NATIONAL ENDOWMENT FOR THE HUMANITIES AND *The* LIBRARY *of* CONGRESS

# The U.S. National Digital Newspaper Program

## *Thinking Ahead, Designing Now*


Helen Agüera (program development)

**US National Endowment for the Humanities**

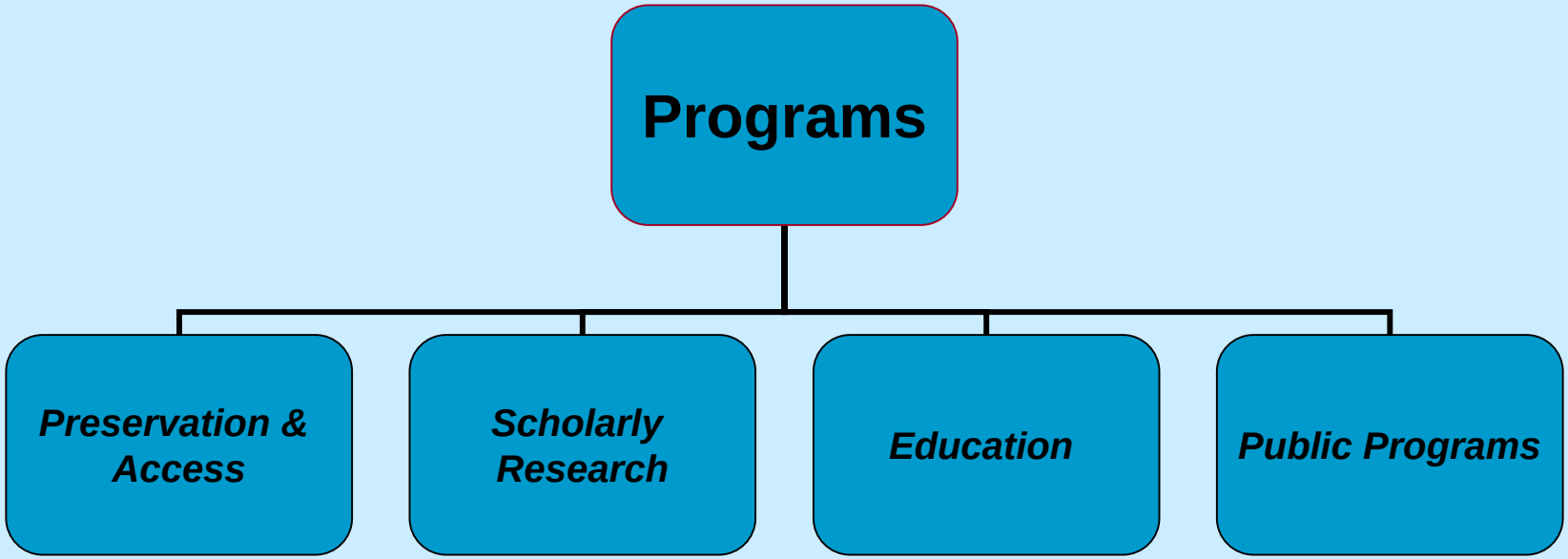
Mark Sweeney (preservation planning)

Ray Murray (technical specifications)

George Schlukbier (repository development)

**US Library of Congress**

# National Endowment for the Humanities

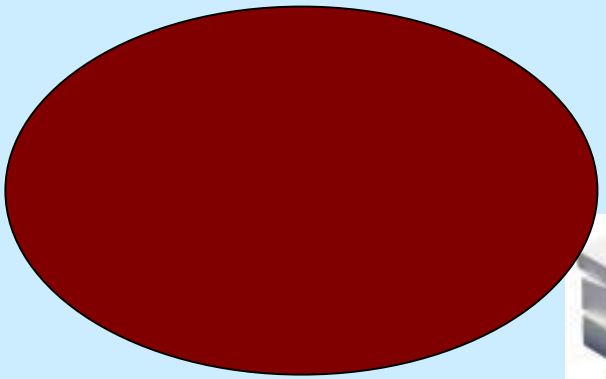


# United States Newspaper Program (USNP)

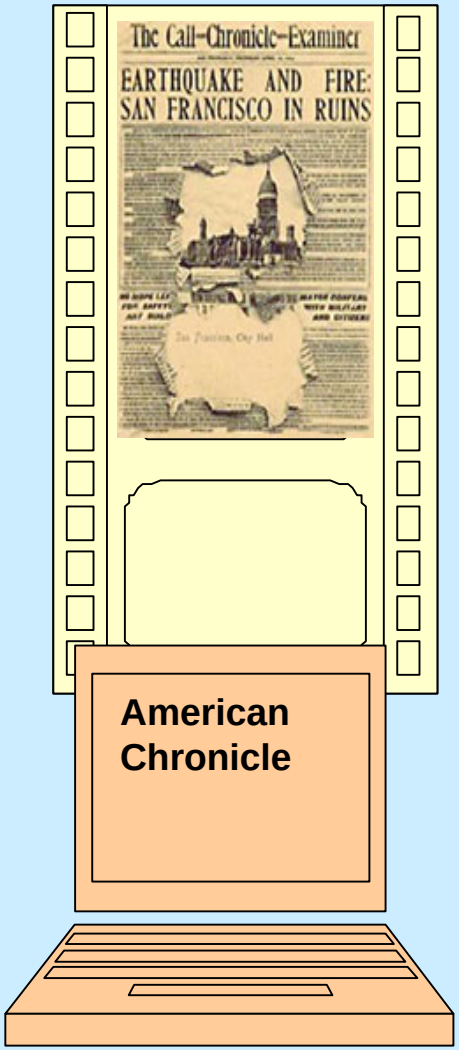
- Started in the 1980s - expected to end in 2007 or shortly after.
- Grants to state projects to inventory and catalog all newspaper holdings within a state.
- Grants to do selected preservation microfilming.
- Partnership between NEH and LC.

# USNP Accomplishments


- Bibliographic records for over 140,000 newspaper titles.
- Access to 70 million pages of newsprint in microfilm.
- NEH funding totaling over $54 million.



# National Digital Newspaper Program (NDNP)




- Provide enhanced access to newspapers, building on USNP.
- Distributed effort involving state projects applying a uniform selection criteria to achieve geographic representation.
- Partnership between NEH and LC.
- NEH will make awards to state projects to convert newspaper microfilm into digital files.
- LC will develop and maintain **American Chronicle** to make digitized papers freely accessible.
- State projects repurpose information for local purposes.

# NDNP Features

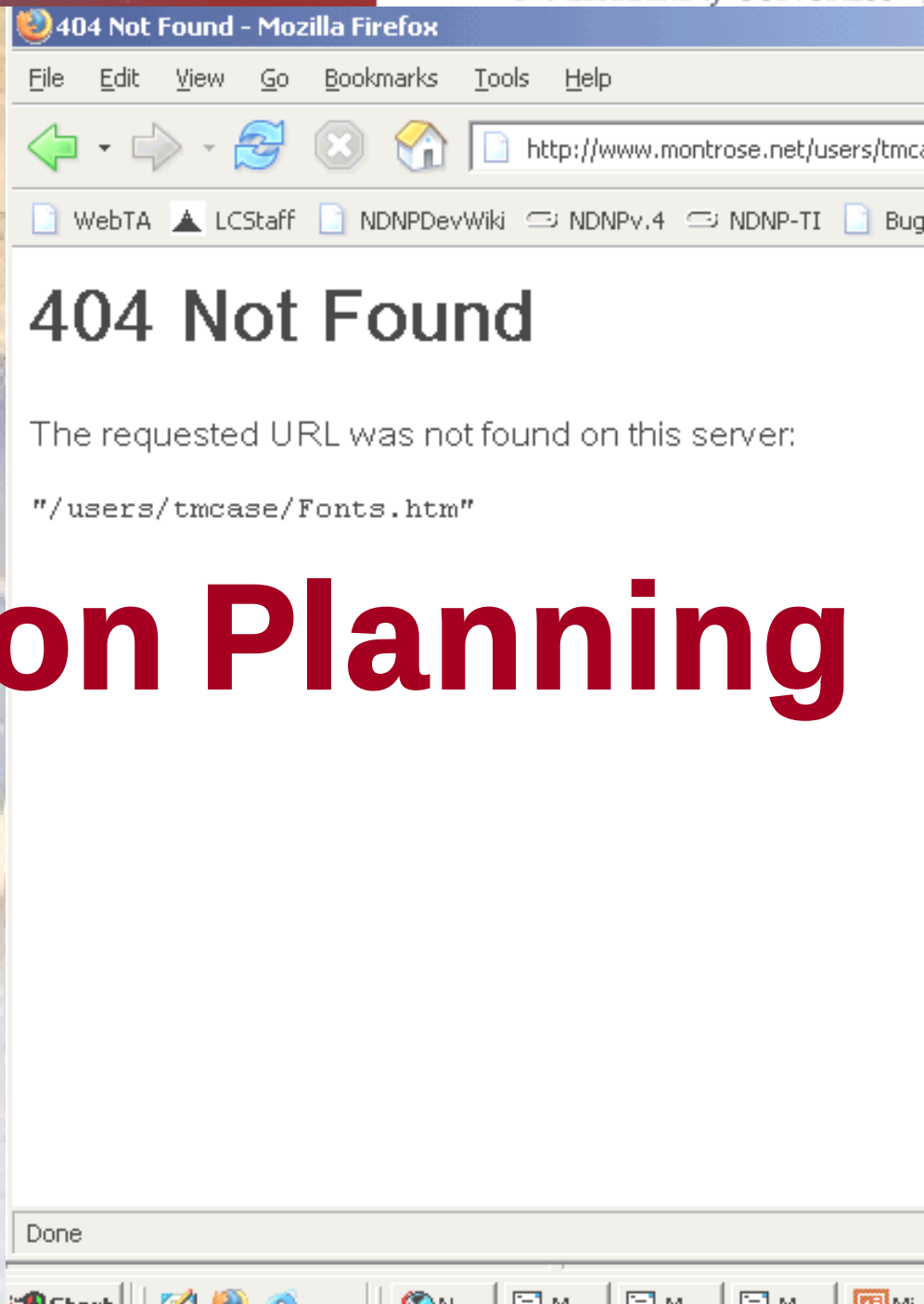
- Focuses on historical newspapers published from 1836 to 1922, the end of public domain.
- Complements other digital resources for earlier historical period.
- Proceeds in phases to encompass millions of pages created by over 50 projects.
- Begins chronological coverage with early 20th century and expands to earlier decades to achieve broad geographical representation.
- Repurposes USNP bibliographic information for uses to locate newspapers in analog formats.

# NDNP: Current Work

- Development phase began in May 2005.
- Six projects selecting titles and digitizing a minimum of 100,000 pages published in California, Florida, Kentucky, New York, Utah, and Virginia from 1900 to 1910.
- LC contributes titles from its collection, aggregates all information, and creates a preservation framework.
- A prototype is launched in September 2006 and the test bed results are evaluated by all partners.

# NDNP: Future Directions

- Awards to state projects with partners that have access to negative microfilm and digital infrastructure.
- An advisory board assists in selecting titles by applying general selection criteria.
- Special consideration is given to “orphan” titles.
- NEH awards cover the costs of selection, digitization, and delivery of information to LC.
- Next deadline is November 1, 2006.




# Preservation Planning

# Guiding Principles

- Aggregate, serve, and preserve
- Consistent with missions and philosophies of NEH and LC
  - Open and perpetual access to the general public and scholarly community
  - Take care to preserve the asset that NDNP builds
  - Must demonstrate good use of taxpayer $
- Phased development
  - Build incrementally – don't close off options

# Open in Many Ways

- We envisage a "system" that is open in many senses
  - freely accessible (a public resource)
  - available to use and re-use
  - deep linking and persistent identification to support citation
  - open technical formats
  - interoperable through support for standard protocols
  - modular architecture
  - software based on open source code to degree possible.

# What is certain? - CHANGE

- Technology
  - Access capabilities will improve and get cheaper
  - OCR will improve
  - Accuracy of automated article segmentation will improve
  - Availability of open source tools will expand
- User expectations
  - What will scholars want?
    - examples: Text mining, time and place analysis
  - What will new user communities want?
    - PDA access, learning tool integration
- Preservation models
  - Openness, trust, tools, stability

# Experience Informing the Future

How can we plan for the Future?

- Content is more important than today's "system"
- Design "system" to be expandable
  - Modular and upgradeable
- Assume interoperability is a requirement
  - A resource that stands alone but plays well with others
- Explicit incorporation of development phase
  - Opportunity for learning
  - Validation of assumptions
  - Develop best practices (perhaps leading to standards)
  - Build corpus that is of value for technical experimentation

# Practical Concerns

- Out-of-the-box solutions have preservation challenges
- Analyze technical options
  - Think carefully about formats
  - Detail specifications to assure consistency
  - Develop means to validate conformance
  - Incorporate metadata to understand context and circumstances of creation
- Build on LC expertise
  - METS, PREMIS, FEDORA, JPEG2000, large-scale and distributed digitization and aggregation, *Stars and Stripes*, NDIIPP
- Expect to learn from awardees

# Specifications

# Archival Needs, Data Needs

- Open Archival Information System (OAIS) Model
- Archive interacts with:
  - Producers, Managers, Consumers
- Archive needs to:
  - Ingest, Manage, Distribute
- OAIS vehicles:
  - Submission Information Package (SIP)
  - Archival Information Package (AIP)
  - Distribution Information Package (DIP)

# Information Object, Data Object

- Information object: original newspaper, microfilm
- Data Object is the digital surrogate:
  - TIFF
  - JP2
  - PDF
  - Optical Character Recognition (OCR) text
  - Structural Metadata
  - Preservation Metadata
  - Need to know the complete data object valid

# Archival Master: TIFF

- Conforms with TIFF 6.0
- 8-bit grayscale
- 400 dpi preferred
- Uncompressed
- Only deskewing should be applied
- Cropped to page edge
- Additional TIFF tags required

# Production Master: JPEG 2000

- Conforms with JPEG 2000, Part 1 (.jp2)
- Use 9-7 irreversible (lossy) filter
- Compressed to 1/8 of the TIFF or 1 bit/pixel
- Tiling, but no precincts
- RDF/Dublin Core metadata in XML box
- Profile prepared with assistance of Rob Buckley, Xerox Labs

# Derivative: PDF

- Compatible with Acrobat 5.0 (PDF 1.4)
- Image with text behind
- Image will be a grayscale, 150dpi JPEG, using a medium (or 40) quality setting
- XMP/RDF/Dublin Core metadata

# OCR text: ALTO

- Conforms with ALTO (Analyzed Layout and Text Object) schema
- ALTO is product of EU-funded METAe project
- Mapping of OCRed text to image coordinates

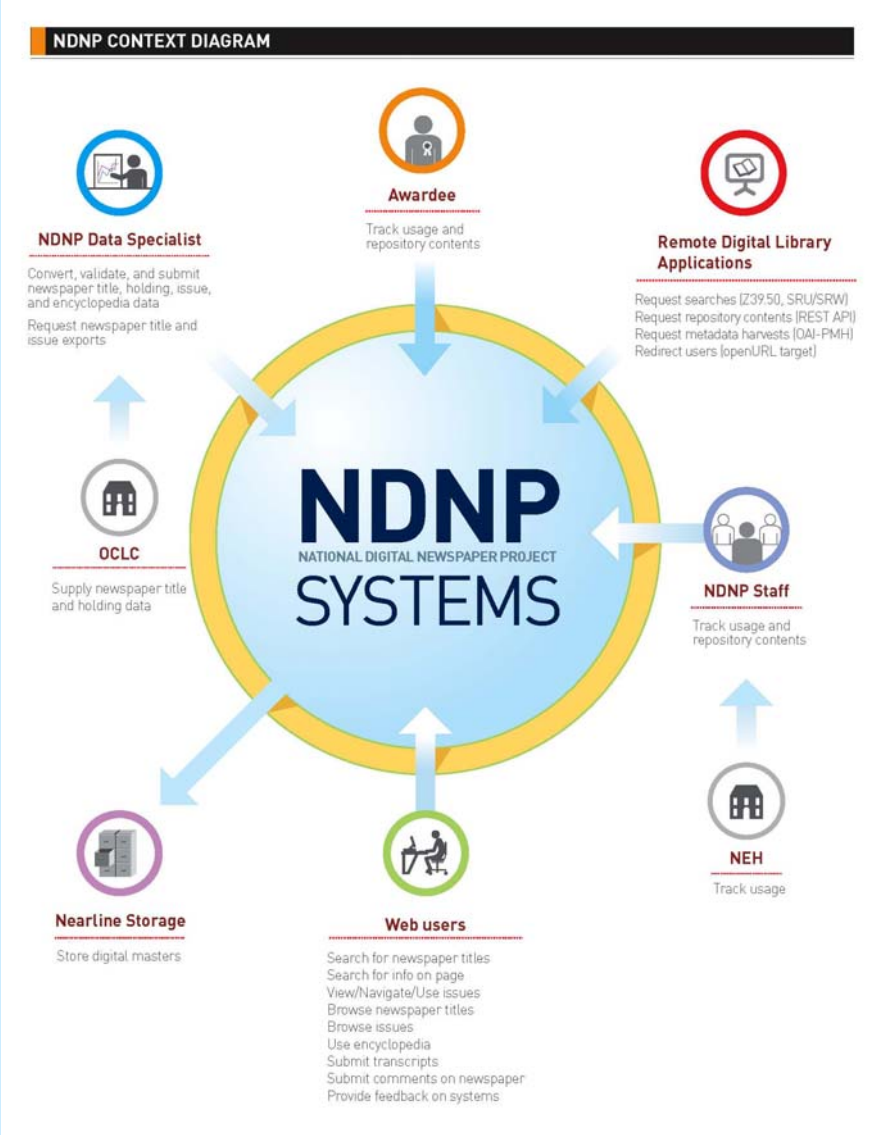
# Structural Metadata

- Metadata Encoding and Transmission Standard (METS)
- Title METS Object
  - Bibliographic and holdings data
  - Corrections, additions (essay, geographic coverage)
- Issue METS Object
  - Issue & page data
- Reel METS Object
  - Reel data & Technical target data

# Validating the SIP

- Digital Viewer and Validator (DVV)
  - Java library for validating SIPs: batches of data
  - Built on JHOVE validation, extends capabilities
  - Can be run from command line or within its GUI viewer interface
  - Digitally signs files as having passed validation
  - Adds preservation metadata to METS
  - Valid SIP may then be ingested

# Architecture

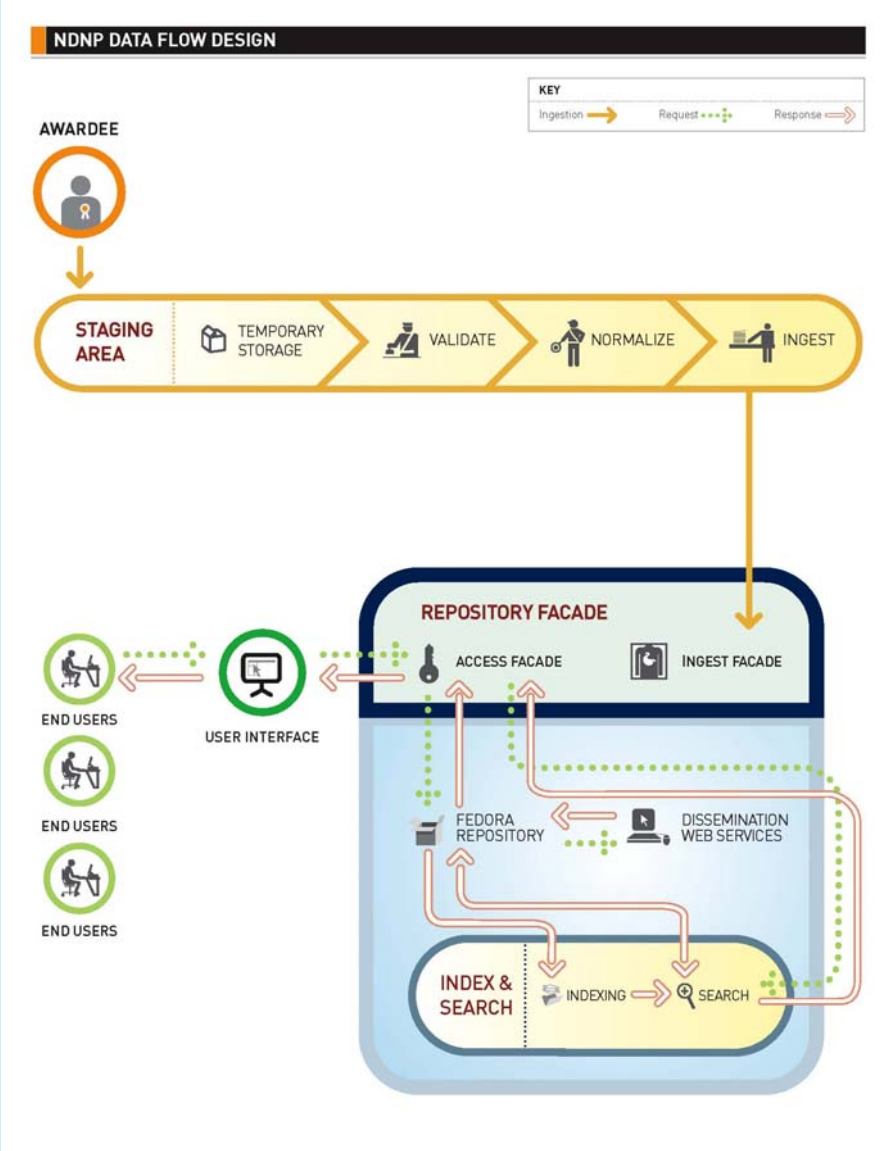

NDNP CONTEXT DIAGRAM
Awardee
Track usage and repository contents
NDNP Data Specialist
Convert, validate, and submit newspaper title, holding, issue, and encyclopedia data
Request newspaper title and issue exports
Remote Digital Library Applications
Request searches (Z39.50, SRU/SRW)
Request repository contents (REST API)
Request metadata harvests (OAI-PMH)
Redirect users (openURL target)
NDNP
NATIONAL DIGITAL NEWSPAPER PROJECT
SYSTEMS
OCLC
Supply newspaper title and holding data
NDNP Staff
Track usage and repository contents
NEH
Track usage
Nearline Storage
Store digital masters
Web users
Search for newspaper titles
Search for info on page
View/Navigate/Use issues
Browse newspaper titles
Browse issues
Use encyclopedia
Submit transcripts
Submit comments on newspaper
Provide feedback on systems

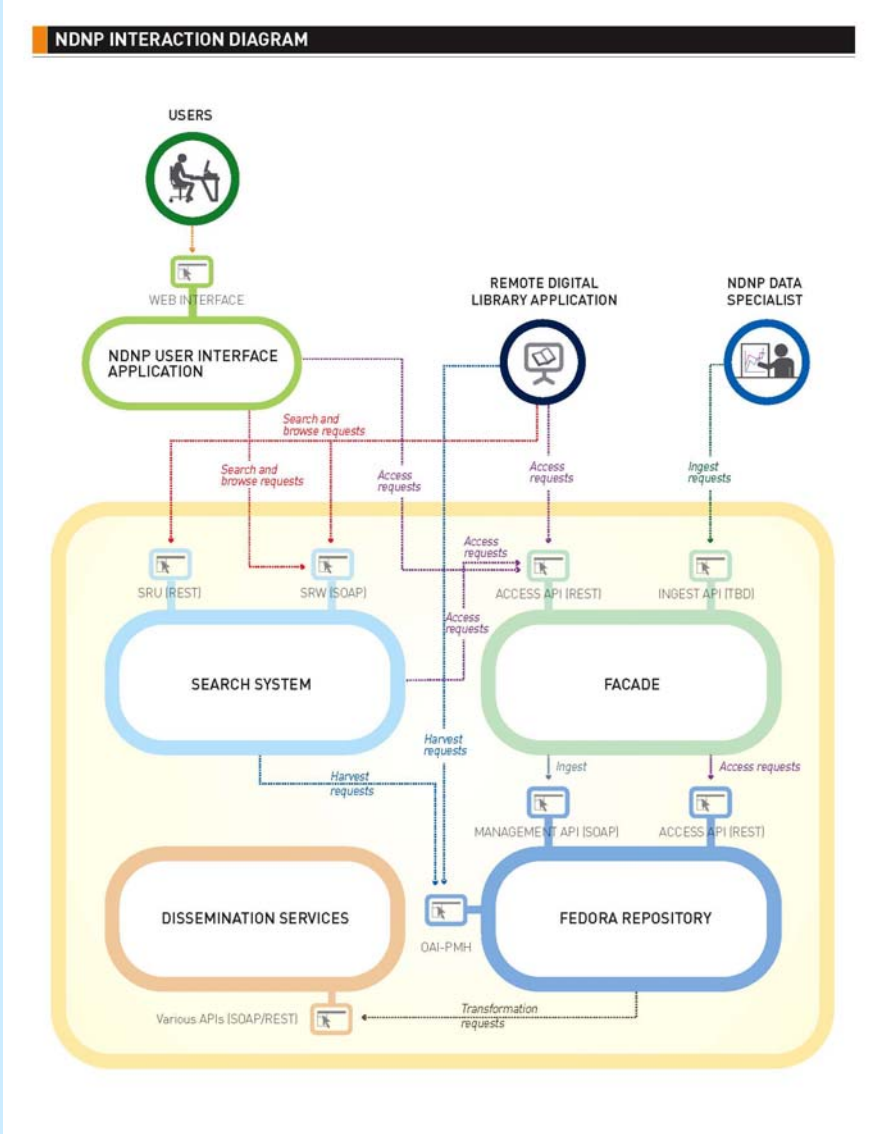
# Interoperability

- Standards-based
  - JPEG
  - XML-METS/ MODS/ALTO
  - TIFF
  - JPEG2000
  - PDF
  - JPEG
- Web exposed API's*
  - Search API- SRW** with extensions
  - Access API
  - Ingest API
  - Relations API
  - (tbd) Export API
  - (tbd) OAI-PMH to enable Metadata Harvesting **
  - (tbd) Client library

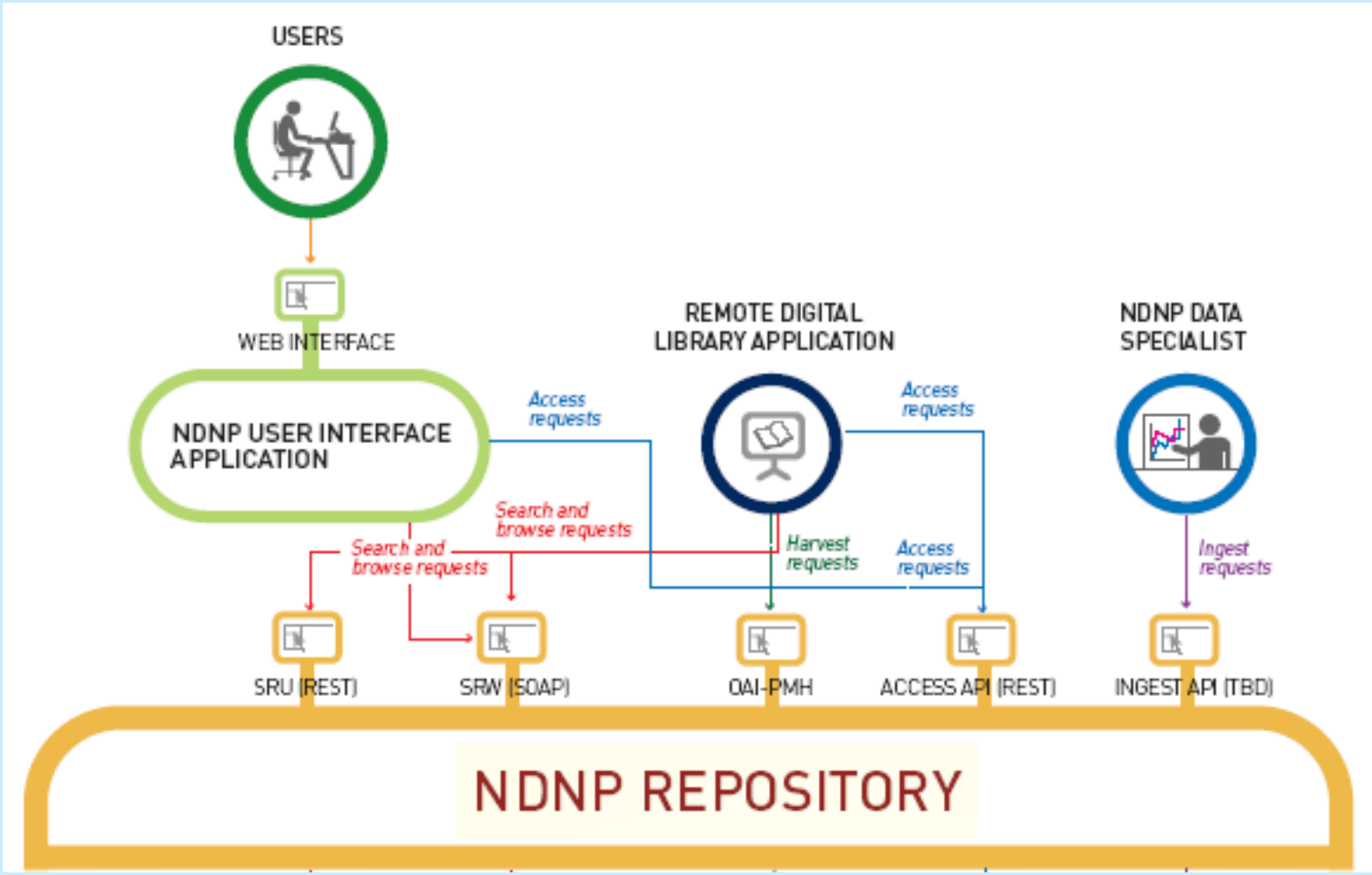
•* Abbreviation of application program interface, a set of routines, protocols, and tools for building software applications. A good API makes it easier to develop a program by providing all the building blocks.

•**SRW is a variation of SRU. Messages are conveyed from client to server, not by a URL, but instead using XML over HTTP via the W3C recommendation, SOAP, which specifies how to wrap an XML message within an XML envelope. The SRW specification tries to adhere to the Web Services Interoperability profile

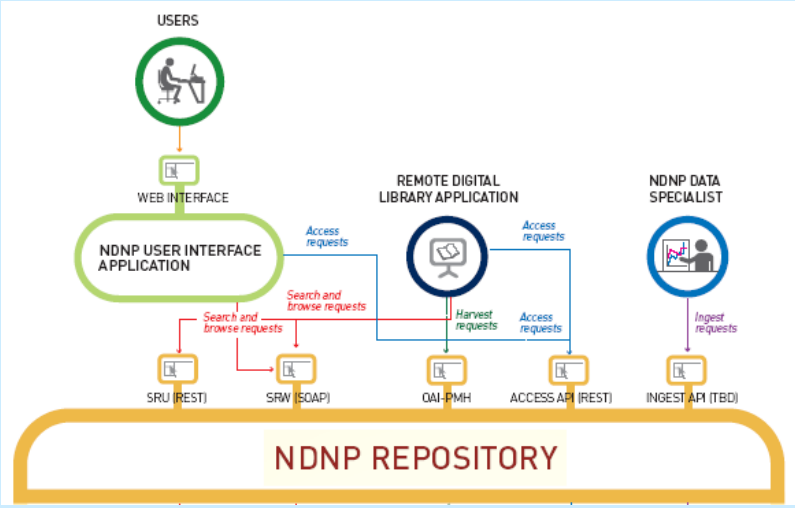

NDNP DATA FLOW DESIGN
KEY
Ingestion
Request
Response
AWARDEE
STAGING AREA
TEMPORARY STORAGE
VALIDATE
NORMALIZE
INGEST
REPOSITORY FACADE
ACCESS FACADE
INGEST FACADE
END USERS
END USERS
END USERS
USER INTERFACE
FEDORA REPOSITORY
DISSEMINATION WEB SERVICES
INDEX & SEARCH
INDEXING
SEARCH


NDNP INTERACTION DIAGRAM
USERS
WEB INTERFACE
NDNP USER INTERFACE APPLICATION
REMOTE DIGITAL LIBRARY APPLICATION
NDNP DATA SPECIALIST
Search and browse requests
Search and browse requests
Access requests
Access requests
Ingest requests
SRU (REST)
SRW (SOAP)
Access requests
ACCESS API (REST)
INGEST API (TBD)
Access requests
SEARCH SYSTEM
FACADE
Harvest requests
Harvest requests
Ingest
Access requests
MANAGEMENT API (SOAP)
ACCESS API (REST)
DISSEMINATION SERVICES
OAI-PMH
FEDORA REPOSITORY
Various APIs (SOAP/REST)
Transformation requests

# DIP – Interfaces

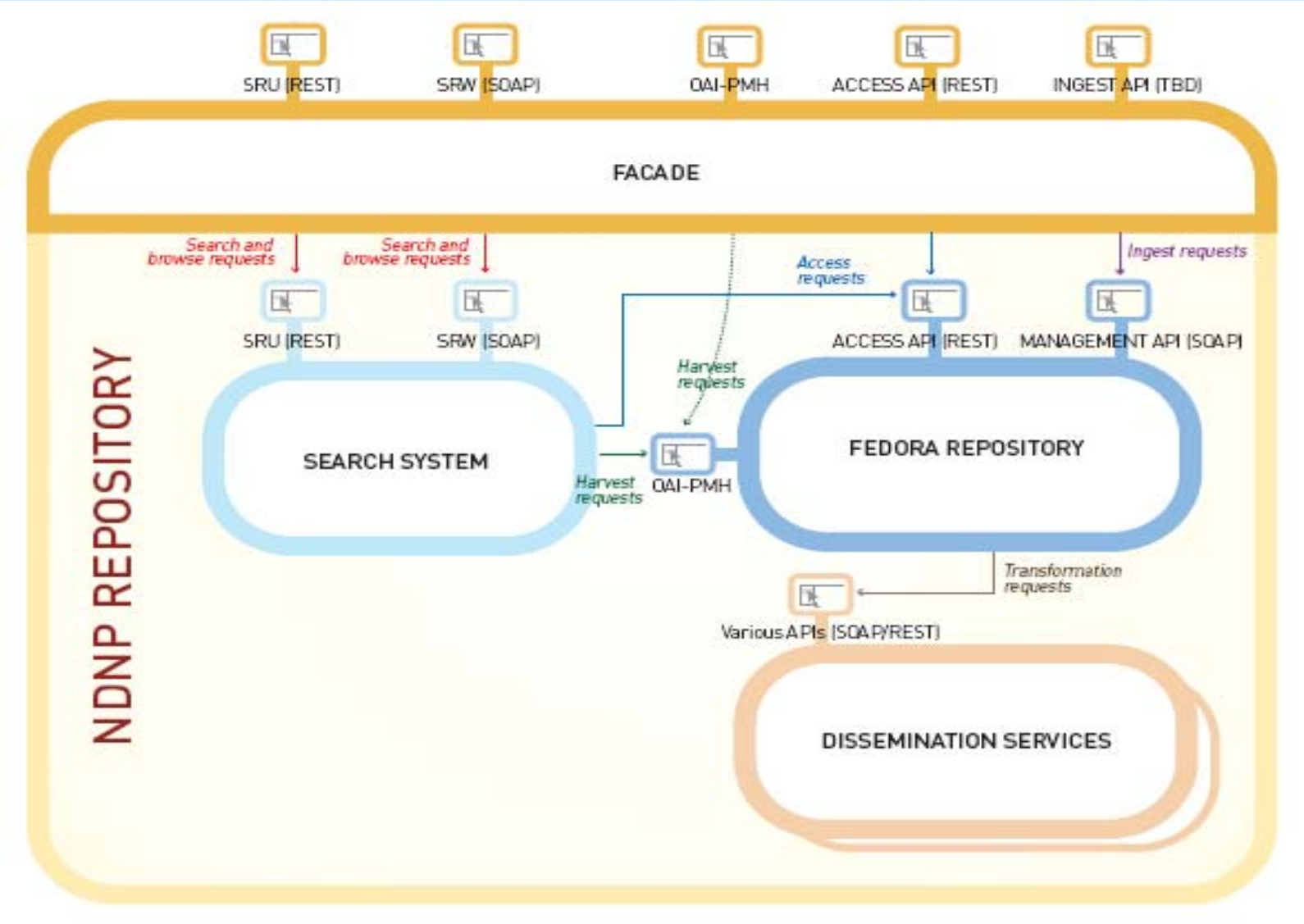


# DIP – Consumers

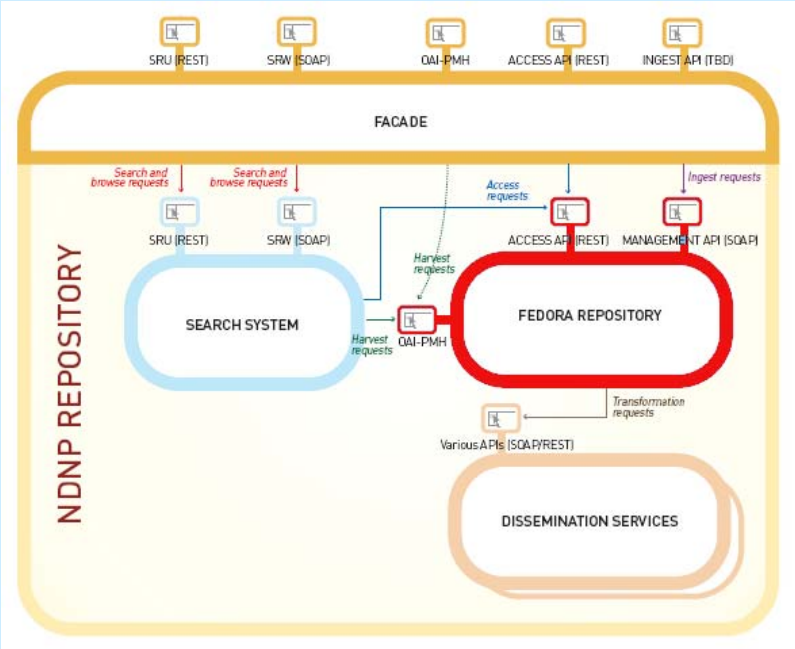



- In-house special presentation user interfaces
- Third-party user interfaces
- Other digital libraries
- Google
- etc….

# Archival Information Package

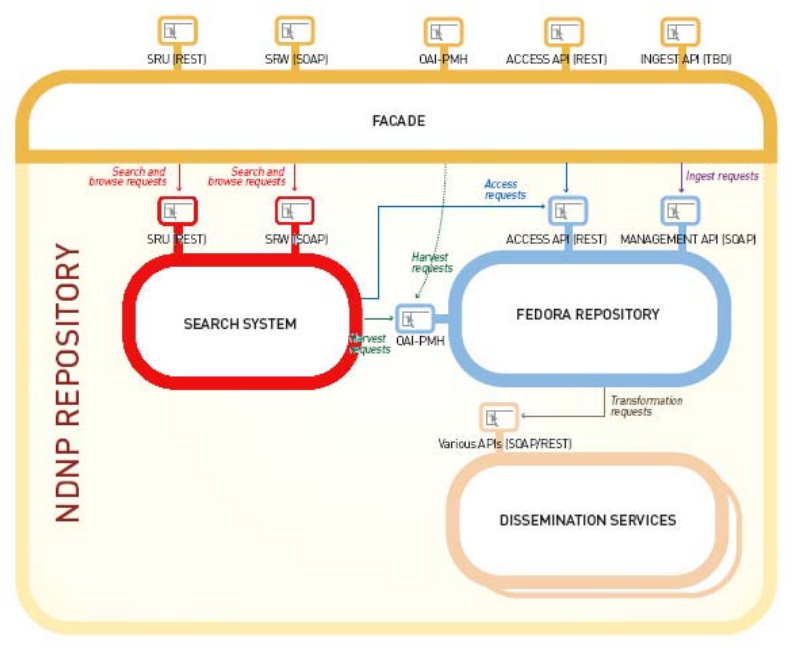


# AIP – Fedora



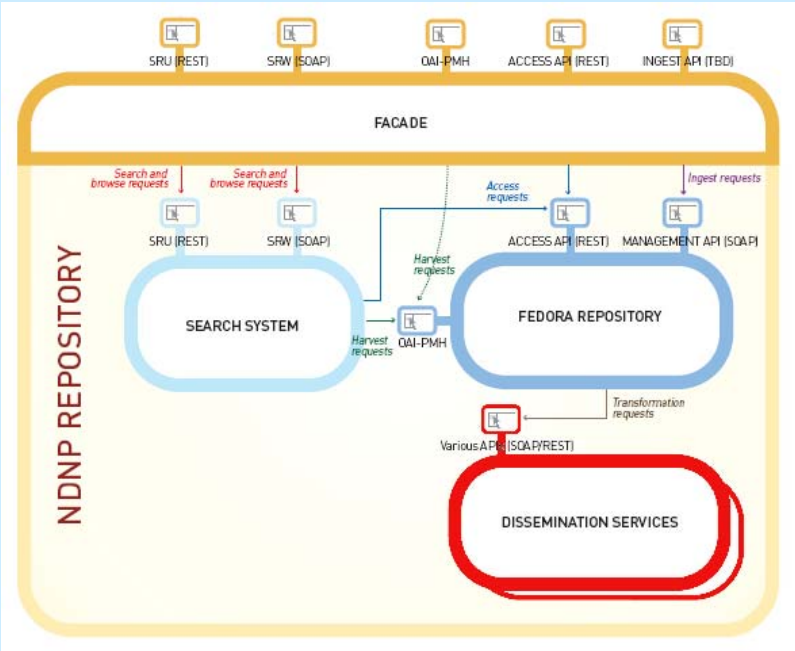

- Flexible Extensible Digital Object Repository Architecture
  - Open source digital library plumbing system
  - Joint venture between UVa and Cornell, funded by Mellon Foundation
  - Low-level storage
  - Digital object polymorphism
  - http://fedora.info

# AIP – Index & Search



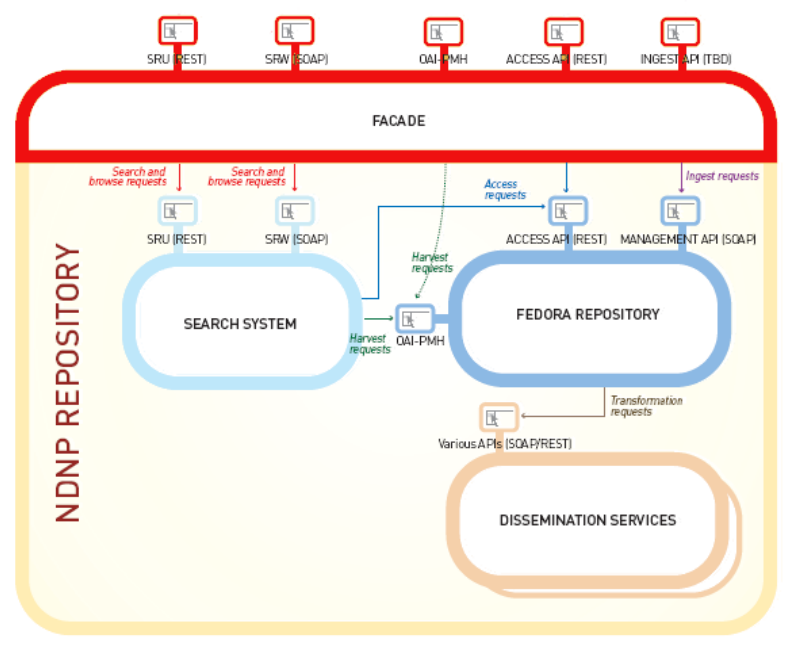

- Apache Lucene
  - Open source
  - Structured metadata searching
  - Full-text searching with Hit Highlighting
  - Virtual collections
  - http://lucene.apache.org

# AIP – Disseminations




- Web services for transforming content for export and display
- Apache Cocoon
  - XML Pipeline Framework
  - Avalon Component Container
  - http://coccoon.apache.org
- Aware JPEG2000 Libraries

# AIP – Repository Facade




- Provides standard interfaces to underlying composited systems
- Orchestrates interaction between other subsystems
- Implementation point for other sundry business logic
  - Digital rights restrictions, Object hierarchy enforcement, SIP Validation
- Apache Cocoon

# LC NDNP Team

**Program Committee:** Mark Sweeney (LS/PRD), Caroline Arms (OSI), Babak Hamidzadeh (OSI), Ray Murray (LS/DCT), George Schlukbier (OSI), Deborah Thomas (LS/PRD)

**Conversion Team:** Ray Murray (LS/DCT), Myron Briggs (LS/DCT), Pete Richey (LS/DCT)

**Development Team:** Babak Hamidzadeh (OSI), George Schlukbier (OSI), Corey Keith (Contractor), Curt Harvey (Contractor), Justin Littman (OSI), Brian Vargas (Contractor), Dave Woodward (ITS), Dave Foulks (ITS) John Williams (Contractor) with additional efforts by: Steve Engratt (OSI) and OSI Web Services.