Reexamining Digital Library Infrastructure at IU

Jon Dunn
Indiana University Digital Library Program

IU Digital Library Brown Bag Series
November 30, 2005

Some IU Digital Library History

- 1995: LETRS – electronic text
- 1996: Variations, DIDO – audio, images
- 1997: Digital Library Program

Digital Library Content Types at IU

- Books
- Manuscripts
- Photographs
- Art images
- Music audio
- Video
- Sheet music
- Musical score images
- Music notation files
- …and more

Current DLP Technical Environment

- Variety of access systems
  - DLXS (University of Michigan)
    - Text
    - Finding Aids
    - Bibliographic information
  - Locally-developed systems
    - Cushman Photograph Collection
    - DIDO: Digital Images Delivered Online
    - Variations/Variations2
    - Page turners (sheet music, METS Navigator)

Current DLP Technical Environment

- Variety of storage systems
  - Local DLP servers
  - DLP Tivoli Storage Manager
  - IU Massive Data Storage System (HPSS)
- No *repository*

What is a digital library repository?

- A system (hardware and software) in which to deposit digital objects (files and metadata) for purposes of access and/or long-term storage.

Repository Purposes

- Access
  - Web access to digital files and metadata
  - Services/applications for searching, browsing, transformation, etc.
- Preservation
  - Secure storage for digital files and metadata
  - Services for file integrity checking (using checksums), migration, conversion, etc.
- Some repositories are single-purpose; some are dual-purpose

Not a New Model…

- Digital Repository
  - Common system for storing, managing, and providing access to digital content and metadata
- Integrated Library System
  - Common system for storing, managing, and providing access to MARC cataloging records

Why do we need a repository?

- Isn’t what we have good enough?
  - Web servers
  - File servers
  - Databases
  - Mass storage systems

Mass Storage Systems

- High-capacity, high-performance data storage
- Hardware
  - Servers
  - Automated tape libraries, e.g. IBM, Storagetek
  - Spinning disk
- Software
  - HSM: hierarchical storage management
  - IU uses HPSS (High Performance Storage System) from IBM


IMATION

Mass Storage Systems

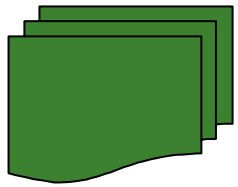
- Typical features
  - Bit-level storage and retrieval of files
  - Security: authentication, authorization
  - Mirroring of data between sites over a network
  - Migration of files to new media types
- Is that enough for digital preservation?

Data Persistence

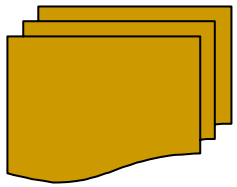
- Key is ***migration***
- Keeping the bits alive
  - Physical media
  - Logical media format
- Keeping the bits understandable
  - File format
  - Metadata
- Digital data must be *actively* managed
- Small “pockets” of digital content pose a problem for migration

Digital Objects: More than just files

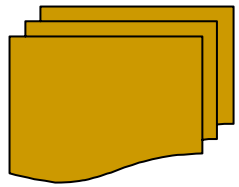
Example: Electronic Book



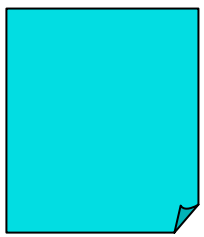
Metadata



Delivery page image files (JPEG)



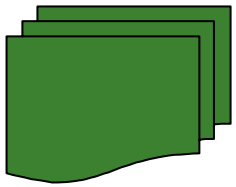
Hi-res page image files (TIFF)



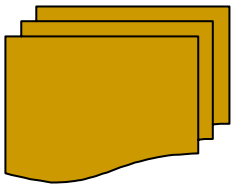
Text transcription (TEI/XML)

Digital Objects: More than just files

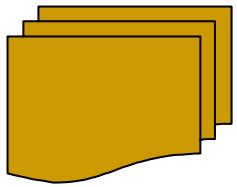
Example: Sound Recording



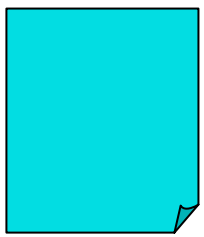
Metadata



Delivery audio files (MP3 or other)



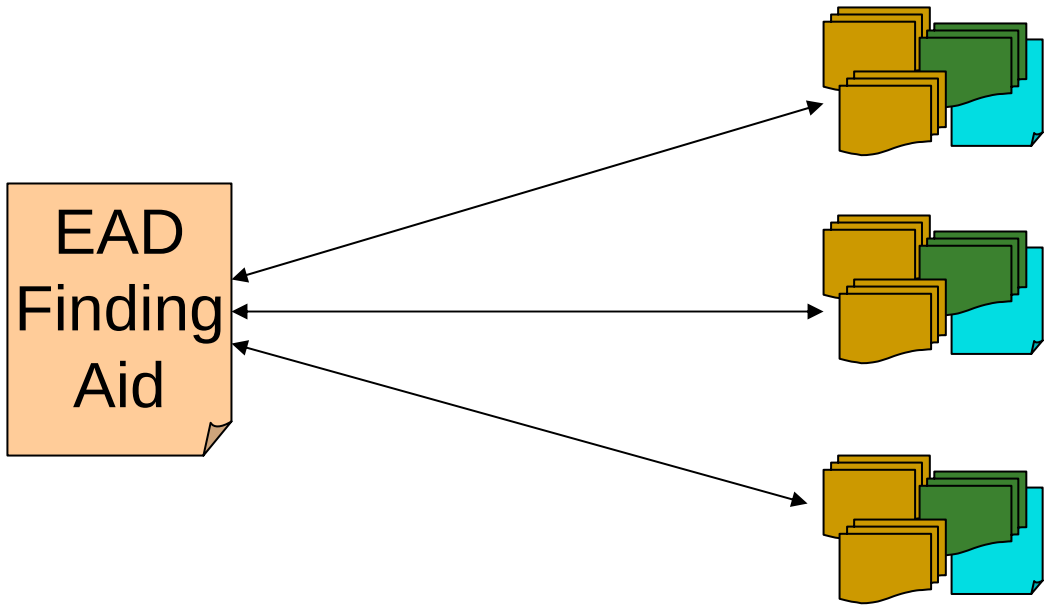
Hi-res audio files (Broadcast WAVE)



Images of labels, jacket, box, etc.

Digital Objects: More than just files

Example: Archival Collection



DL Objects

- Digital library “objects” have many parts
  - Metadata
    - Descriptive, administrative, structural, preservation, …
  - Preservation/archival files (several)
  - Delivery files (several)
  - Persistent identifier
- How do we keep them connected and organized?
  - Now: Good practice in file naming, directory organization, project documentation -not scalable!
  - Future: Digital object repository

A Word About Metadata

- Descriptive
  - Used for discovery and identification
- Technical
  - Technical characteristics of the object and its components
  - Used for preservation and for delivery
- Digital Provenance
  - How an object got to be what it is today
- Structural
  - How the parts of an object relate to each other

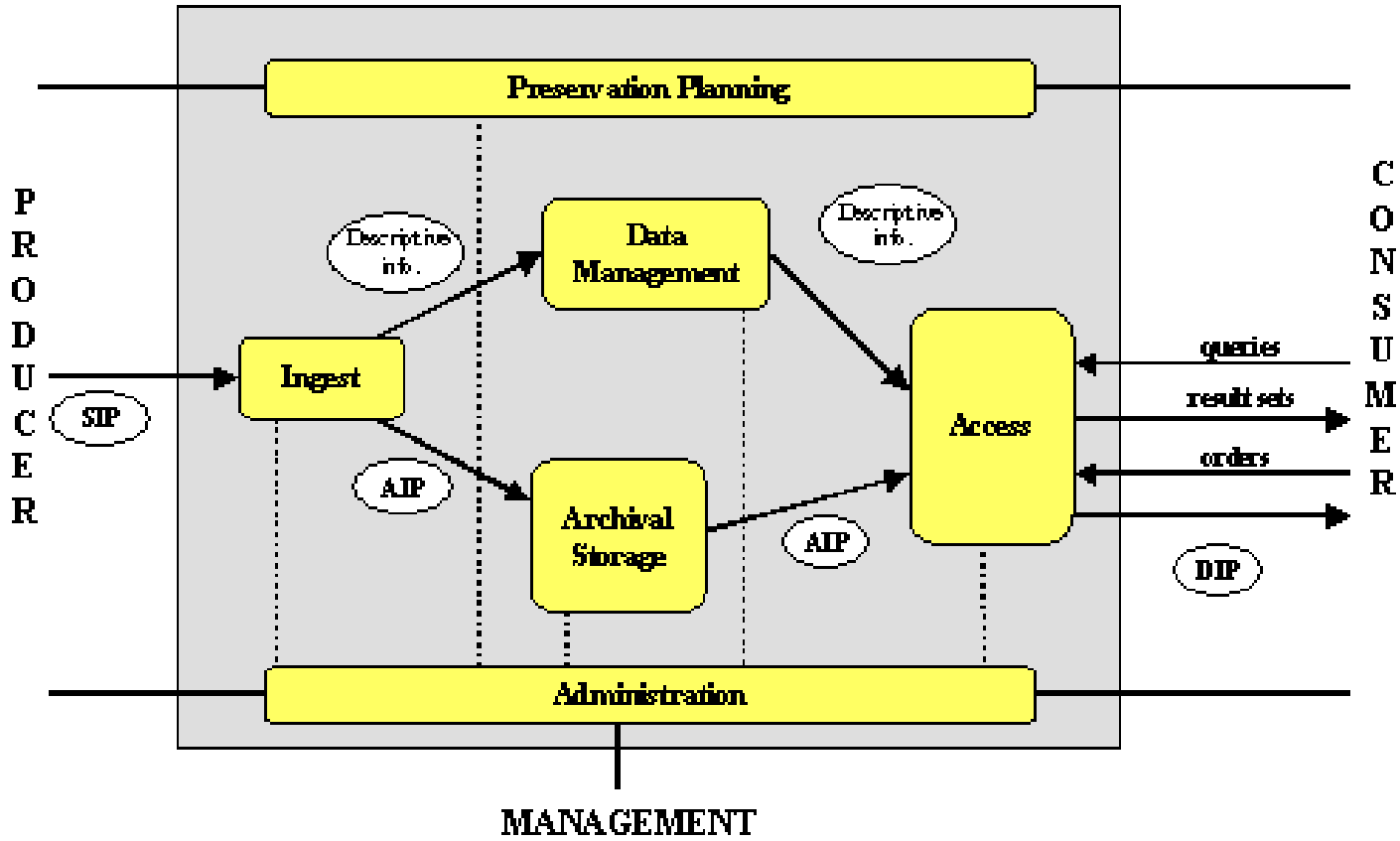
Some Relevant Metadata Standards

- Descriptive
  - MARC, MARCXML, Dublin Core, MODS, VRA Core, EAD
- Technical
  - MIX, PREMIS
- Digital Provenance
  - PMD, PREMIS
- Structural
  - METS, MPEG-7, MPEG-21

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*

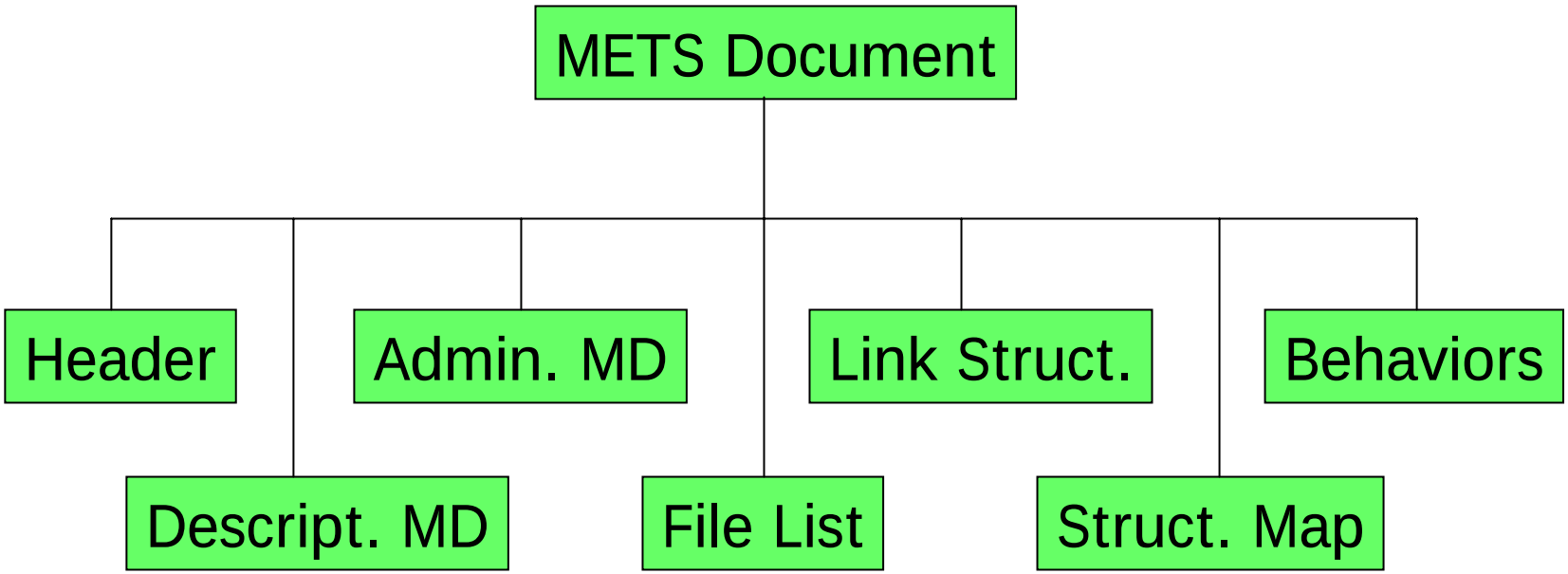
OAIS Reference Model



Object Packaging Standards: Content and Metadata

- Functions in OAIS model
  - Submission Information Package (SIP)
  - Archival Information Package (AIP)
  - Dissemination Information Package (DIP)
- Two main competitors
  - METS
    - Metadata Encoding and Transmission Standard
  - MPEG-21 DIDL
    - Digital Item Declaration Language

METS


METS Document
Header
Admin. MD
Link Struct.
Behaviors
Descript. MD
File List
Struct. Map

Digital Object Repository Software Platforms

- Commercial digital asset management / content management / document management systems
  - e.g. IBM Content Manager, Artesia TEAMS, FileNet, Documentum
- Open source systems
  - e.g. Fedora (University of Virginia and Cornell)
- Homegrown systems
  - e.g. Harvard, California Digital Library
- Commercial services
  - e.g. OCLC Digital Archive

“Digital Repository” vs. “Institutional Repository”

- Digital repository
  - Common storage for digital content and metadata
  - Basic infrastructure component: “plumbing”
  - e.g. Fedora
- Institutional repository
  - Often implies focus on one application: institutional content, research output
  - e.g. MIT DSpace:
    - “capture, store, index, preserve, and redistribute *the intellectual output of a university’s research faculty* in digital formats”

Motivation for a Digital Repository at IU

- Many pockets of digital content and metadata
- Difficult to sustain
  - Variable tech support, replacement funding
  - Harder to preserve, migrate data forward to new software and hardware
  - Harder to budget for
- Difficult to build common services and applications
  - Cross-collection search
  - Standard interfaces for viewing and playing content
  - Interfaces to course management and other IT services
  - OAI data providers
  - Preservation services (integrity checks, etc.)

Questions In Repository Planning at IU

- Scope
  - Just library?
  - Museums and archives?
  - All campuses?
  - Other digital content
    - Instructional (e.g. faculty materials in OnCourse)
    - Business (PR, Athletics, etc.)
- Funding model
- Standards
  - Minimum requirements for content formats and metadata
- Tools/services/applications
  - What else is needed to make a repository useful/usable for preservation and access?

Repository Evaluation Criteria

- Flexibility
  - Not a rigid data model
  - Support for many media types, complex digital objects
  - Not locked into one technology platform (OS, database)
- Extensibility
  - Use of modern technologies
  - Easy integration with other systems/tools
  - Means of extension/modification
  - Support for DL standards, particularly metadata
- Sustainability
- Supportability
- Usability
- Cost

Fedora

- FEDORA
  - Flexible
  - Extensible
  - Digital
  - Object and
  - Repository
  - Architecture

Fedora - Background

- Began as CS research project at Cornell – 1997-98
  - Architecture
  - Reference implementation
- UVa Libraries became interested – 2000
  - Trying to create a DL architecture
  - No commercial solutions found
- Mellon-funded project – 2001-2003
  - Joint UVa/Cornell project
  - Update technologies
  - Make use of relational database
  - Make more production-ready
  - IU member of “deployment group” engaged in testing

Fedora - Technical Environment

- Open Source software
- Written in Java
- OS Platforms:
  - Windows
  - Linux / Unix
  - Mac OS X (not yet officially supported)
- Database support:
  - MySQL
  - McKoi
  - Oracle8i , Oracle9i

What does Fedora do?

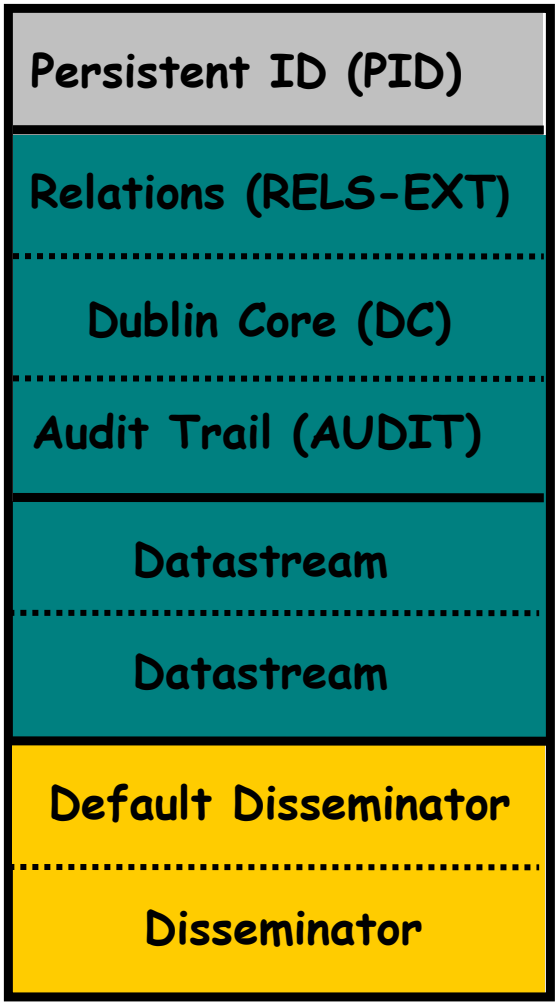
- Manages files or references to files that make up digital objects
- Manages associations between objects and interfaces
- Invokes behaviors of objects

What does Fedora not do (yet)?

- Searching/browsing of metadata and content
- End-user UI for display/navigation of metadata and content
- Cataloging tools
- Preservation services
- …

- Fedora is DL “plumbing”… Not an out-of-the-box complete DL system

Fedora Object Model



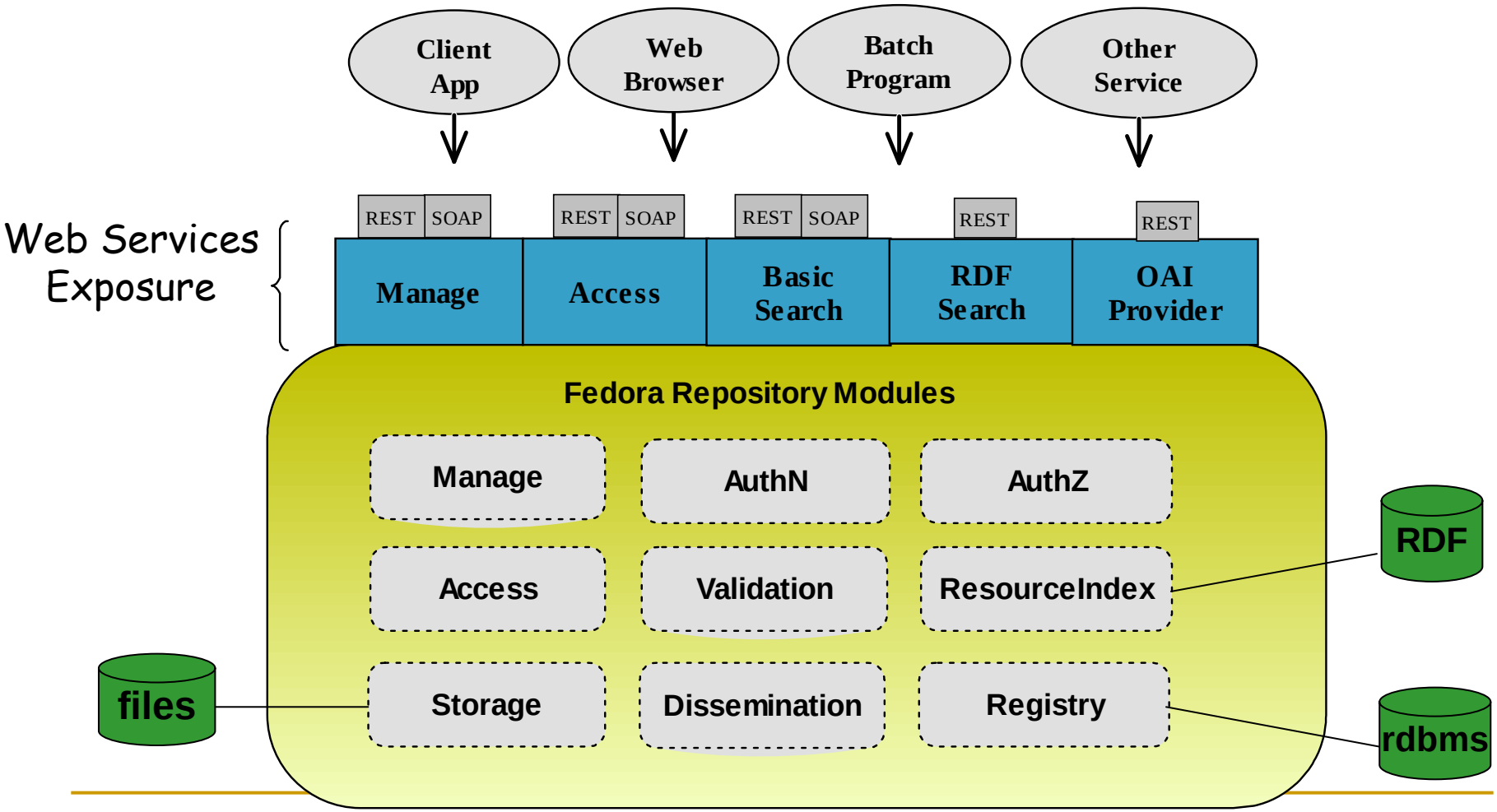
Digital object identifier

Reserved Datastreams
Key object metadata

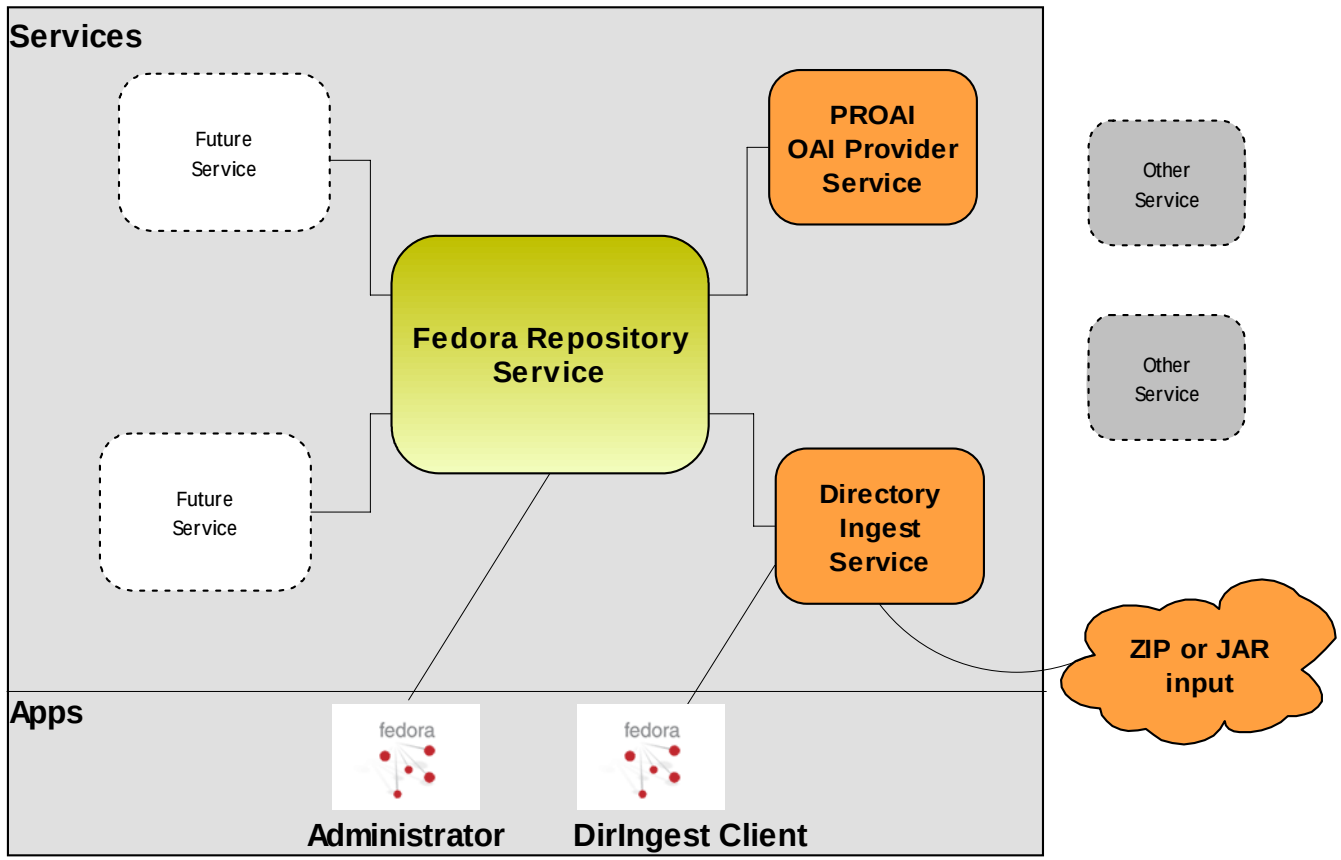
Datastreams
Aggregate content or metadata items

Disseminators
Pointers to service definitions to provide service-mediated views

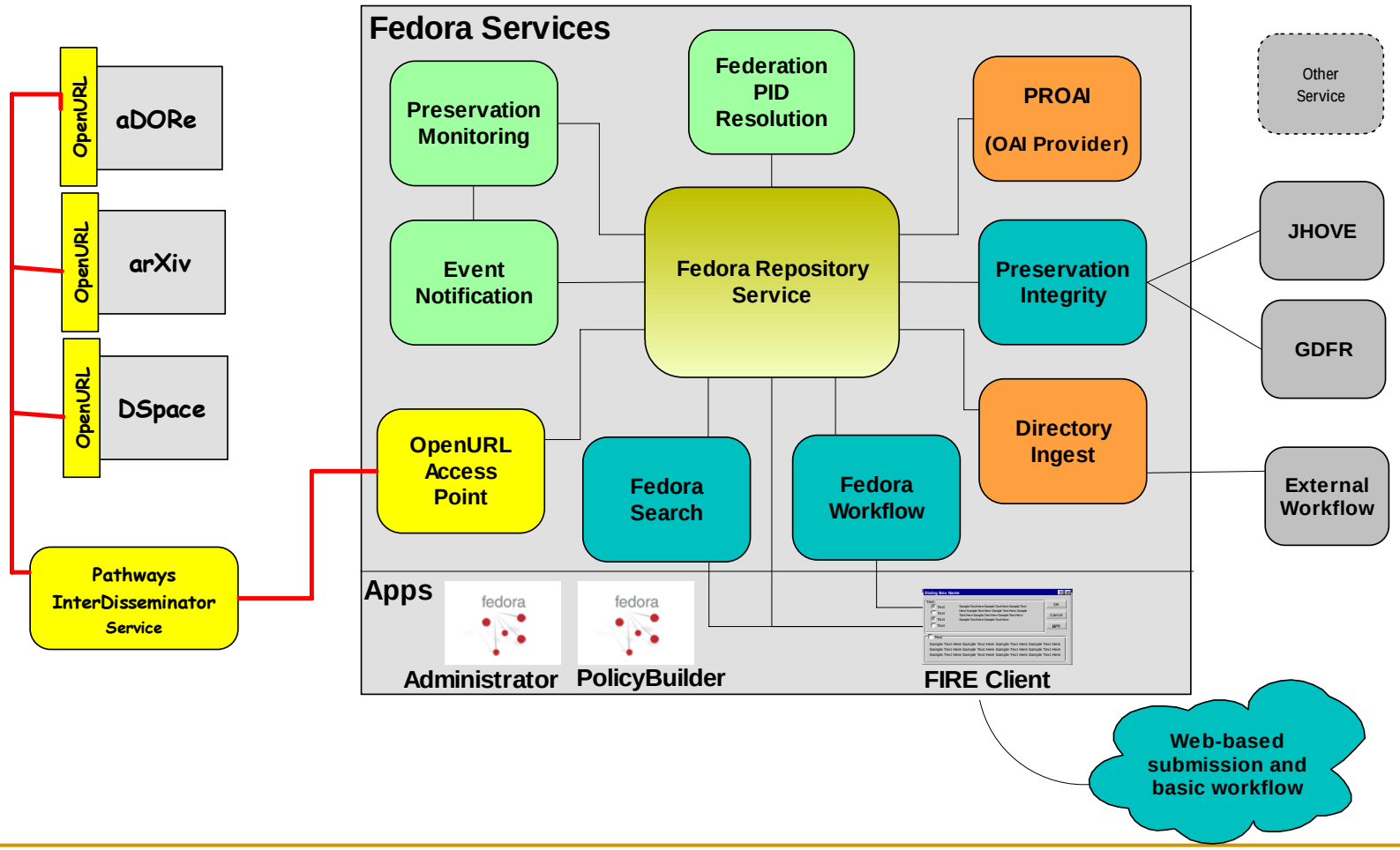
Fedora Repository and Web Services



Fedora Service Framework (Fedora 2.1)



Fedora Service Framework (2005-07)



Content models

- A content model describes the internal structure of a class of Fedora objects
  - Number & type of datastreams
  - Number & type of disseminators
- Benefits of a content model
  - A method to describe the structure of similar Fedora objects
  - Facilitate the creation of “batches” of objects
  - Standardize handling of Fedora objects by tools outside the repository

Content model goals

- Maintain consistency with other Fedora users
- Standardize disseminators across objects, shifting the implementation to suit the needs of the collection
  - Makes it easier to build collection-independent applications on top of Fedora
  - It's possible to change implementations behind the scenes (JPEG2000?)
- Maintain functionality of existing collections

Standard disseminators

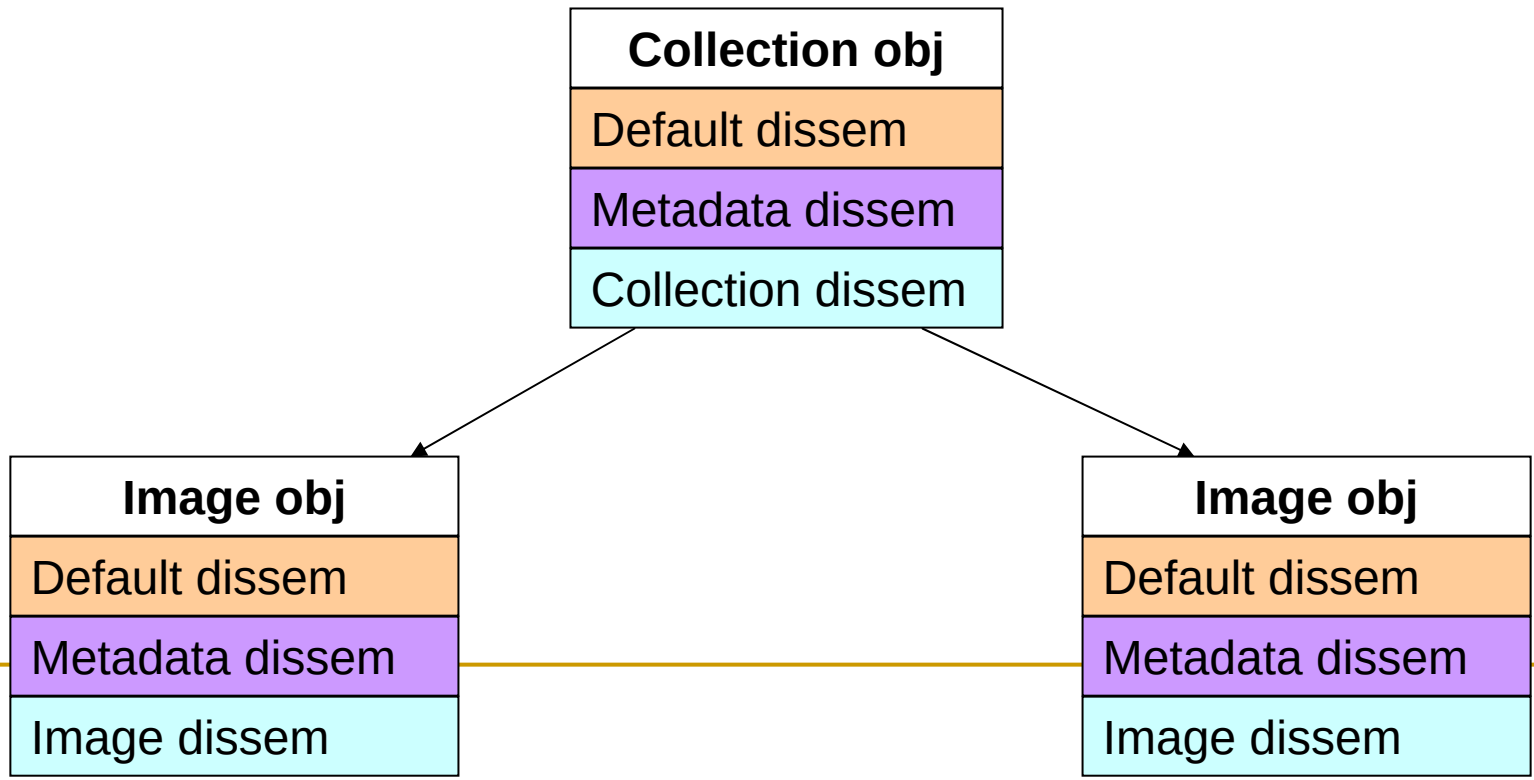
- All objects implement the default disseminator
- Most objects implement the metadata disseminator
- Most objects implement type-specific disseminators

Default dissem
getLabel
getDefaultContent
getPreview
getFullView

Metadata dissem
getDC
getMetadata(type)

Content model for simple images

- Each image is a single Fedora object
- Images are available in a variety of sizes
- Each image belongs to a collection



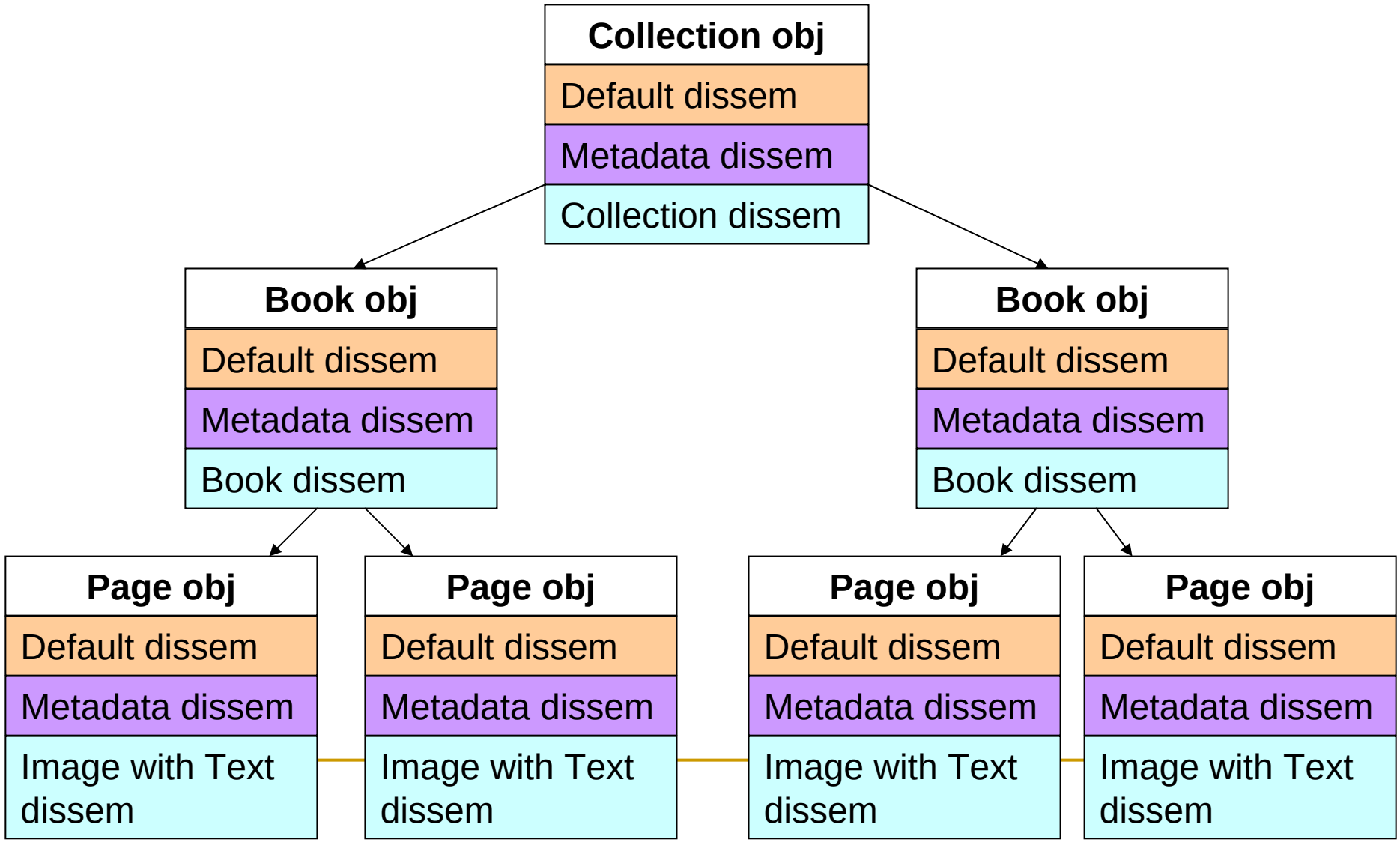
Handling metadata

- All metadata is stored in a single datastream
- All metadata is wrapped in a METS document
- Authoritative metadata is stored at the “natural location”
- Derived metadata may be stored elsewhere for technical reasons

Fedora Demos

- Hohenberger collection
- IU test server (Fedora native interface)
  - Horseshoe players
  - Hohenberger collection
- Fedora at Tufts

More complex models



Infrastructure Project Progress

- New staff hired with support from UITS
- Scope defined
  - Start with IUB Libraries
- Fedora selected as repository
- Initial planning work on DIDO2 started
  - Evaluation of tools
- Content modeling work begun
- Test import of some existing image collections

Infrastructure Project: Next Steps

- Finalize project sequencing
  - DIDO2
  - Documentary photography
  - Multi-page image objects
  - TEI text
- Define content, metadata standards
- Define and implement tools
  - Validation/loading/“ingestion”
  - Cataloging/metadata creation
  - Searching/browsing/discovery
  - Use
- Ongoing process

Infrastructure Project Challenges

- Time and resources vs. scope of work
- Sorting out old collections – digital archeology
- Implementing new infrastructure while continuing to do new projects
- Generalization
- Metadata entry / cataloging tool design
- Integration with MDSS/HPSS

Thank You!

- Contact info:
  - Jon Dunn jwd@indiana.edu
  - Ryan Scherle rscherle@indiana.edu
  - Eric Peters erpeters@indiana.edu

- Thanks to the Fedora project for diagrams.