Newspaper Digitization: Issues and Opportunities

Phil Sager
OHS

OHIO HISTORY

Ohio Historical Society

MISSING YOUTHS FOUND BY HERDER

Boys, Seeking Body Of Lion, Become Lost in Hills,

After straying from a Moosheart picnic at the head of Indian canyon on the Duchesne highway at noon Sunday, two Helper boys, Jack Williams, 9, and Thomas Pegano, 5, became lost and were not found until 7:15 that evening. The youths were found by L. Tatton, a sheepherder, in Avintaqin canyon, eight miles from the picnic grounds.

The Story Behind the Story

News Advocate (Price, UT)
September 4, 1930

MISSING YOUTHS FOUND BY HERDER

Boys, Seeking Body Of Lion, Become Lost in Hills.

After straying from a Moosehart picnic at the head of Indian canyon on the Duchesne highway at noon Sunday, two Helper boys, Jack Williams, 9, and Thomas Pagano, 5, became lost and were not found until 7:15 that evening. The youths were found by L. Tatton, a sheepherder, in Avintaqin canyon, eight miles from the picnic grounds.

The Story Behind the Story

News Advocate (Price, UT)
September 4, 1930

John, I saw your article in the Tribune the other day and put it to the test. This morning I found the article I thought might exist on me and my uncle getting lost near Duchesne 74 years ago. My name then was Thomas Pagano and now is Thomas Billis. My family is ecstatic over the find. My daughter who is a genealogist went absolutely bonkers and said it was actually almost unbelievable to find such information. She said " but Dad you were only 5 years old then and are 78 now". Thank you for sharing the information that you guys exist and how we can use it.
Thomas W. Billis (2/10/04)

Why Digitize Historic Newspapers?

- Fosters interest and pride in a community's history
- Birth, death, marriage announcements
- New technologies to experience newspapers online
- Wealth of historical information
- 24/7 access, worldwide

Why more attention lately?

- National Digital Newspaper Project (NDNP)
  - "Ultimately, over a period of approximately 20 years, NDNP will create a national, digital resource of historically significant newspapers from all the states and U.S. territories published between 1836 and 1922"
  - http://www.neh.gov/projects/ndnp.html

NDNP

- 2005 awardees:
  - University of California, Riverside
  - University of Florida Libraries, Gainesville
  - University of Kentucky Libraries, Lexington
  - New York Public Library, New York City
  - University of Utah, Salt Lake City
  - Library of Virginia, Richmond

Why more attention lately?

- The technology has improved
- Though still relatively few, projects digitizing historic newspapers are increasing in number
- Public is finding out through news, announcements, google, etc.

So What's So Hard About Putting Historic Newspapers Online?

- Readability issues (human and OCR)
- Conservation issues
- Digitization / image processing
- OCR processing
- Recording metadata
  - structural, descriptive
- Storage

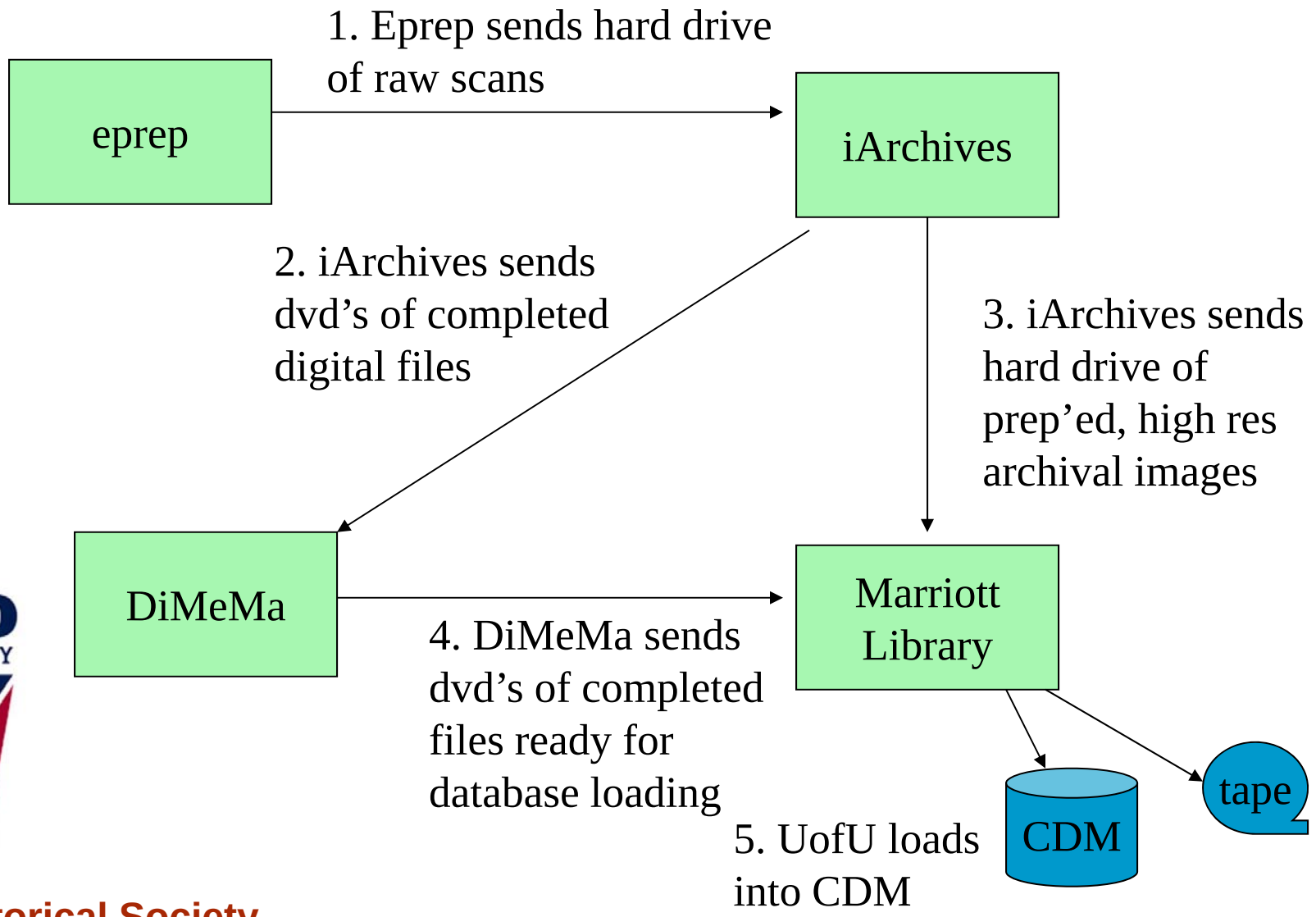
Examples

- Utah Digital Newspapers
  - http://www.lib.utah.edu/digital/unews/
- Brooklyn Daily Eagle
  - http://www.brooklynpubliclibrary.org/eagle/

High-Level Process

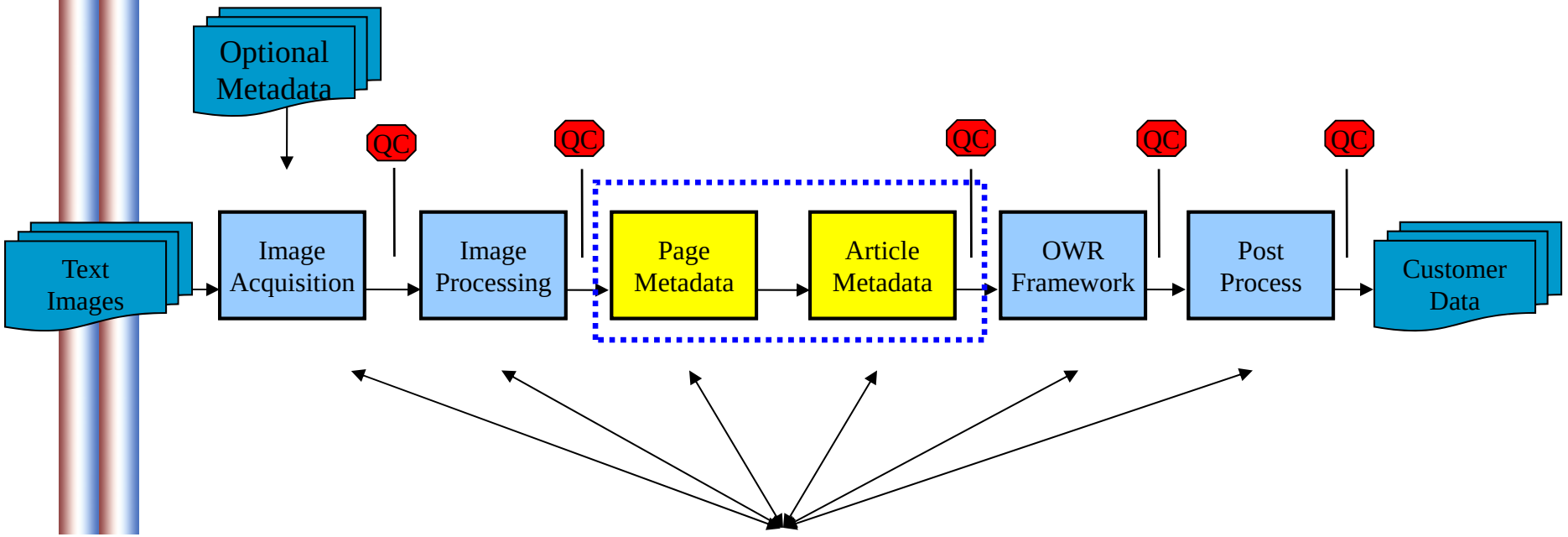
Select titles	UofU w/ Adv Bd
Obtain source materials	UofU w/
– Originals	owner
– Film	BYU
Repair originals (if needed)	UofU
Scan	
– Originals	eprep
– Film	iArchives
Zone / metadata / OCR	iArchives
Database index	DiMeMa
Database load	UofU
Web development	

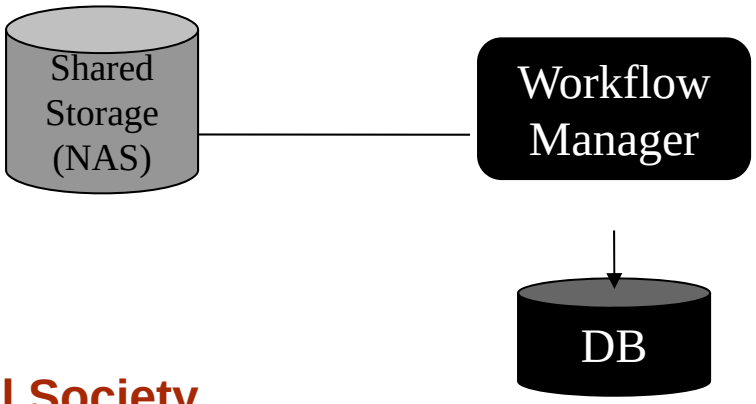
Flow of Original Papers



Ohio Historical Society

Image processing workflow

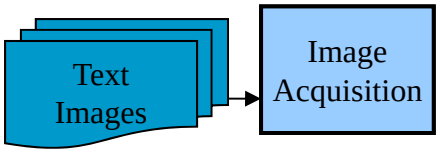




- Data
- Automatic process [image processing, OCR, …]
- Manual process [image + article metadata]
- Quality Control
- Metadata entry Delhi and Coimbatore, India

Ohio Historical Society

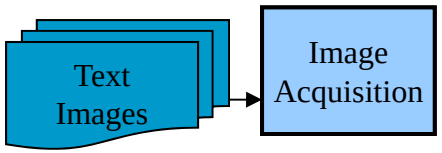
Image acquisition



The quality of the original image has direct effect on every other aspect of the newspaper digitization process.

- Photo realism
- Text readability
- Metadata collection
- Searchable word accuracy

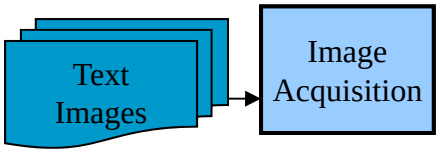
Image Acquisition



Selecting the **image type** is the first decision to be made when considering the viewing experience.

- Color, grey, or black-and-white (Bit-depth)
- Image resolution (ppi)
- Format (TIF, JPG, JP2, GIF, PDF)
- Compression type and level

Image Acquisition



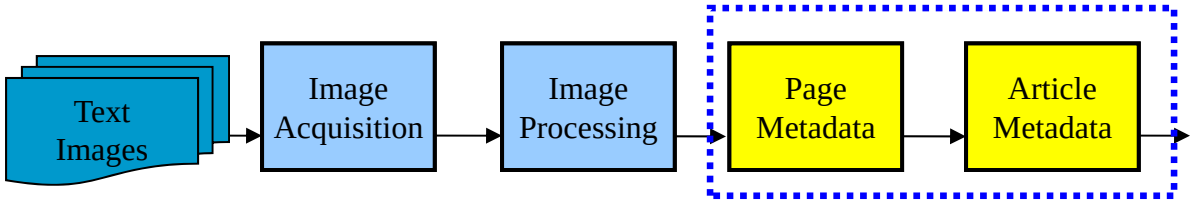
Several factors come into play when selecting the image type to be used for an online newspaper.

- End-user bandwidth
- Original media type (microfilm, microfiche, paper)
- Original material format (page size, font size, photos, etc)

Image Processing

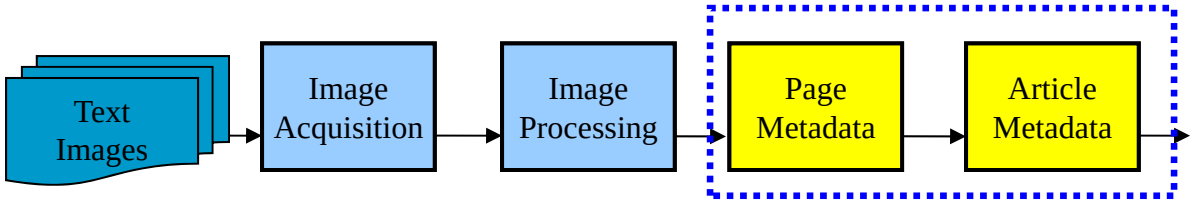
- Cropping
- Deskewing
- Despeckling
  - removal of “salt and pepper” noise that prevents good OCR
- Binarization (for OCR)
  - converts greyscale and color images into black and white images

Metadata



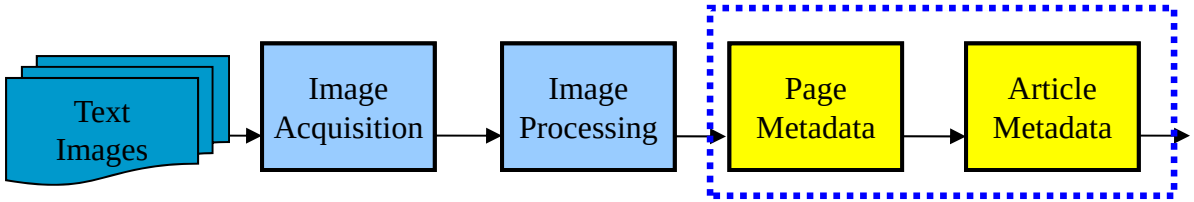
- There are Two levels of metadata in newspapers
  - Page-level
    - NDNP at present
    - Structural, Text Layout
  - Article-level
    - Structural, Text Layout
    - Title, subject, etc...
    - Others including UofU

Metadata



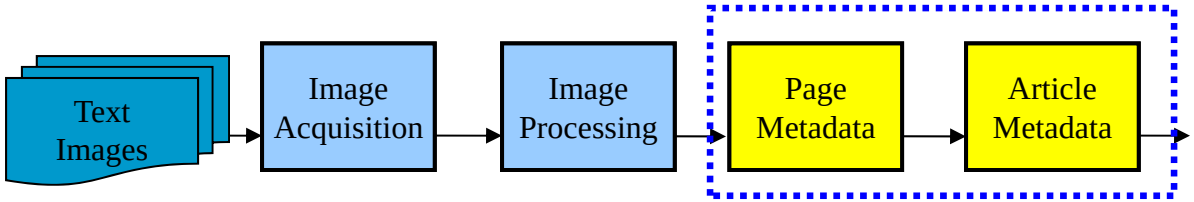
- Page-level metadata
  - Inexpensive to capture
  - OK for browsing
  - OK for searching
  - Relatively small amount of data
    - Article metadata and zoning not present

Metadata



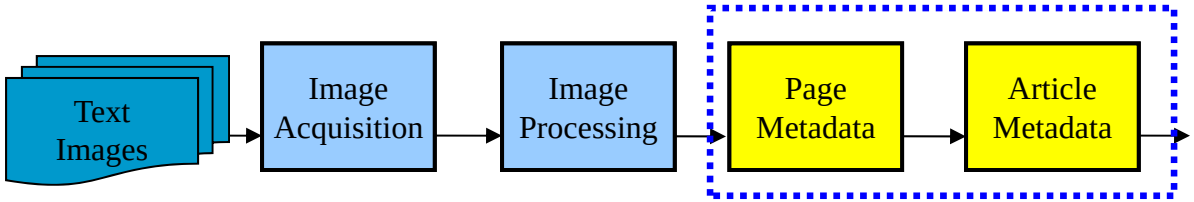
- Article-level metadata
  - More expensive to capture
  - Good for browsing
  - Good for searching
  - Relatively large amount of data
    - Article metadata and zoning present

Metadata



- Page level view
  - For example, NDNP
  - http://www.loc.gov/ndnp/xml/issueTemplate.xml

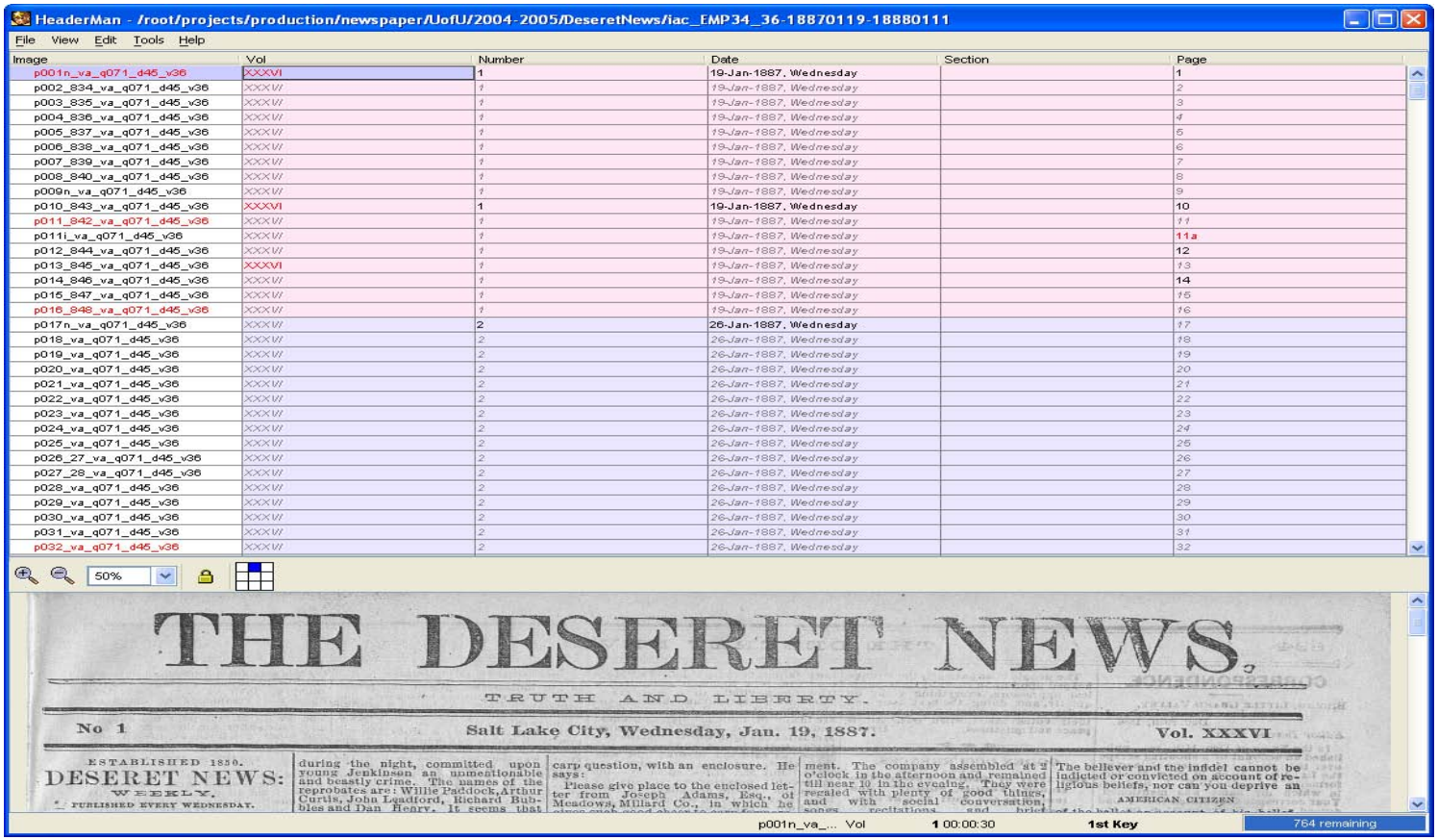
Metadata



- Article level view
  - For example, University of Utah
  - XML includes all page-level data, plus:
    - Article headline
    - Article category
      - birth announcement name
      - death announcement name
      - marriage announcement name
    - County

Tool Set - HeaderMan

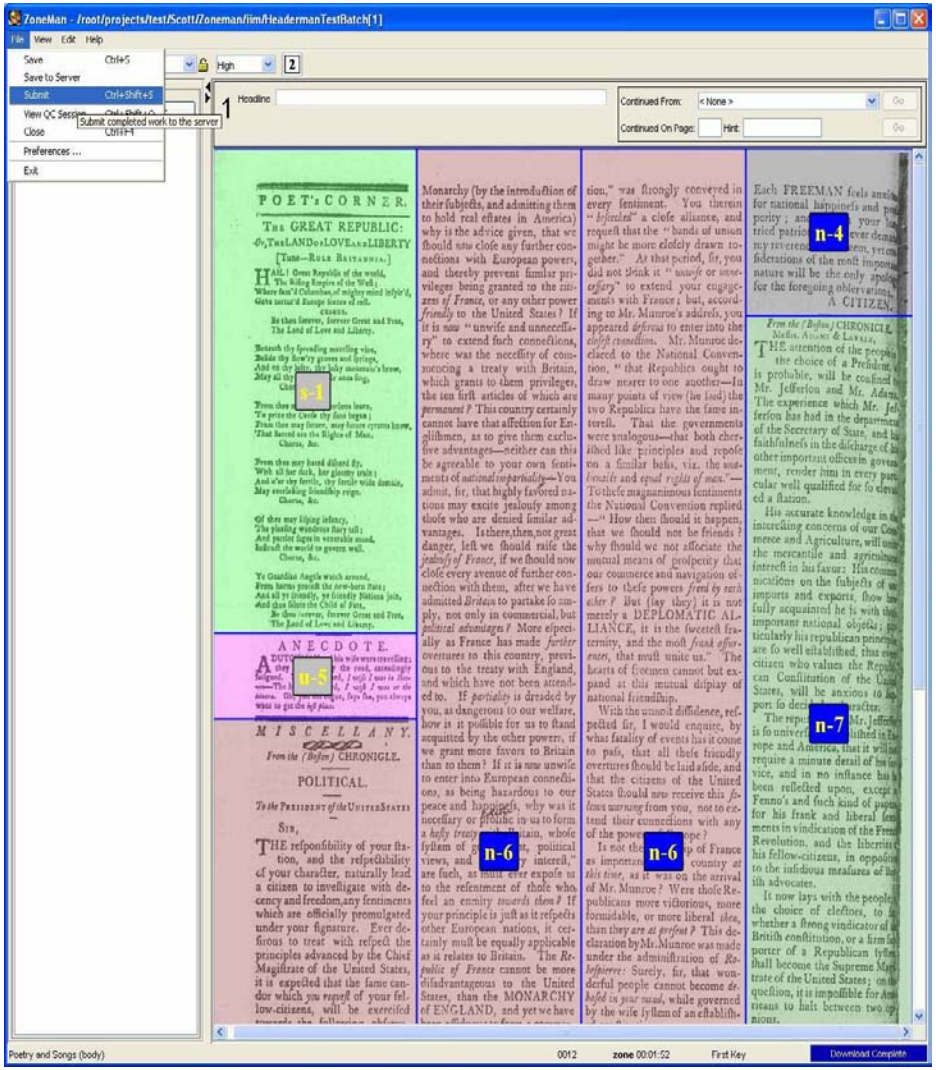
HeaderMan is a very efficient means of entering meta data to the page level. It requires 2 entries and reconcile to maximize accuracy.



Ohio Historical Society

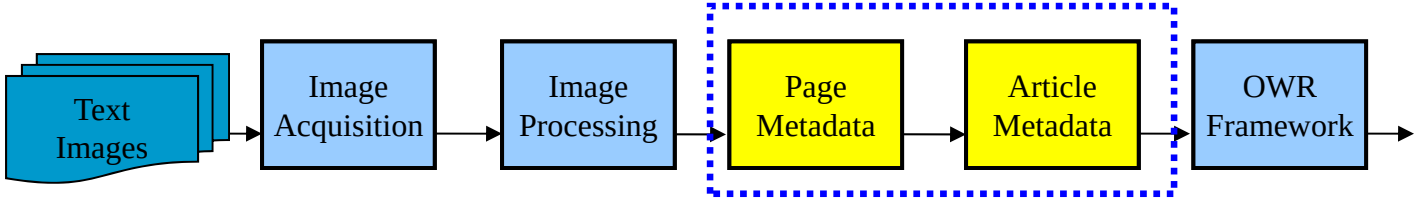
Tool Set - ZoneMan

ZoneMan is a
flexible, efficient
means to collect
customized
article-level
meta data.



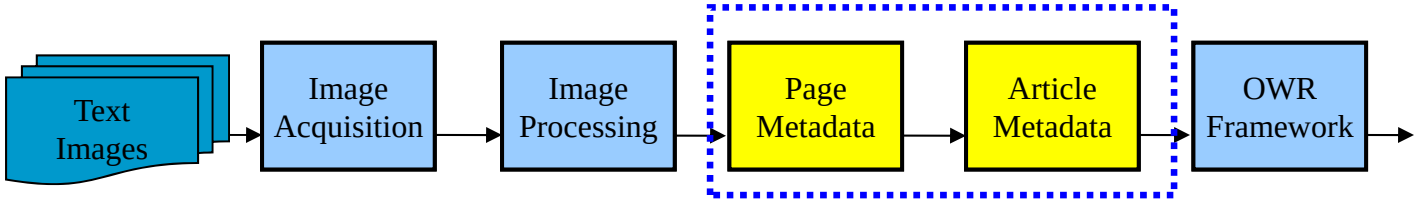
Ohio Historical Society

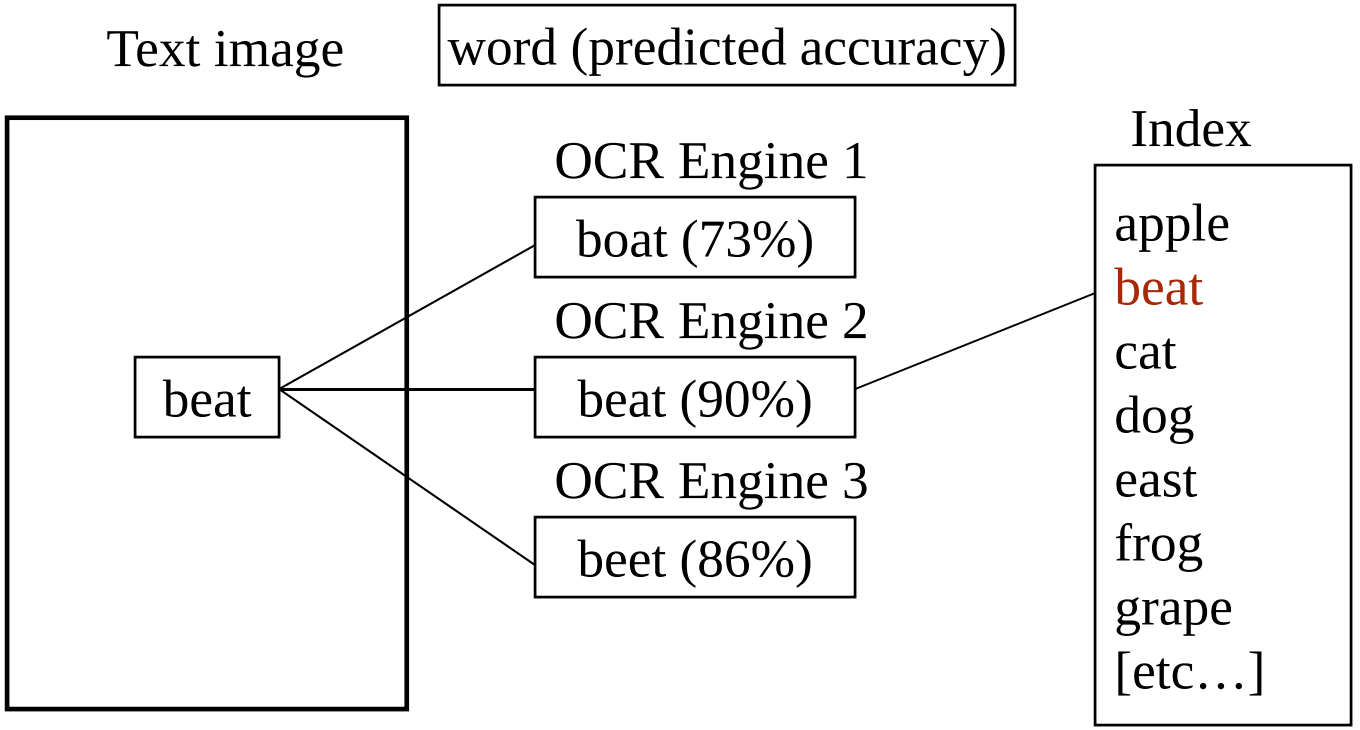
OCR and OWR



- OWR = Optical Word Recognition
- First, it's important to realize that OCR/OWR does not yield article "transcriptions"
- The text OCR'd from the images of historic newspapers is used for searching purposes

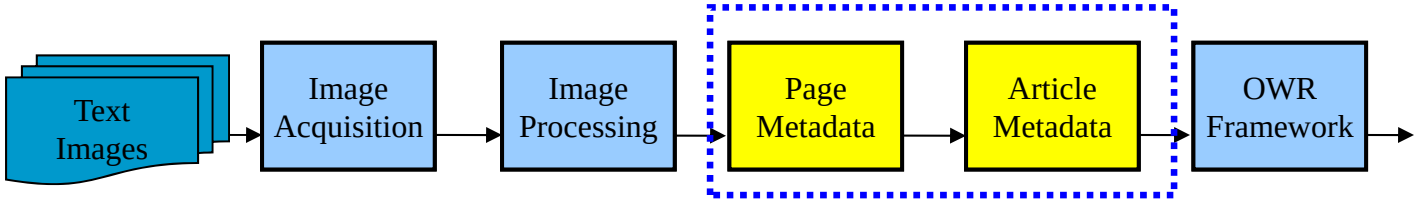
OCR and OWR



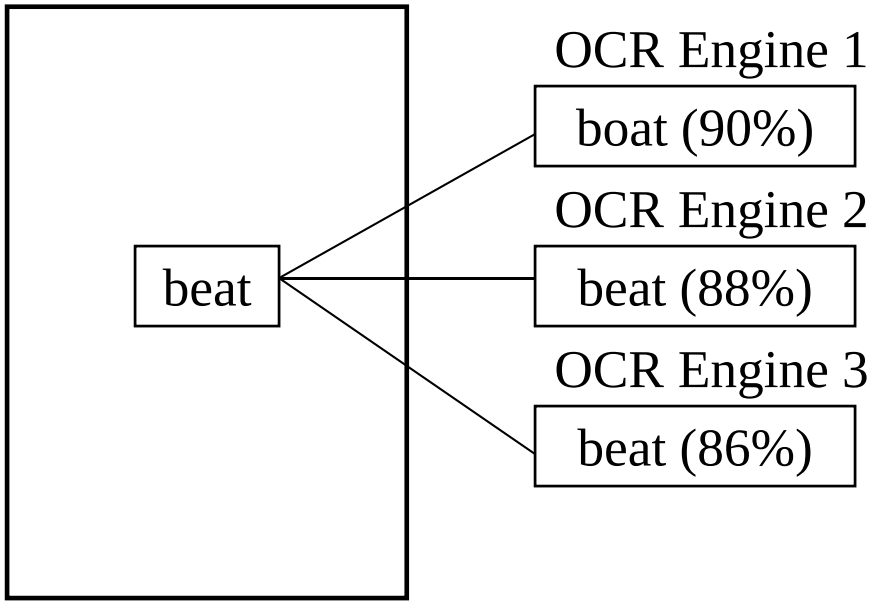


Ohio Historical Society

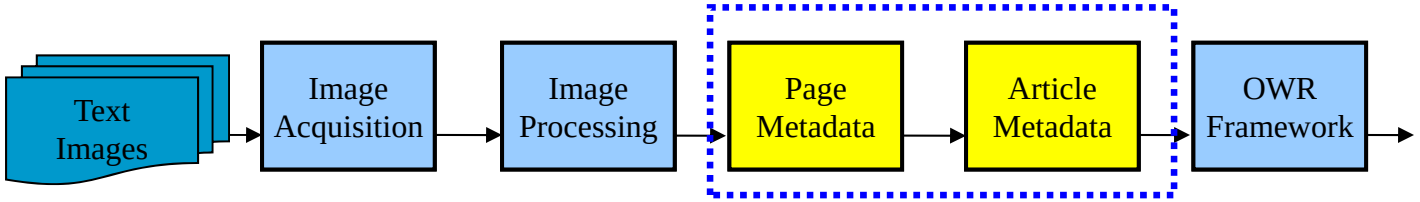
OCR and OWR



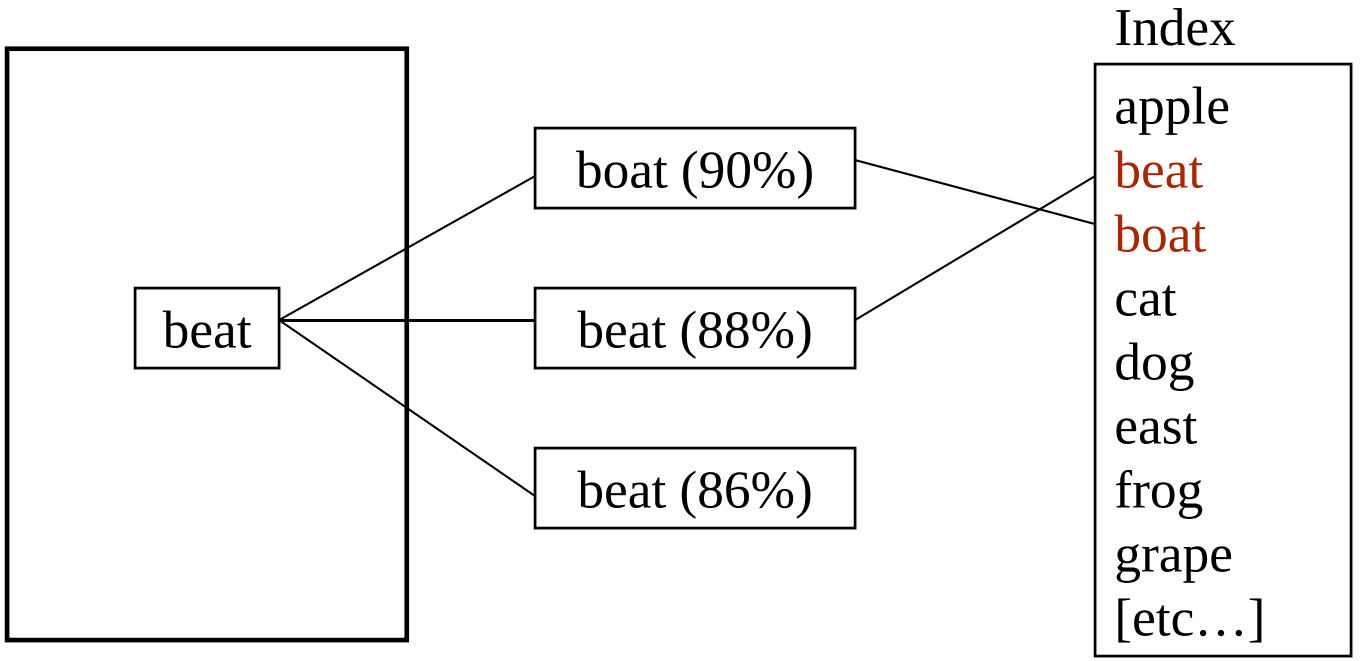
What if there are OWR ties?



OCR and OWR

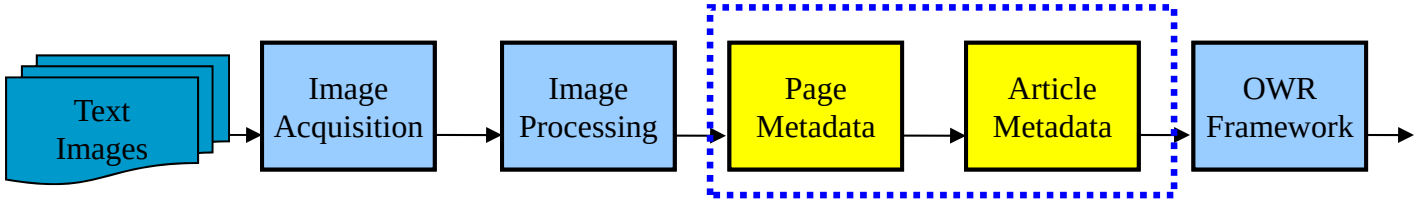


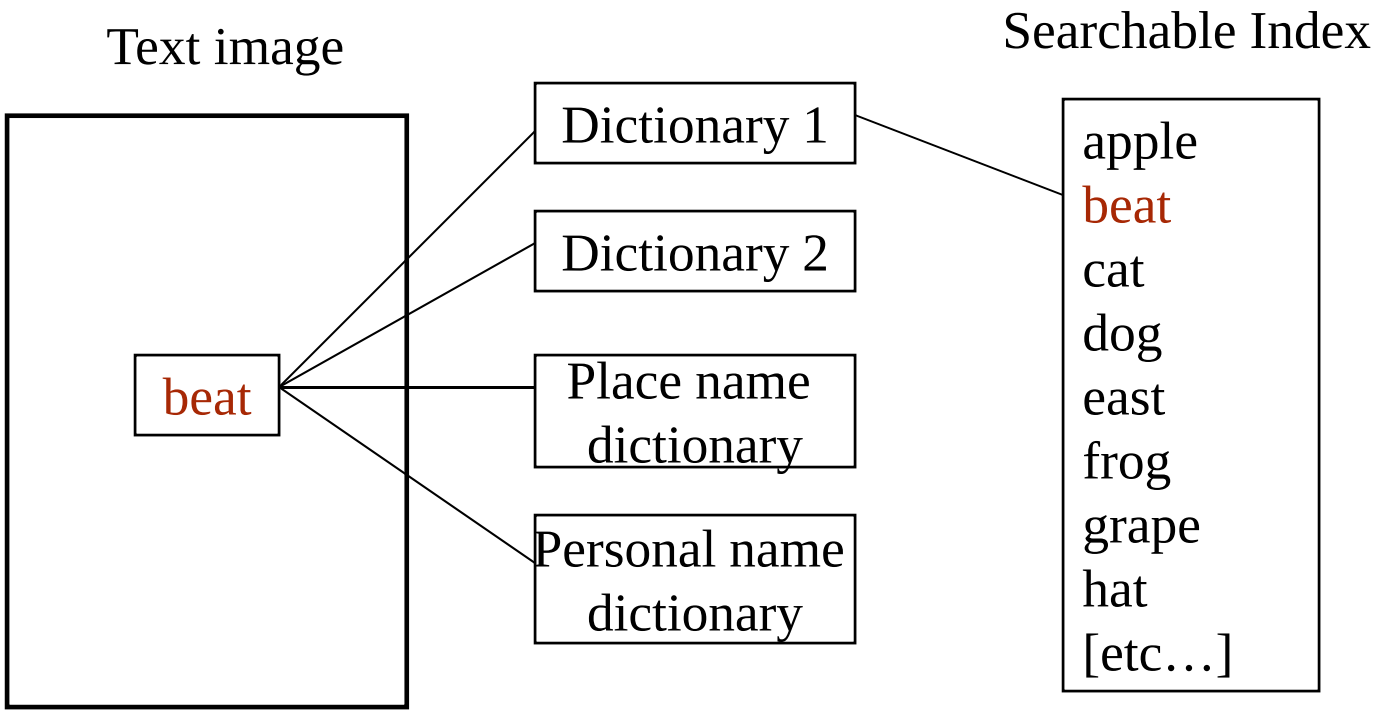
Can put two (or more) words in the searchable index



Ohio Historical Society

OCR and OWR





Ohio Historical Society

Digital Library (NDNP)

```
<?xml version="1.0" encoding="UTF-8" ?>
- <mets xmlns="http://www.loc.gov/METS/" xmlns:mix="http://www.loc.gov/mix/" xmlns:mods="http://www.loc.gov/mods/v3"
  xmlns:ndnp="http://www.loc.gov/ndnp" xmlns:premis="http://www.oclc.org/premis" xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" LABEL="National Forum, 1910-05-28" PROFILE="urn:library-of-
  congress:mets:profiles:ndnp:issue:v1.5" TYPE="urn:library-of-congress:ndnp:mets:newspaper:issue">
  <!-- METS HEADER -->
  - <metsHdr CREATEDATE="2005-09-06T17:25:40" RECORDSTATUS="Validated">
    - <agent ROLE="CREATOR" TYPE="ORGANIZATION">
        <name>Library of Congress</name>
      </agent>
    </metsHdr>
    <!-- DESCRIPTIVE METADATA -->
  - <dmdSec ID="issueModsBib">
    - <mdWrap LABEL="Issue metadata" MDTYPE="MODS">
      - <xmlData>
        - <mods:mods>
          - <mods:relatedItem type="host">
              <mods:identifier type="lccn">sn82015056</mods:identifier>
            - <mods:part>
              - <mods:detail type="volume">
                  <mods:number>I</mods:number>
                </mods:detail>
              - <mods:detail type="edition">
                  <mods:number>1</mods:number>
                </mods:detail>
              </mods:part>
            </mods:relatedItem>
          - <mods:originInfo>
              <mods:dateIssued encoding="iso8601">1910-05-28</mods:dateIssued>
            </mods:originInfo>
            <mods:note type="noteAboutReproduction">Present</mods:note>
          </mods:mods>
        </xmlData>
      </mdWrap>
```

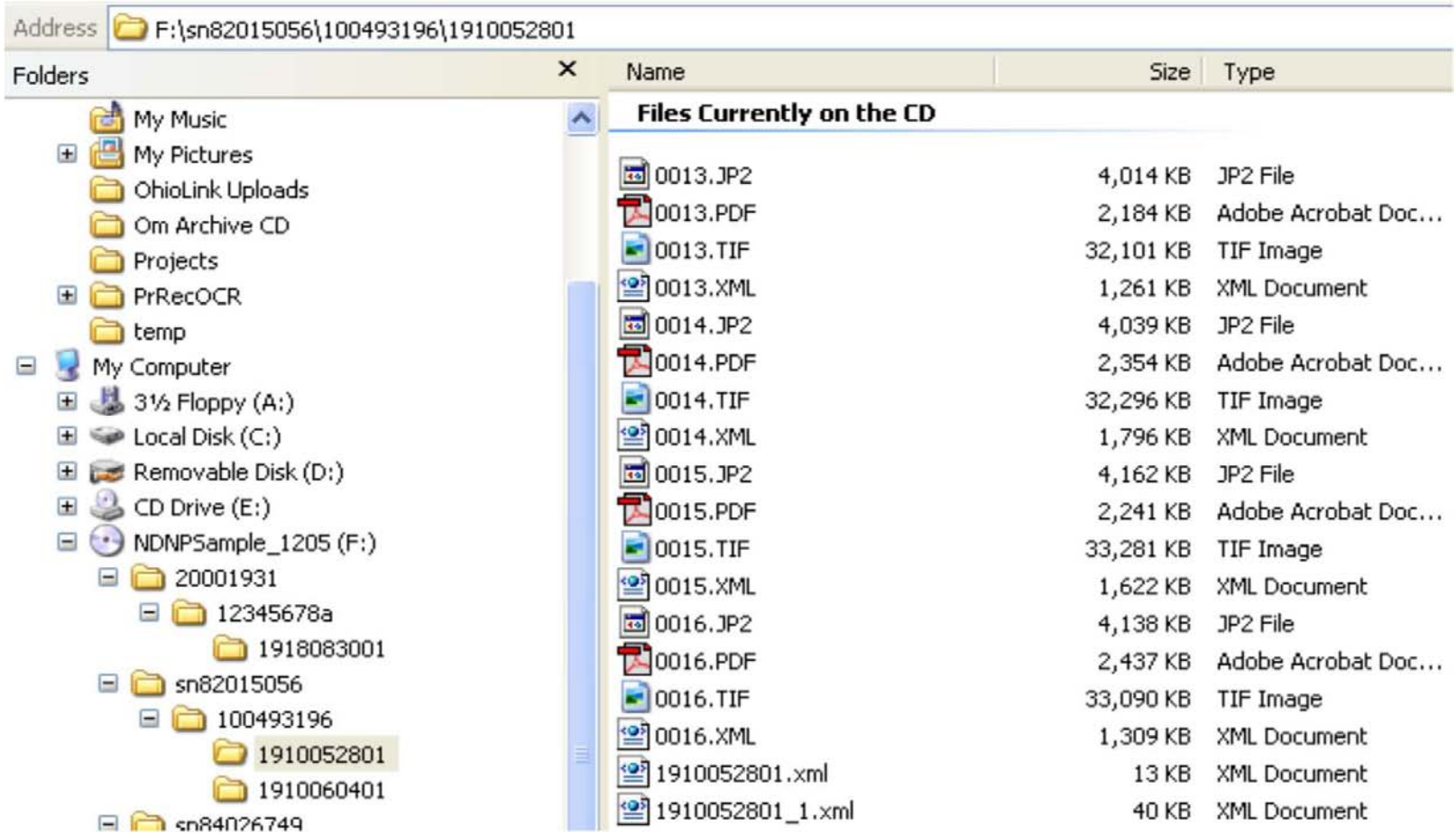
Ohio Historical Society

Digital Library (NDNP)

```
- <TextLine HEIGHT="144.0" WIDTH="1324.0" HPOS="4096.0" VPOS="10528.0">
    <String STYLEREFS="ID3" HEIGHT="104.0" WIDTH="364.0" HPOS="4096.0" VPOS="10556.0" CONTENT="duped"
      WC="0.95238096" />
    <String STYLEREFS="ID5" HEIGHT="72.0" WIDTH="156.0" HPOS="4096.0" VPOS="10584.0" CONTENT="elI"
      WC="0.95238096" />
    <SP WIDTH="-364.0" HPOS="4460.0" VPOS="10556.0" />
    <String STYLEREFS="ID5" HEIGHT="100.0" WIDTH="448.0" HPOS="4324.0" VPOS="10556.0" CONTENT="cl1lnto"
      WC="0.8095238" />
    <SP WIDTH="72.0" HPOS="4252.0" VPOS="10584.0" />
    <String STYLEREFS="ID3" HEIGHT="100.0" WIDTH="256.0" HPOS="4516.0" VPOS="10556.0" CONTENT="Into"
      WC="0.95238096" />
    <SP WIDTH="-256.0" HPOS="4772.0" VPOS="10556.0" />
  - <String STYLEREFS="ID3" HEIGHT="144.0" WIDTH="588.0" HPOS="4832.0" VPOS="10528.0" CONTENT="Hospitals"
      WC="0.95238096">
      <ALTERNATIVE>UOSllltals</ALTERNATIVE>
    </String>
    <SP WIDTH="60.0" HPOS="4772.0" VPOS="10556.0" />
  </TextLine>
- <TextLine HEIGHT="132.0" WIDTH="2188.0" HPOS="3740.0" VPOS="10720.0">
    <String STYLEREFS="ID5" HEIGHT="100.0" WIDTH="124.0" HPOS="3740.0" VPOS="10740.0" CONTENT="In"
      WC="0.96825397" />
    <String STYLEREFS="ID5" HEIGHT="104.0" WIDTH="204.0" HPOS="3952.0" VPOS="10732.0" CONTENT="the"
      WC="1.0" />
    <SP WIDTH="88.0" HPOS="3864.0" VPOS="10740.0" />
  - <String STYLEREFS="ID5" HEIGHT="128.0" WIDTH="592.0" HPOS="4248.0" VPOS="10724.0" CONTENT="operating"
      WC="0.95238096">
      <ALTERNATIVE>operatIns</ALTERNATIVE>
      <ALTERNATIVE>operations</ALTERNATIVE>
    </String>
    <SP WIDTH="92.0" HPOS="4156.0" VPOS="10732.0" />
    <String STYLEREFS="ID5" HEIGHT="80.0" WIDTH="380.0" HPOS="4924.0" VPOS="10748.0" CONTENT="rooms"
      WC="1.0" />
    <SP WIDTH="84.0" HPOS="4840.0" VPOS="10724.0" />
  - <String STYLEREFS="ID5" HEIGHT="100.0" WIDTH="128.0" HPOS="5388.0" VPOS="10720.0" CONTENT="ot"
      WC="0.82539684">
      <ALTERNATIVE>of</ALTERNATIVE>
```

Ohio Historical Society

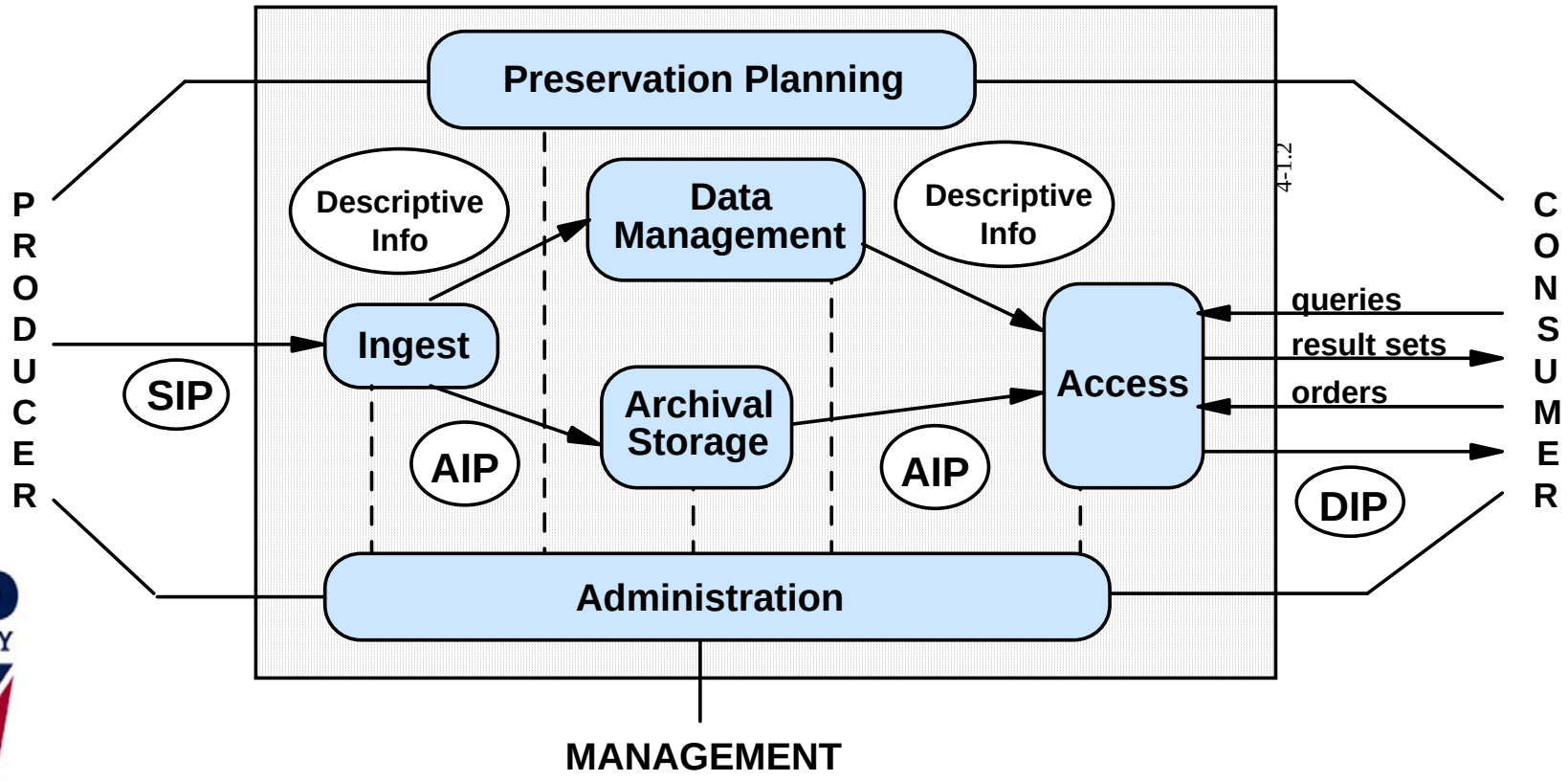
Digital Library (NDNP)



OHIO HISTORY

Ohio Historical Society

Digital Library Architecture



Ohio Historical Society

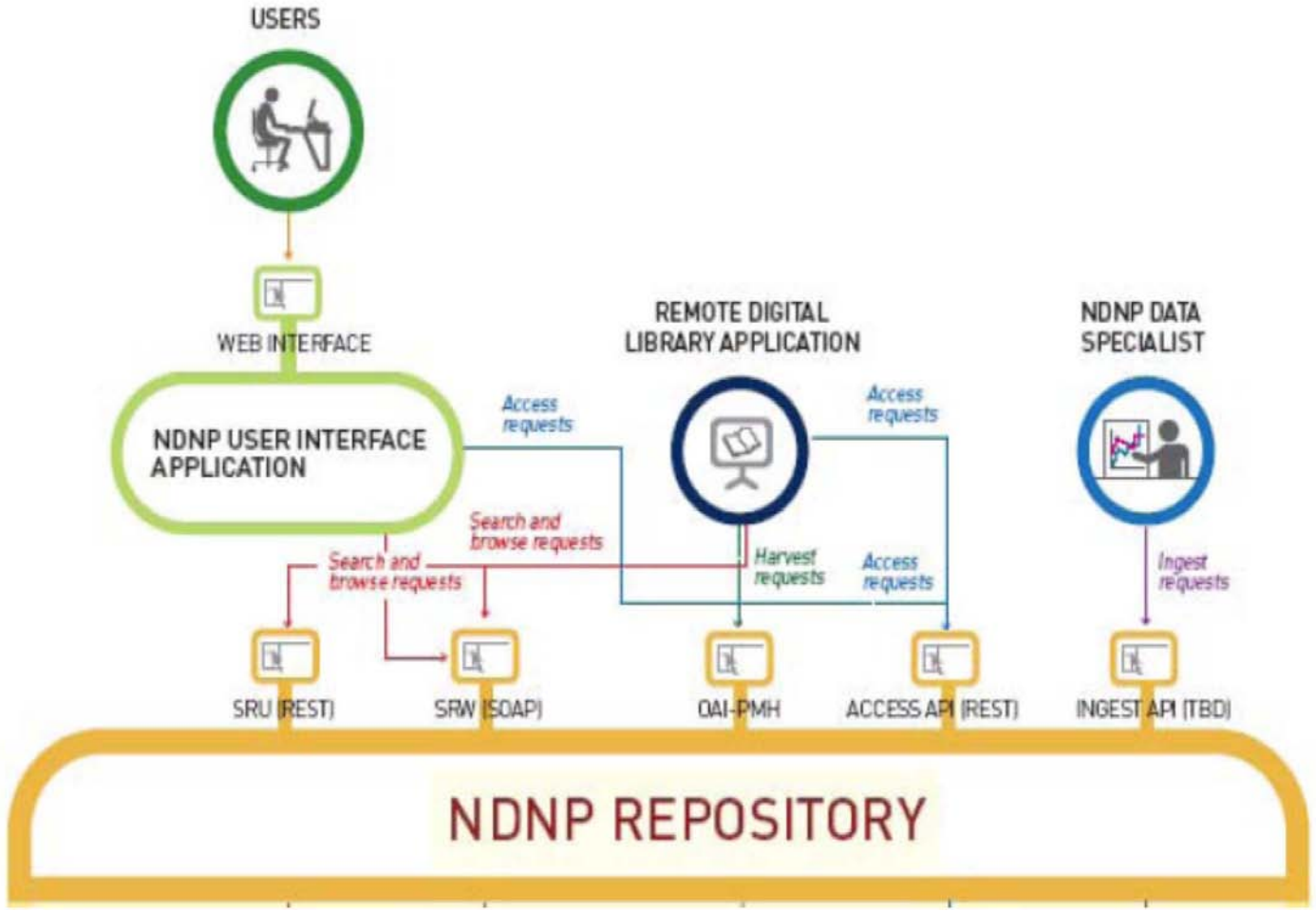
Digital Libraries

- ContentDM
- Olive (ActivePaper platform)
- Greenstone
- Fedora
- DSpace
- “Homegrown”

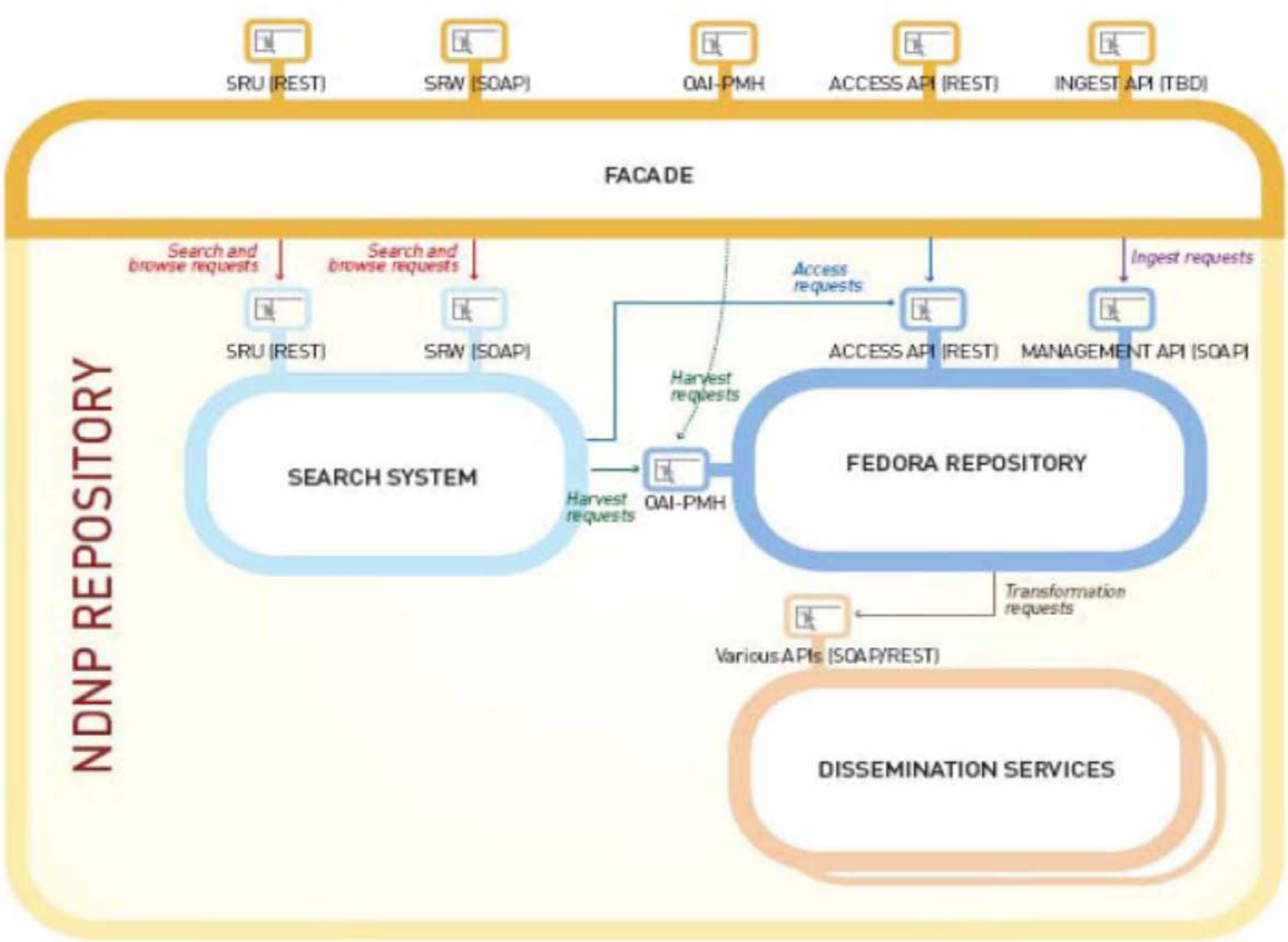
NDNP Digital Library (LoC)

- Gentoo Linux (operating system)
- Fedora (digital repository)
- Tomcat (servlet container)
- Cocoon (web application framework)
- Lucene (index/search engine)
- MySQL (database)
- Apache (web server)
- JDK 1.5 (java environment)

NDNP Digital Library



NDNP Digital Library



OHIO HISTORY

Ohio Histori

The End

Phil Sager
OHS

OHIO HISTORY

Ohio Historical Society