The Data Capacitor and Digital Libraries at IU

Jon Dunn
Associate Director for Technology
IU Digital Library Program

February 22, 2006

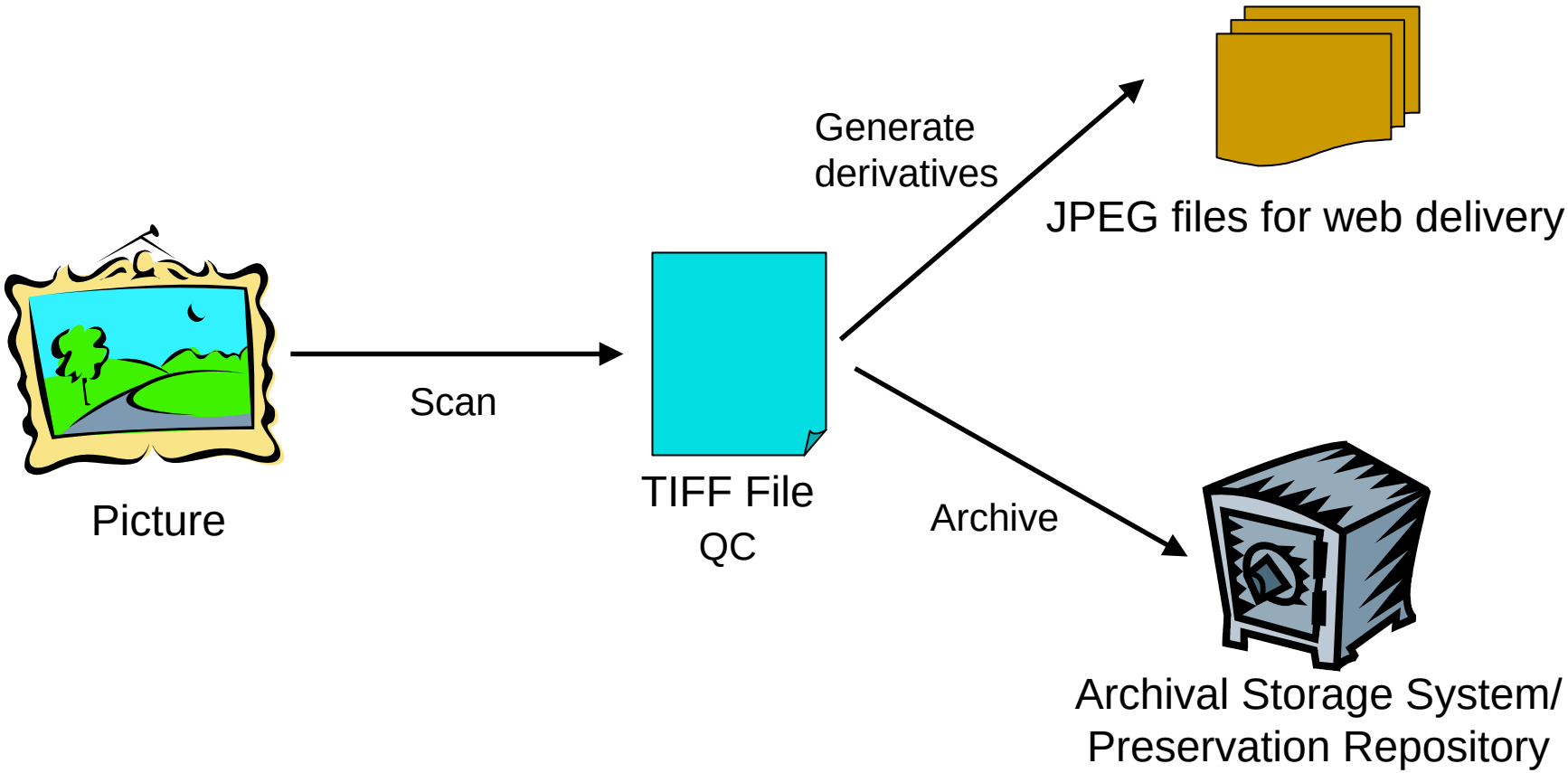
Scientific Data vs. Traditional Digital Library Content

- Scientific instruments creating data
  - Digitization creating digital files
- Post-processing of data
  - Quality control and derivative creation
- Archiving of data
  - Archiving
- Metadata
  - Metadata

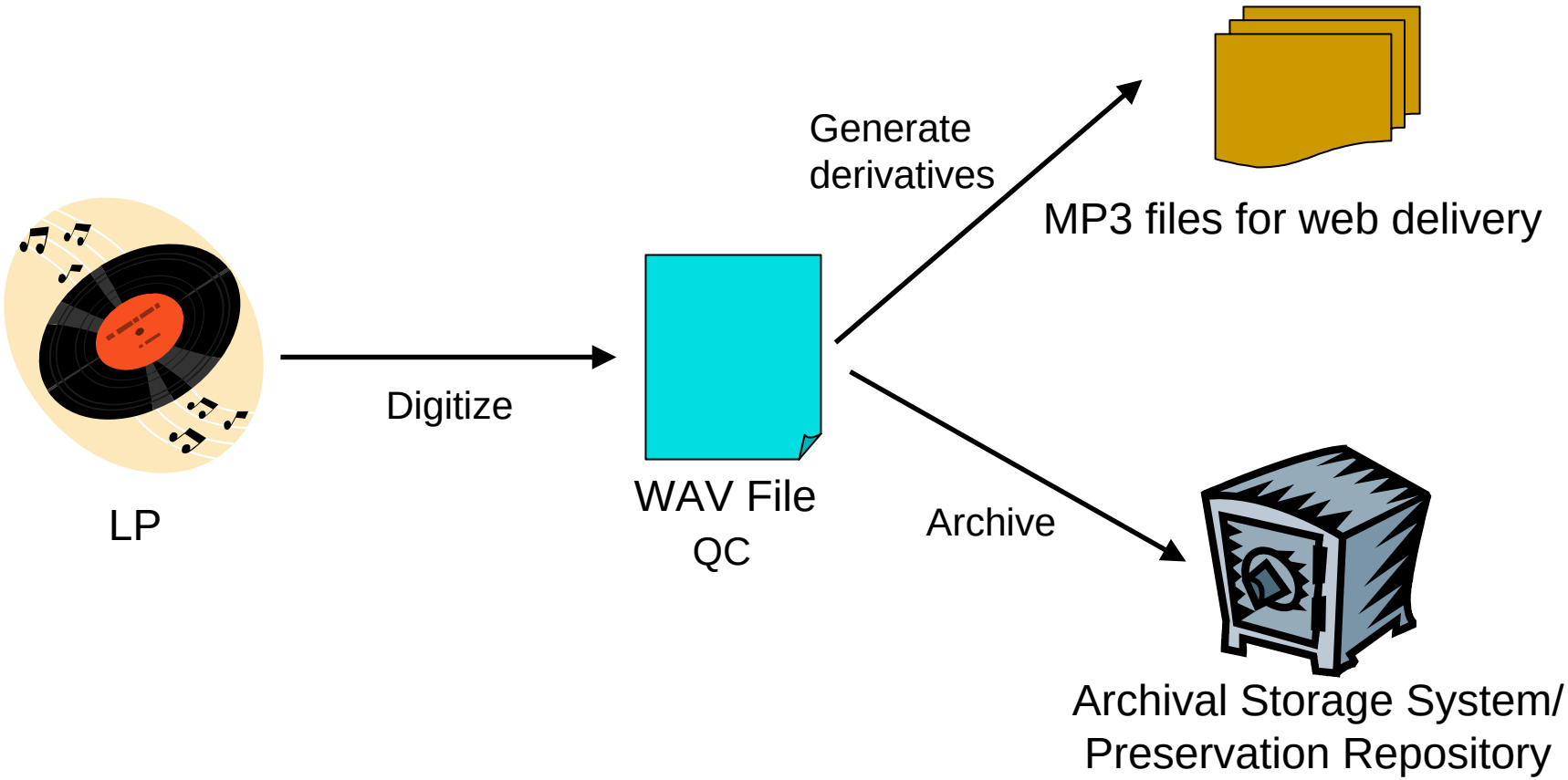
Steps in a Digital Library Ingestion Workflow

- Transfer
  - Digitization
- Quality control
- Processing
- Derivative creation
- Archiving

Digital Imaging Workflow



Audio Digitization Workflow



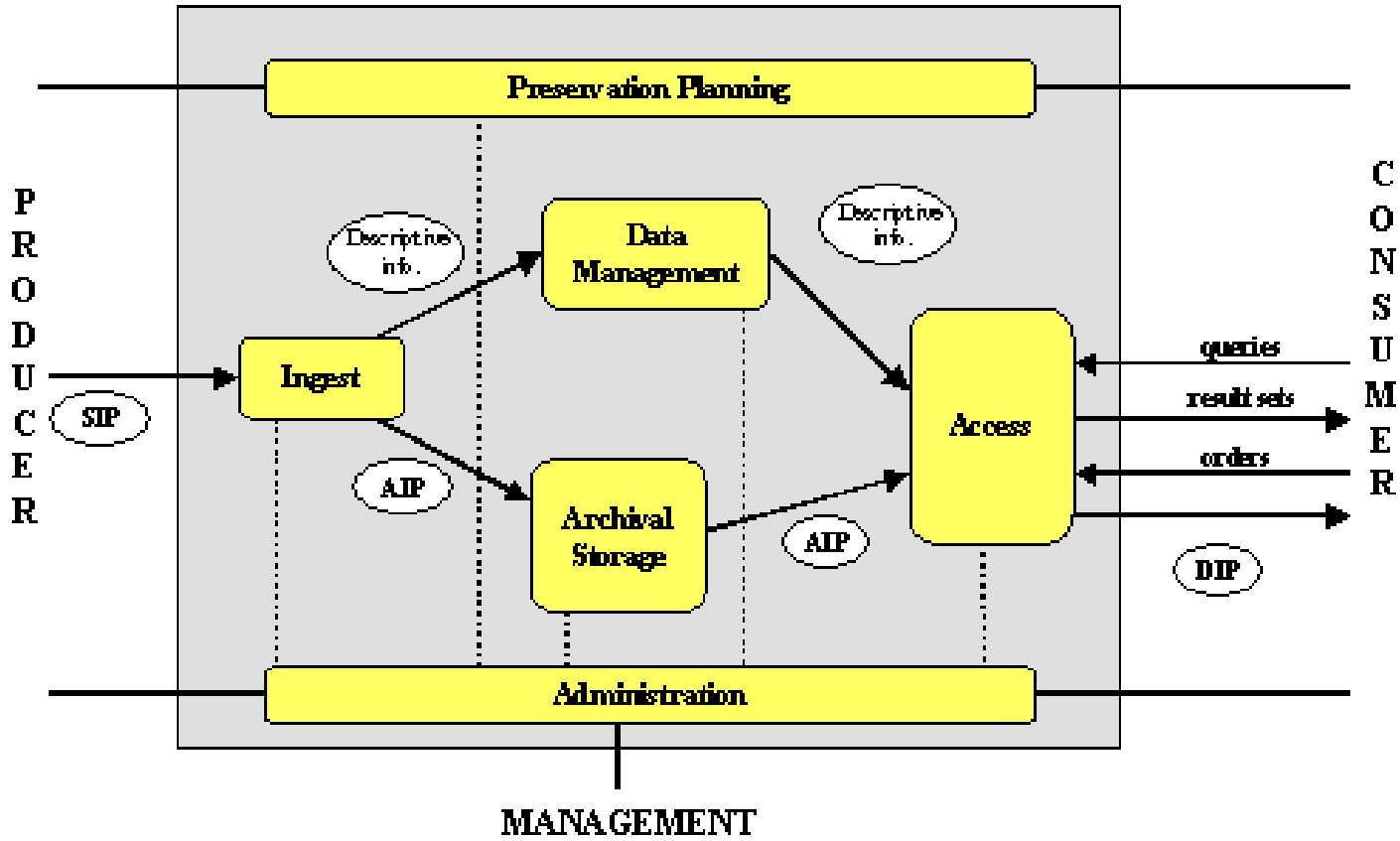
Digital Library Infrastructure Reengineering Project

- Began August 2005
- Funded for two years
- Implement a coherent technical infrastructure for digital libraries at IU
- Content storage – metadata storage – preservation – delivery – interoperability
- Using Fedora digital repository as a basis
  - www.fedora.info

OAIS: *Open Archival Information System*

- Conceptual framework for an archival system dedicated to preserving and maintaining access to digital information over the long term
- Origins in space science community
- Discusses interactions that *producers*, *consumers*, and *managers* have with a repository
- Basis for much current thinking on repositories in digital library community
  - OCLC/RLG *Trusted Digital Repositories: Attributes and Responsibilities*
  - RLG/NARA *Audit Checklist for Certifying Digital Repositories*

OAIS Reference Model



Digital Library Infrastructure: Three Kinds of Storage Needed

- Working storage
  - For quality control, post processing, derivative generation
  - Fast, convenient, secure network access from wide variety of platforms
  - Temporary
- Archival storage
  - Secure, long-term storage
  - OK if access is slow
  - Preservation integrity checking
  - Migration to handle media and file format obsolescence
- Delivery storage
  - Fast, Web accessible, secure (in some cases)

Digital Library Infrastructure: Three Kinds of Storage Needed

- Working storage
  - DLP servers?
  - Data capacitor?
- Archival storage
  - Massive Data Storage Service – MDSS?
- Delivery storage
  - DLP servers

Some Numbers…

Local sheet music scanning

- 20MB/page
  x 20 pages/hour
  x 8 hours/day

  = 3.2 GB/day

- Time to process: about 3 hours
- Not a problem

But… Vendor Outsourced Digitization

- Items shipped to vendor for digitization
- Images returned on CD-R, DVD-R, or portable hard drive

- Hypothetical sheet music example:

  5000 titles
  x 6 pages/title
  x 20MB/page
  = 600 GB at once

  - Time to process: 20-30 days

Phase One P45 Camera Back

- 39 megapixel digital back for Hasselblad medium format camera
- 39,052,992 pixels x 48 bits/pixel = 224MB/image
- 10 images/hour x 8 hours/day = 18 GB/day
- 30 images/hour x 8 hours/day = 54 GB/day

Video = Really Scary

- Uncompressed 4:4:4 video
  - 720 x 486 resolution = 349,920 pixels per frame
  - 349,920 pixels x 10 bits/sample x 3 samples/pixel = 10,497,600 bits per frame
  - 10,497,600 bits/frame X 29.97 frames/second = 314,613,072 bits per second
  - 314,613,072 bps x 3600 seconds = ~141.58 GB/hour
  - For 1920x1080 HDTV, more like 840 GB/hour

EVIA Digital Archive

- Ethnomusicological Video for Instruction and Analysis
- Video digitized at University of Michigan
- Concentrated period of production – couple of months

- 12 scholars contribute 10 hours/video each
- 12 scholars x 10 hours x 142 GB/hour = 17 terabytes

More Information…

- Data Capacitor
  - http://datacapacitor.org/
- Steve Simms
  - ssimms@indiana.edu
- Jon Dunn
  - jwd@indiana.edu