  - Develop recovery strategies
  - Develop an IT contingency plan
  - Plan testing, training, and exercises
  - Plan maintenance. (NIST, 2002)
- ***A digital repository should institute an annual contingency plan drill***



# OAIS Archival Storage Model



# Keep the Files Useable

- This is a much harder problem
- File formats
  - Complex
  - Variable
  - Bound to a technology

“The concept of representation format permeates all technical aspects of digital repository architecture and is, therefore, the foundation of many, if not all, digital preservation activities” (Abrams, 2004).



# Keep the Files Useable (2)

- Formats differ by levels of use
- Risk assessment
  - Library of Congress' 7 sustainability factors
  - National Archives of England, Wales and the United Kingdom's 7 risk factors
  - National Archives of Australia's 8 step evaluation process
  - OCLC's 6 risk factors
- Open
  - Freely available
  - Transparent
- Well documented
- Well supported
- Widely used



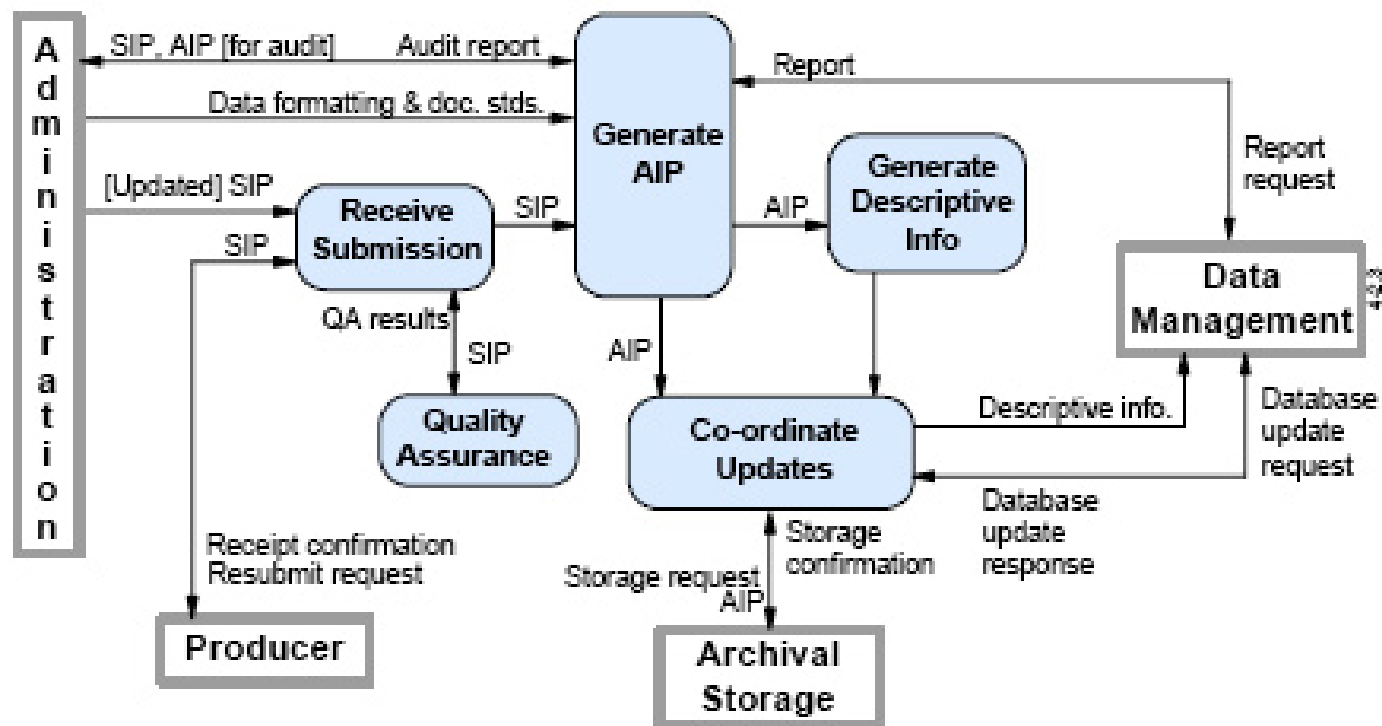
# Keep the Files Useable (3)

## Format Registries

- ***PRONOM*** from the National Archives of England, Wales and the United Kingdom
- ***Format Registry*** of the Digital Curation Centre
- ***Global Digital Format Registry*** (GDFR) sponsored by the Digital Library Federation
  
- Goal – document formats for automatic processing



# Data Ingest



## Keep the Files Useable (4)

- Format Identification
  - File extensions are insufficient
- Format Validation
  - Both well formed and valid
- JHOVE – a joint Harvard/JSTOR project
- DROID –used in conjunction with PRONOM registry
- ***A digital repository should validate digital objects when submitted for ingest.***

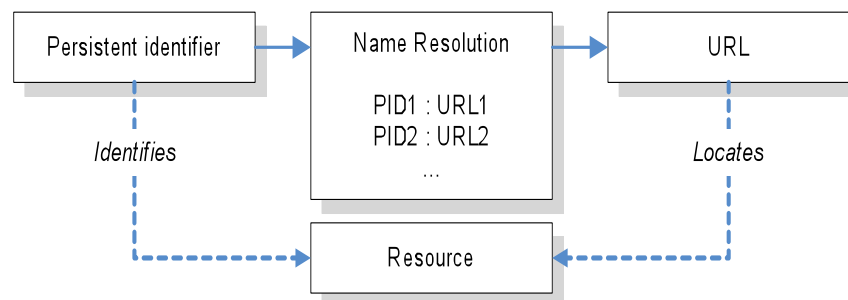


## Keep the Files Useable (5)

- Provenance “broadly refer[s] to a description of the origins of a piece of data the process by which it arrived in a database”
- Currently, Provenance description languages are domain and/or format specific
- ***A digital repository should provide provenance information for all of its digital objects.***



# Keep the Files Useable (6)



## ***Persistent Identifier (PID)***

- Persistent URLs - PURLS
- Digital Object Identifiers - DOIs
- Handles
- Archival Resource Key - ARK
- Name Resolution Service - NRS
- ***A digital repository should implement a persistent identifier service to insure long-term unambiguous access to its digital objects.***



# Keep the Integrity of the Object

*Maintaining the intellectual wholeness of a digital object*

- Implicit Metadata
  - Directory structures and file names
- Explicit Metadata
  - Metadata Encoding and Transmission Standard - METS
  - OAI-ORE
- ***A digital repository needs to maintain the relationships between all of the components of an object.***



# Keeping the Context of the Object

## Preservation System Models

- Digital Repository
  - Dark Archives
  - Integrated access and archiving
- Institutional Repository
- Both require “most essentially an organizational commitment to the stewardship of ... digital materials, including long-term preservation where appropriate, as well as organization and access or distribution” (Lynch, 2003, p. 2).



# Preserving Library Objects

- Static objects
  - Books (images and text)
  - Photographs (images)
  - Time based media (audio, video)
- Dynamic or interactive objects
  - Games
  - Websites
  - Databases



# Digital Preservation Projects at IU

- Most digitization projects have a preservation aspect
  - High quality, high resolution master files
  - Well known file formats
  - Standardized metadata
- Several projects with a focus on digital preservation R&D



# Digital Preservation Projects at IU

- Sound Directions
- Digital Audio Archives Project (DAAP)
- EVIA Digital Archive
- CIC Floppy Disk and SUDOC CD-ROM projects (Lou Malcomb, Geoffrey Brown)
  
- All involve audio, video, and/or born-digital content
  - Loss of existing carriers





# Digital Audio Archives Project (DAAP)

- IMLS-funded partnership 2004-2006:
  - Johns Hopkins University Digital Knowledge Center
  - IU Digital Library Program
  - IU Cook Music Library
  - IU Jacobs School of Music
- R&D: Workflow for efficient high-quality audio digitization
- Digitizing items from JSoM performance archive
- Led to funding of ongoing reformatting operation
- Born-digital recording of new performances



# EVIA Digital Archive

- Mellon-funded partnership:
  - IU Department of Folklore and Ethnomusicology
  - IU Archives of Traditional Music
  - IU Digital Library Program
  - IU UITS Digital Media Network Services
  - University of Michigan
- Preservation of and access to field video
- Video segmentation/annotation tool
- Web access searching and browsing
- <http://www.indiana.edu/~eviada/>



# Local Infrastructure at IU

- Storage
  - Massive Data Storage Service (MDSS)
- Repositories
  - DSpace
  - Fedora



# IU Massive Data Storage System (MDSS)

- Hierarchical storage management
  - Some storage on hard disks
  - Much more storage on automated tape
- Managed by UITS Research Technologies
- Servers in Bloomington and Indianapolis connected via I-Light high-speed fiber link
- Total capacity: 2+ petabytes



# Digital Repositories

- Centrally-managed systems for storage (and delivery) of digital information
- Leverage economies of scale for storage and management costs
- Support preservation integrity functions (migration, replication, validation)
- Much easier to manage than many little pockets of digital information
- Potential single point of failure



# From OAIS to Trusted Digital Repositories

- 2002 OCLC-RLG task force report:
  - Trusted Digital Repositories: Attributes and Responsibilities
- What are the attributes of a trusted repository?
  - OAIS compliance
  - Administrative responsibility
  - Organizational viability
  - Financial sustainability
  - System security
  - Procedural accountability



# Trusted Digital Repositories: Auditing and Certification

- Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist
  - OCLC/NARA/CRL report
  - <http://www.crl.edu/PDF/trac.pdf>
- Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)
  - <http://www.repositoryaudit.eu/>



# Three DLP-supported Repositories

- **IUScholarWorks Repository (DSpace)**
  - [scholarworks.iu.edu](http://scholarworks.iu.edu)
  - “Institutional repository” for preserving and providing access to IU’s research output: articles, conference papers, dissertations, etc.
- **Archives of Institutional Memory (DSpace)**
  - [institutionalmemory.iu.edu](http://institutionalmemory.iu.edu)
  - Repository of IU documents managed by Archives
- **IU Digital Library Repository (Fedora)**
  - [www.dlib.indiana.edu/collections](http://www.dlib.indiana.edu/collections)
  - General-purpose digital content repository

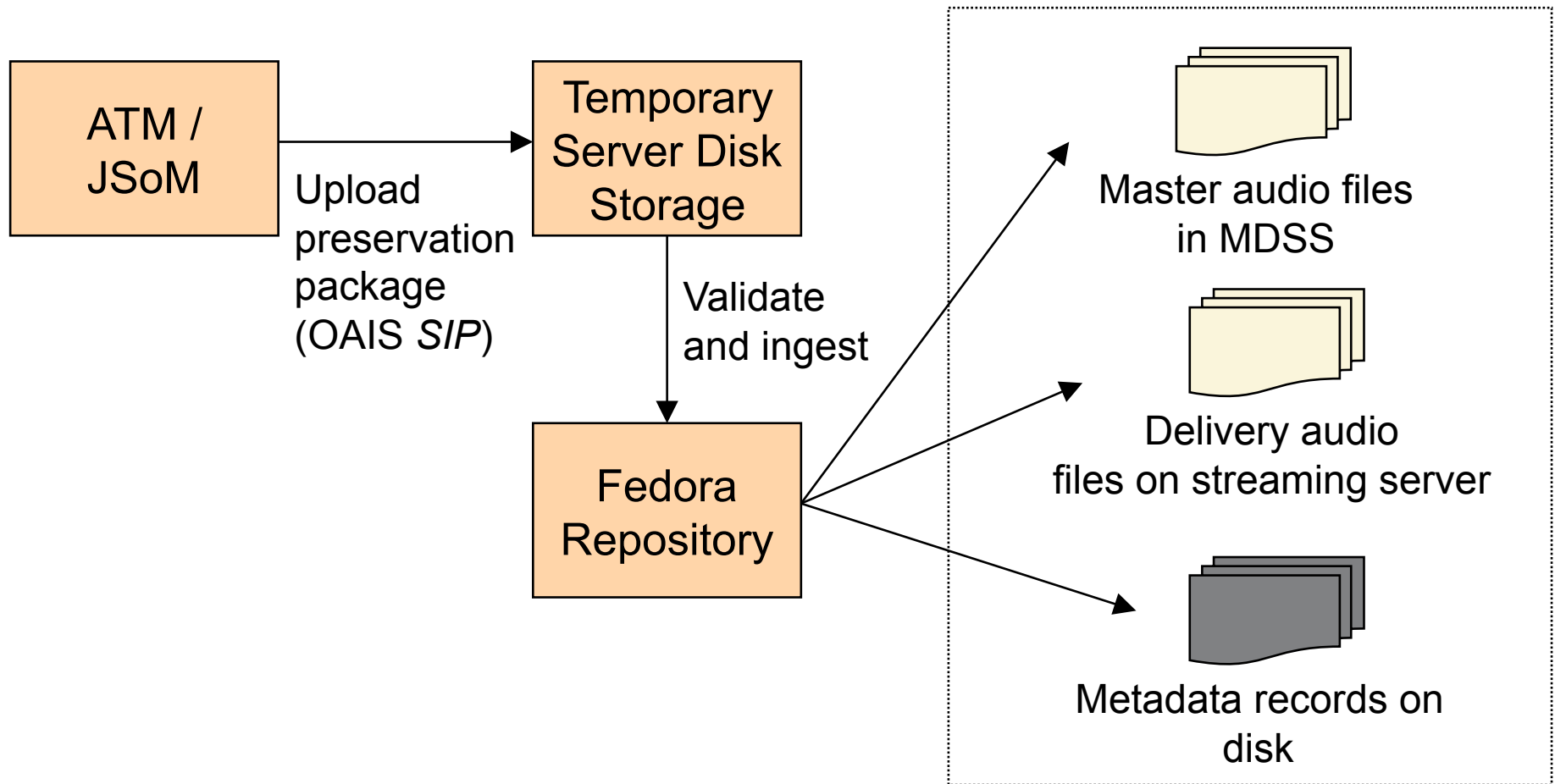


# Fedora

- **F**lexible **E**xtensible **D**igital **O**bject  
**R**epository **A**rchitecture
- Open source digital repository software developed by Cornell and the University of Virginia
- Supported by new organization:  
Fedora Commons
- Basis for IU Digital Library Repository
- Eventual backend for IUScholarWorks Repository and AIM as well



# Moving Audio into a Preservation Repository – Idealized Workflow



# Fedora at IU: Toward a Preservation Repository

- Need to add:
  - File integrity validation
  - Integration with MDSS – replication of data
  - Eventually, file format obsolescence monitoring and migration (for certain file types)
- Minimum requirements for file formats and metadata (descriptive, technical, digital provenance)
- Self-audit and/or external certification as Trusted Digital Repository
  - DRAMBORA, TRAC



# Collaborative Infrastructure

- HathiTrust
- LOCKSS/CLOCKSS
- Portico



# HathiTrust

- Repository for digitized books and journals from the CIC (and potentially other partners)
- Based on University of Michigan MBooks system; hardware at UMich and IU
- Supporting access and preservation
- Trusted Digital Repository certification:
  - Response to TRAC checklist
  - Undergoing DRAMBORA audit
- See <http://www.hathitrust.org/accountability>



# LOCKSS:

## Lots of Copies Keeps Stuff Safe

- Stanford-based peer-to-peer decentralized preservation infrastructure
- Harvests Web content via crawling
- Distributed copies compared against each other; damaged or incomplete copies are repaired automatically
- CLOCKSS: Joint venture between libraries (including IU) and publishers to preserve e-journal content using LOCKSS technology
- [www.lockss.org](http://www.lockss.org)



# Portico

- Also focused on archiving journal content
- Began as part of JSTOR
- Supported by libraries (including IU), publishers, and Mellon
- Centralized approach
- Publishers deposit content in PDF, XML, or SGML format
- [www.portico.org](http://www.portico.org)



# Questions?

- [skowalcz@indiana.edu](mailto:skowalcz@indiana.edu)
- [jwd@indiana.edu](mailto:jwd@indiana.edu)
- [www.dlib.indiana.edu](http://www.dlib.indiana.edu)

