Digital Libraries: from metadata to mega-collections

Presented by
Mark Sullivan
Systems Department
Digital Library Center, UF

Roadmap

1) What is a digital library
2) [Digital] Library Standards
3) Trends in Digital Libraries
4) Digital Library Technologies
5) UF Digital Collections Technologies
6) Example Digital Collections
7) DLC Workflow

What is a Digital Library?

Confusion continues, last 25 years gave us these definitions:

- Machine readable data files related to technical & scientific research
- Components of the National Information Infrastructure: this was tied to metadata repositories created for spatial data sets collected by Federal agencies
- Online databases/CD products
- Networked library systems

Current terminology doesn’t help:

- electronic library
- electronic archives
- digital archives
- digital repository
- digital collection
- virtual collection
- virtual library
- library without walls

CURRENT DEFINITIONS IN VOGUE:

Digital libraries basically store materials in electronic format and manipulate large collections of those materials effectively.

--University of Illinois at Urbana-Champaign Digital Libraries Initiative

"...a library that maintains all, or a substantial part, of its collection in computer-processible form as an alternative, supplement, or complement to the conventional print and microform materials that currently dominate library collections."

--"Do Digital Libraries Need Librarians," by Lisa Matson and David Bonski, Online Magazine, November 1997. http://www.onlinemag.net/NovOL97/matson11.html

Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities.

--Digital Library Federation. A working definition of digital library. 1998. ***http://www.clir.org/diglib/dldefinition.htm****.*

So much for a definition...

Perhaps we can safely suggest that digital libraries are collections of electronic resources that are aggregated to serve a definable function.

Now we are getting some where

Digital libraries differ from traditional libraries in that access rather than ownership becomes a major consideration.

This has a major impact on libraries where success of a library has always depended on its size. The bigger a library's collection the greater its success.

The Impact of ACCESS

In a 2001 Nature article (v.411, no.6837, p.521) Steve Lawrence writes about a use analysis study of 119,924 conference articles in computer science and related disciplines:

“The mean number of citations to offline articles is 2.74, and the mean number of citations to online articles is 7.03, or 2.6 times greater than the number of offline articles.”

(http://www.neci.nec.com/~lawrence/papers/online-nature01/)

Could we build a digital library without owning a single digital object?

- Yes
- No

- Maybe so and here's why

What can be included in digital libraries?

The California Digital Library contains:

- **University of California Digital Collections- all the Web sites from UC campuses**
- **Counting California-contains numeric data about California**

- Directory of CDL-Licensed Content

- eScholarship – includes texts published by UC scholars including books, journals, journal articles, prepublication materials, etc.

- Melvyl union catalog-holdings of all 10 UC campus libraries +

- Online Archive of California (OAC)-federates access to museums, archives, and libraries

Major digital library initiatives:

- American Memory Project
- Colorado Digitization Project
- Cultural Heritage Materials
- American South
- National Science Digital Library(SMETE)
- Geospatial Data Clearinghouse
- Eisenhower National Clearinghouse for Mathematics and Science Education (ENC)

Major “digital library” initiatives:

- Google Books
- Yahoo
- Gallica (Bibliothéque nationale de France)

UF Digital Library Initiatives

- PALMM Project
- International Children's Digital Library
- University of Florida Digital Collections
- National Digital Newspaper Program

[Digital] Library Standards

- MARC Records
- New Metadata Formats
- Sharing Your Resources

MARC Records

So now we can build a record (metadata) of a site

We are most familiar with the catalog records we see in WebALEPH. These are based on a metadata format called MARC.

The MARC metadata format is extremely rich in the number of information fields it offers and is also extremely time consuming to produce.

New Metadata Formats Emerge

- METS
- Dublin Core
- MODS

Sharing of Your Resources

- Real-time searches v. Harvesting
- Z39.50
- OAI-PMH

Trends in Digital Libraries

- Full Text Searching (v. subject keywords)
- Geographic placing of resources
- Who’s harvesting my harvester while I was out harvesting...

Trends in Digital Libraries

- Consolidation
  - Within Universities
- Collaboration
  - Collections
  - Open Source Software

Trends in Digital Libraries

- Institutional Repositories
- Open Access

- Learning Modules
- Adaptive Libraries

Digital Library Technologies

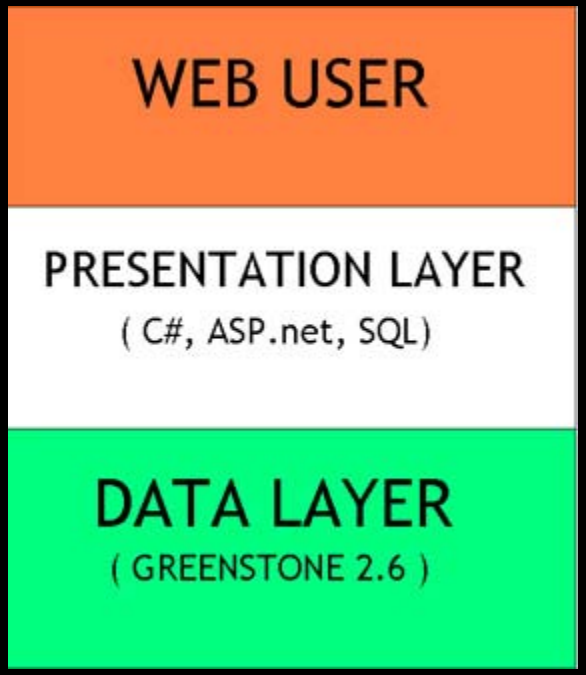
- DigiTool (exLibris)
- ContentDM

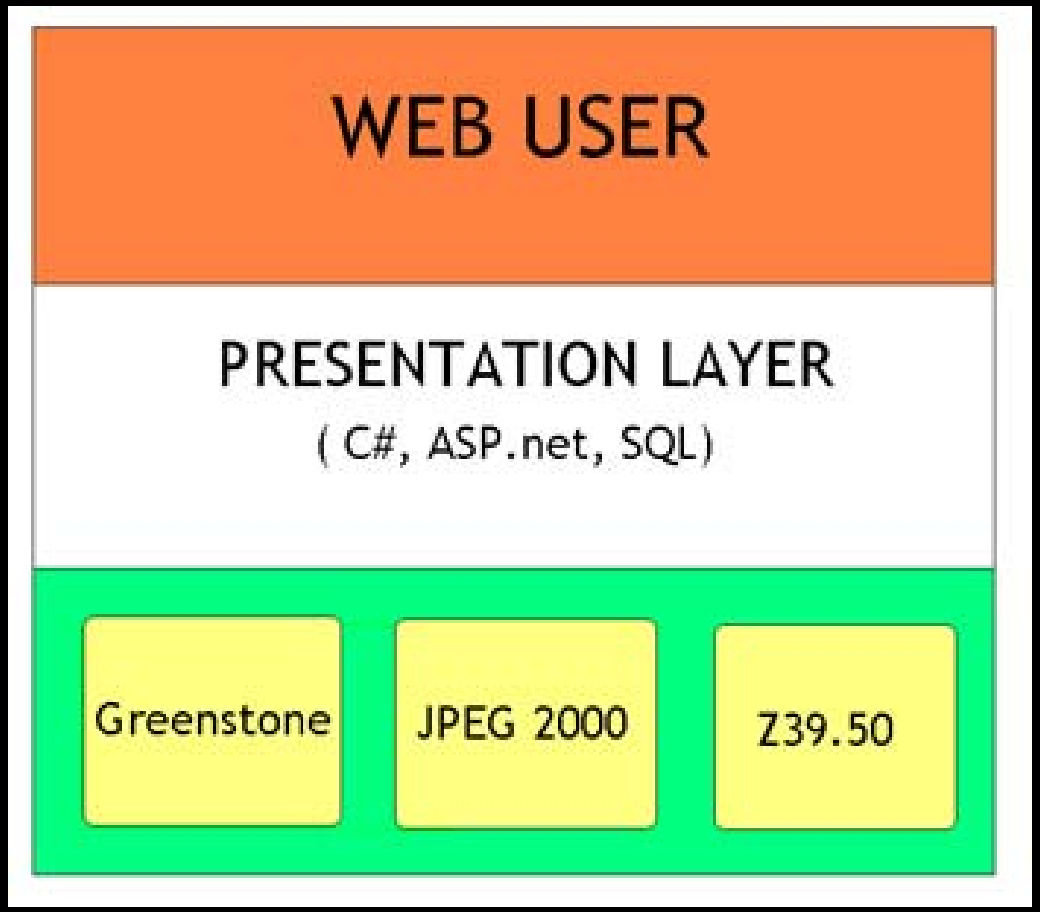
- Greenstone Digital Library
- Fedora

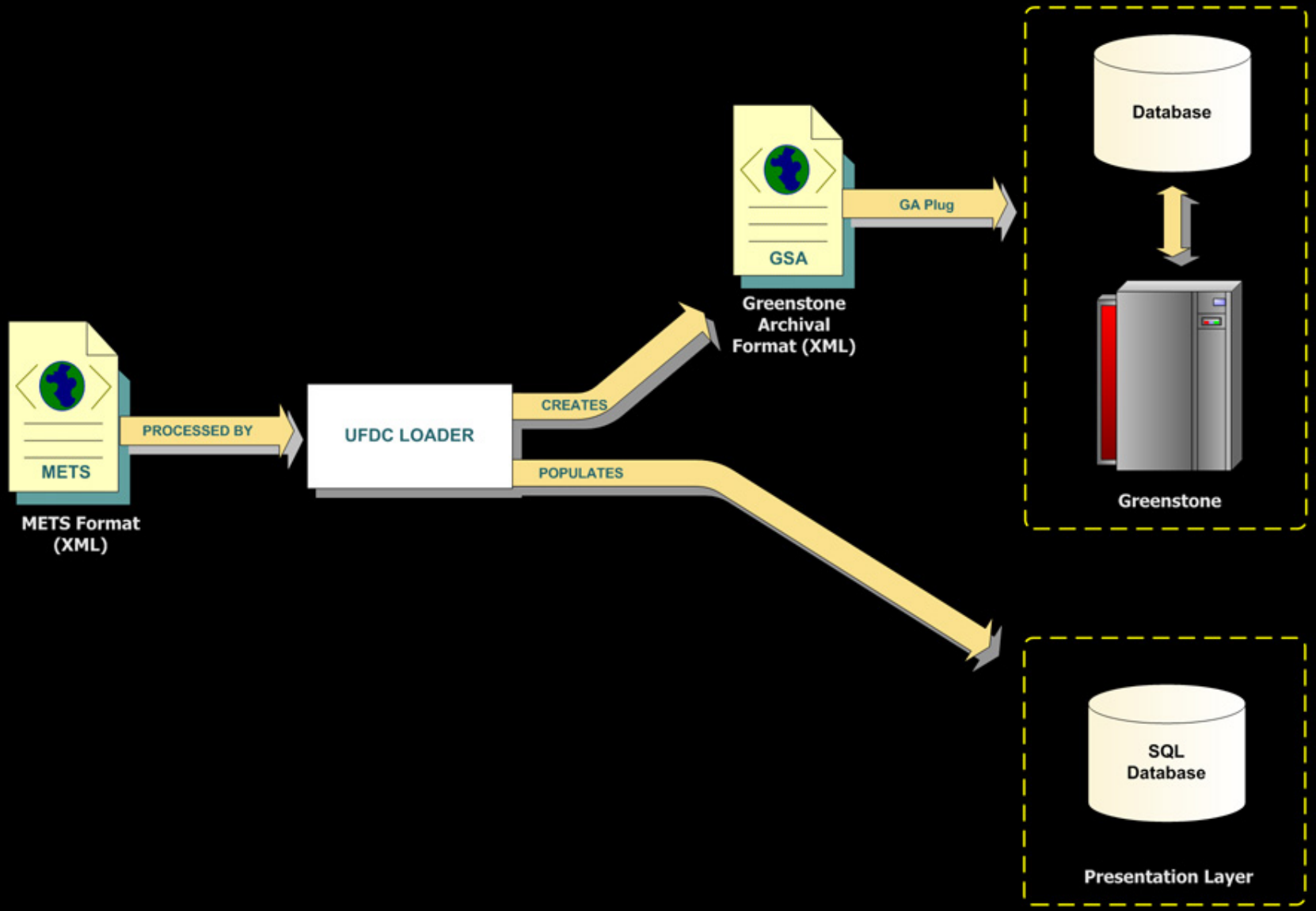
UF Digital Collections

http://www.uflib.ufl.edu/ufdc

Basic architecture




WEB USER
PRESENTATION LAYER
(C#, ASP.net, SQL)
Greenstone
JPEG 2000
Z39.50


Database
GA Plug
GSA
Greenstone
Archival
Format (XML)
METS
PROCESSED BY
UFDC LOADER
CREATES
POPULATES
METS Format
(XML)
Greenstone
SQL
Database
Presentation Layer

Greenstone Digital Library Software

- Indexes metadata and full text
- Resource discovery
- Stores complete bibliographic record

UFDC Database

- Item display information
- Related downloads for each item
- Collection Hierarchy
- Interfaces
- Authority Database (new)
- Database Diagram
- Details

Example Digital Collections

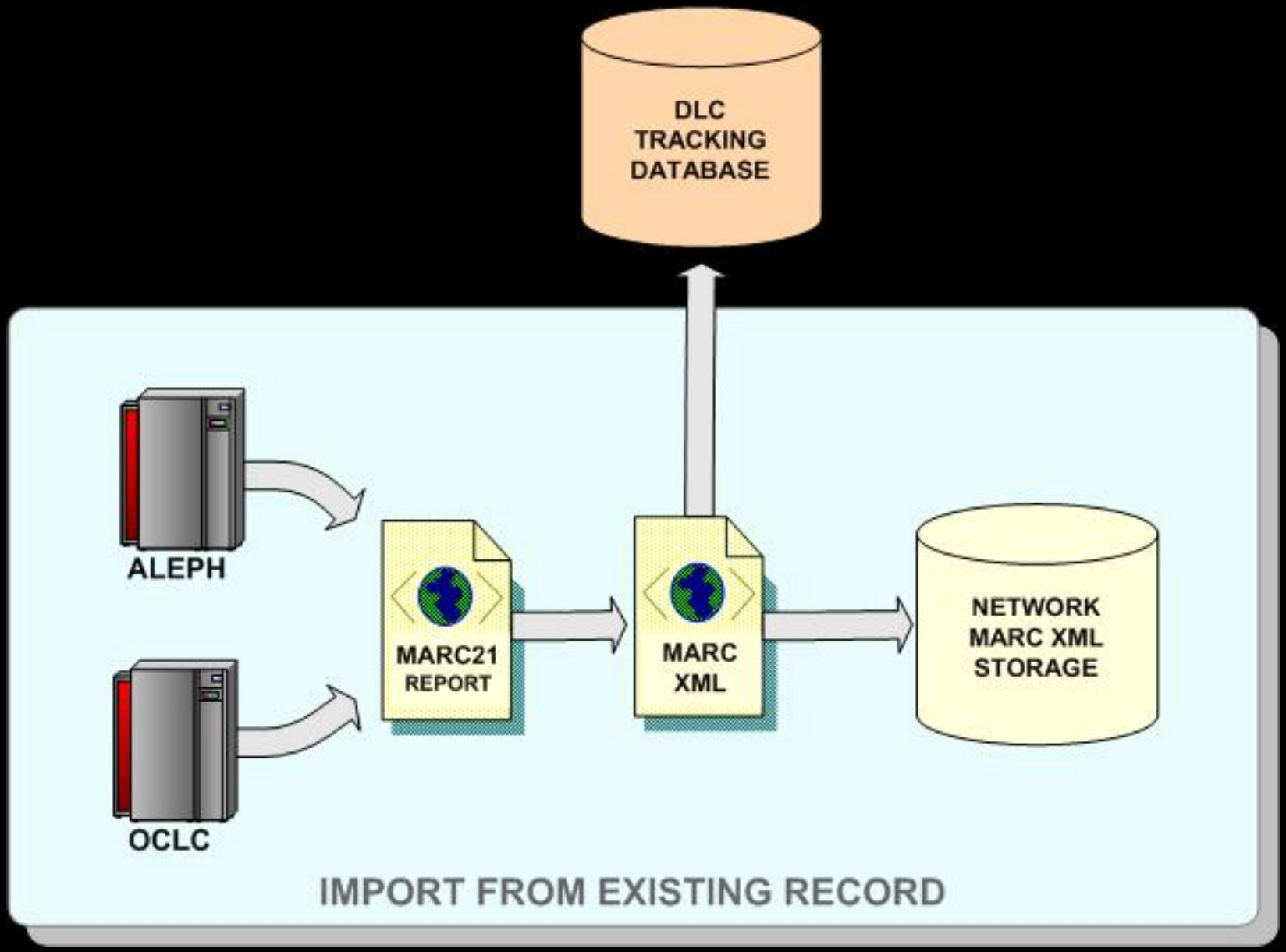
- Baldwin Collection of Children's Literature
- Florida Digital Newspaper Project
- Digital Library of the Caribbean
- Ephemeral Cities Project

Digital Library Center Workflow

- Importing any existing data sources
- Scanning / Digitization
- Post-Capture Processing
- Quality Control
- Optical Character Recognition [OCR]
- Final preparation and review for UFDC

Data Import

- Catalog Records (MARC)
  - ALEPH
  - OCLC
- Custom Spreadsheets
- Tracking Databases


DLC
TRACKING
DATABASE
ALEPH
OCLC
MARC21
REPORT
MARC
XML
NETWORK
MARC XML
STORAGE
IMPORT FROM EXISTING RECORD

Scanning / Digitization

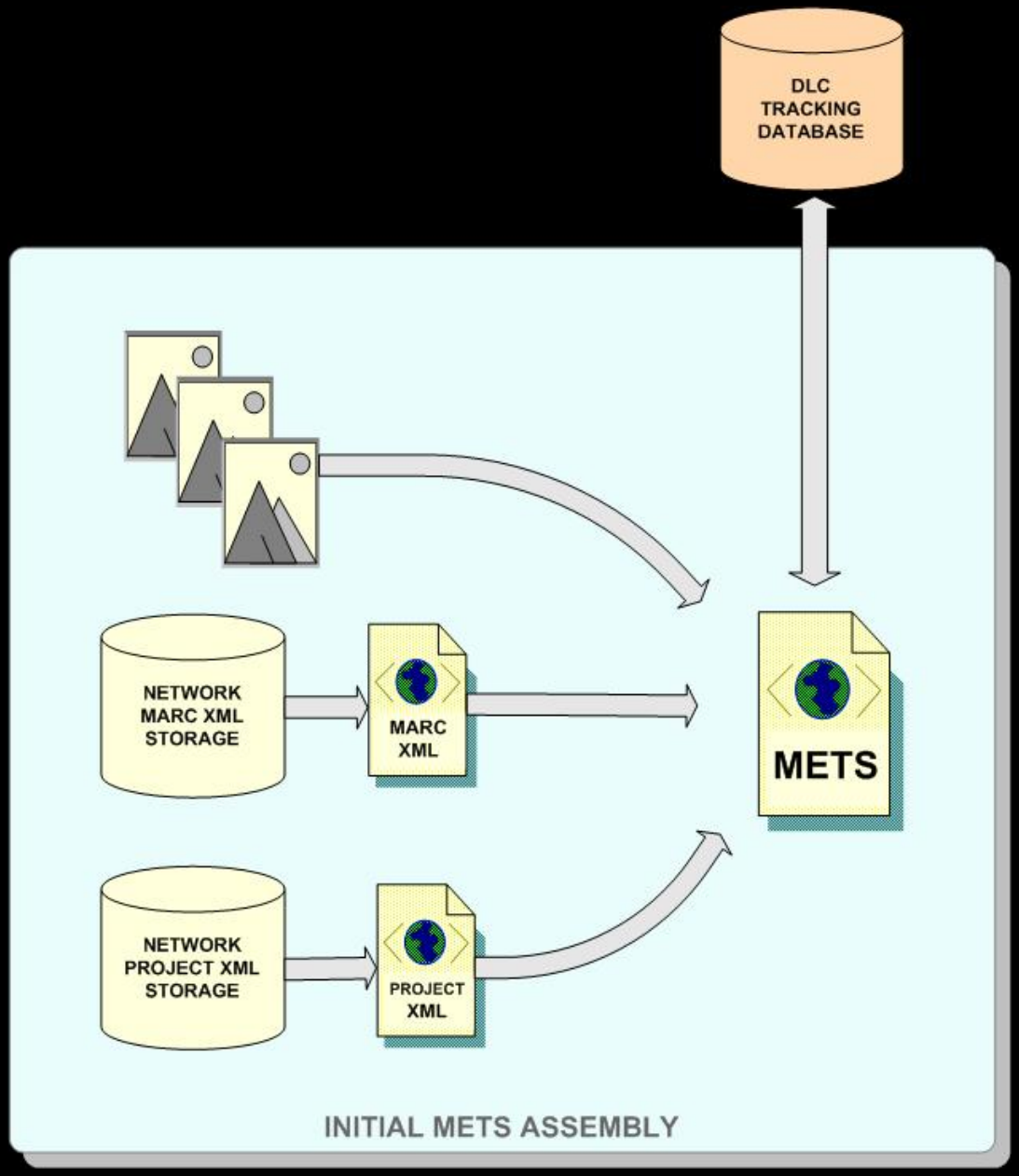
- Microtek Flat-Bed Scanners
- Panasonic High-Speed Scanners
- Large Format Scanner
- CopiBook Scanners

Post-Capture Processing

- Image Processing
  - Project Specific
    - Newspaper project
    - Funeral Records
  - General Image Clean-up

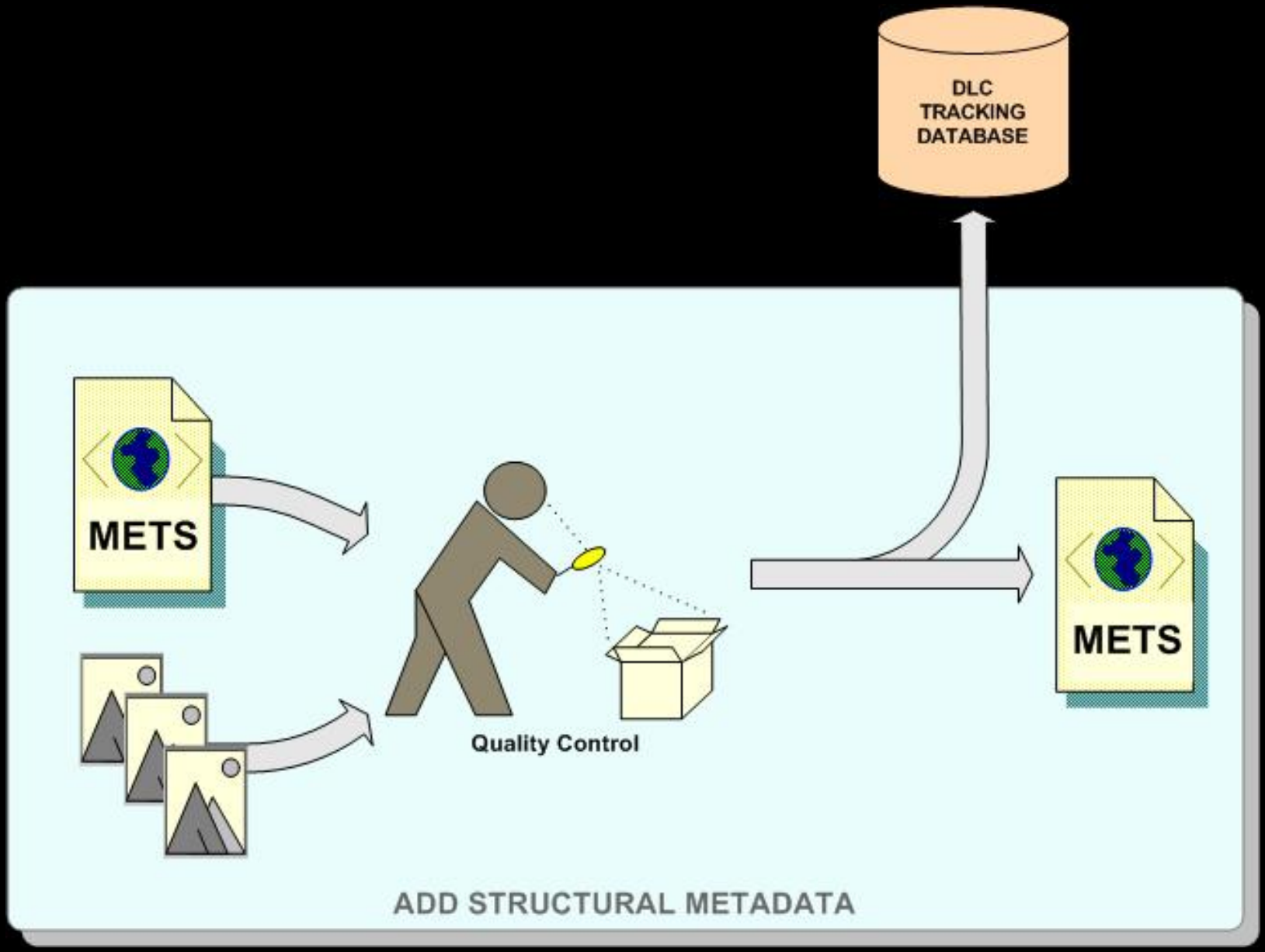
Post-Capture Processing

- Metadata (METS) Assembly
  - Imported records
  - Images
  - Project-level metadata
  - Tracking database


DLC
TRACKING
DATABASE
NETWORK
MARC XML
STORAGE
MARC
XML
METS
NETWORK
PROJECT XML
STORAGE
PROJECT
XML
INITIAL METS ASSEMBLY

Quality Control

- Image QC and Clean-Up
- Structural Metadata Created
- Metadata Reviewed


DLC
TRACKING
DATABASE
METS
Quality Control
METS
ADD STRUCTURAL METADATA

Digital Library Quality Control Application

File Actions Help

UF00023488 : 00001

Title: Grandmother Puss, or, The grateful mouse.

00001

Page:

D: Main

00002

Page:

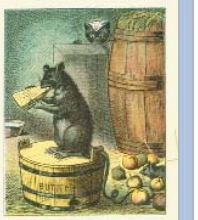
D: Main

00003

Page:

D: Main

00004



Page:

D: Main

00005

Page:

D: Main

00006

Page:

D: Main

00007

Page:

D: Main

00008

Page:

D: Main

Critical Volume Error: No Volume Level Error

Last QC'd: QC'd on 2/22/2006 by SMATHERSLIB\marsull

Cancel Save

Digital Library Quality Control Application

File Actions Help

UF00023488 : 00001

Title: Grandmother Puss, or, The grateful mouse.

00001

Page: Page 1

D: Front Cover

00002

Page: Page 2

D: Chapter

00003

Page: Page 3

D: Grandmother Puss,...

00004

Page: Page 4

D: Grandmother Puss,...

00005

Page: Page 5

D: Grandmother Puss,...

00006

Page: Page 6

D: Grandmother Puss,...

00007

Page: Page 7

D: Grandmother Puss,...

00008

Page: Page 8

D: Grandmother Puss,...

Critical Volume Error: No Volume Level Error

Last QC'd: QC'd on 2/22/2006 by SMATHERSLIB\marsull

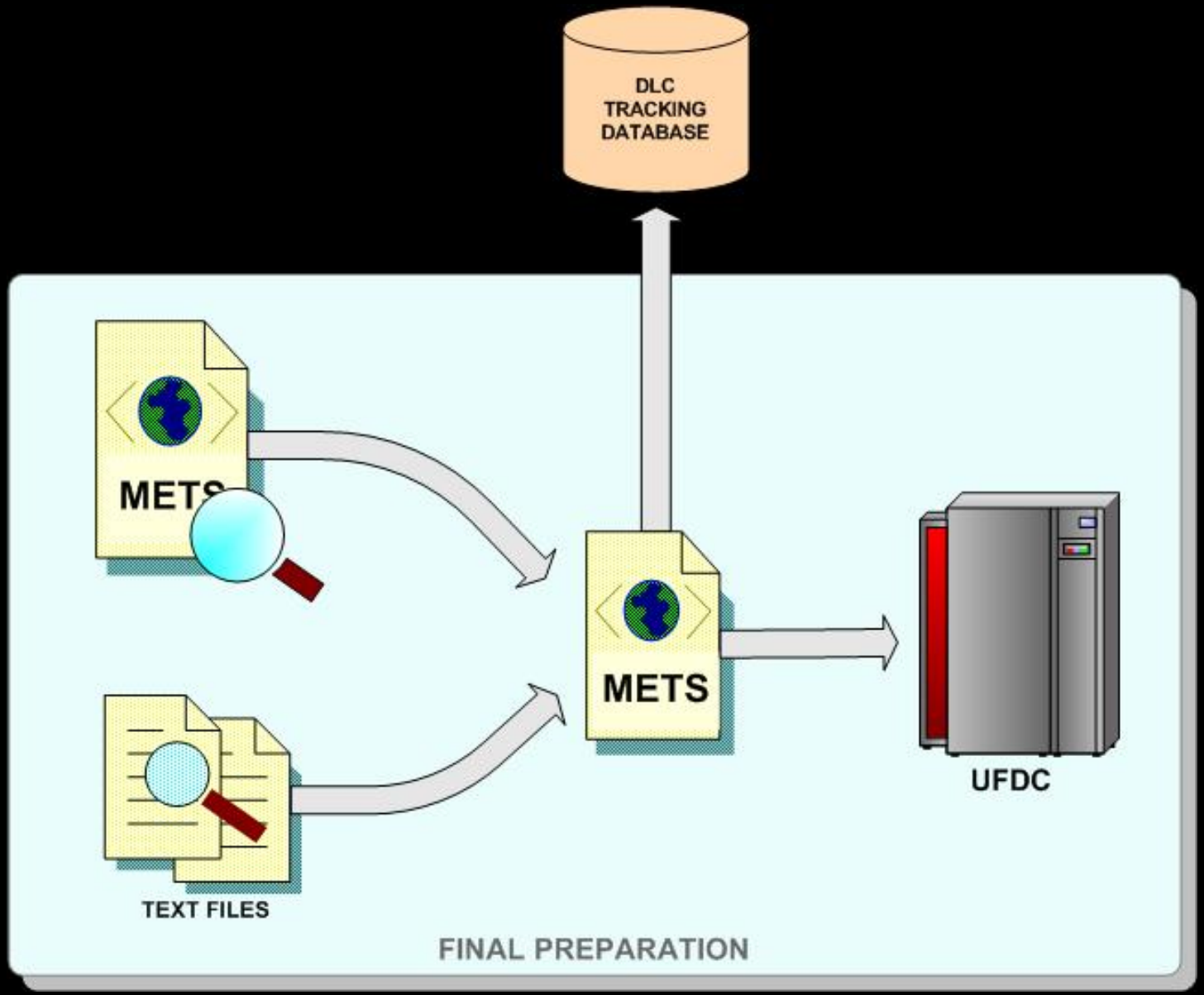
Cancel Save

Optical Character Recognition [OCR]

- Full text created for nearly every item
- Prime Recognition OCR Software
- Future data mining

Final Preparation and review for UFDC

- OCR examined
- Any additional metadata work completed
- Final METS package assembled
- Submitted to UFDC


DLC
TRACKING
DATABASE
METS
METS
TEXT FILES
UFDC
FINAL PREPARATION

Digital Library Center

Tour

Digital library initiatives:

Digital Library Initiatives Across Europe
by David Raitt
http://www.infotoday.com/cilmag/nov00/raitt.htm

D-Lib Magazine
http://www.dlib.org/