

Newspaper Digitization: Issues and Opportunities

Phil Sager
OHS



Ohio Historical Society



MISSING YOUTHS FOUND BY HERDER

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

After straying from a Mooseheart 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 Avintain 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 Mooseheart 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 Avintaquin 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

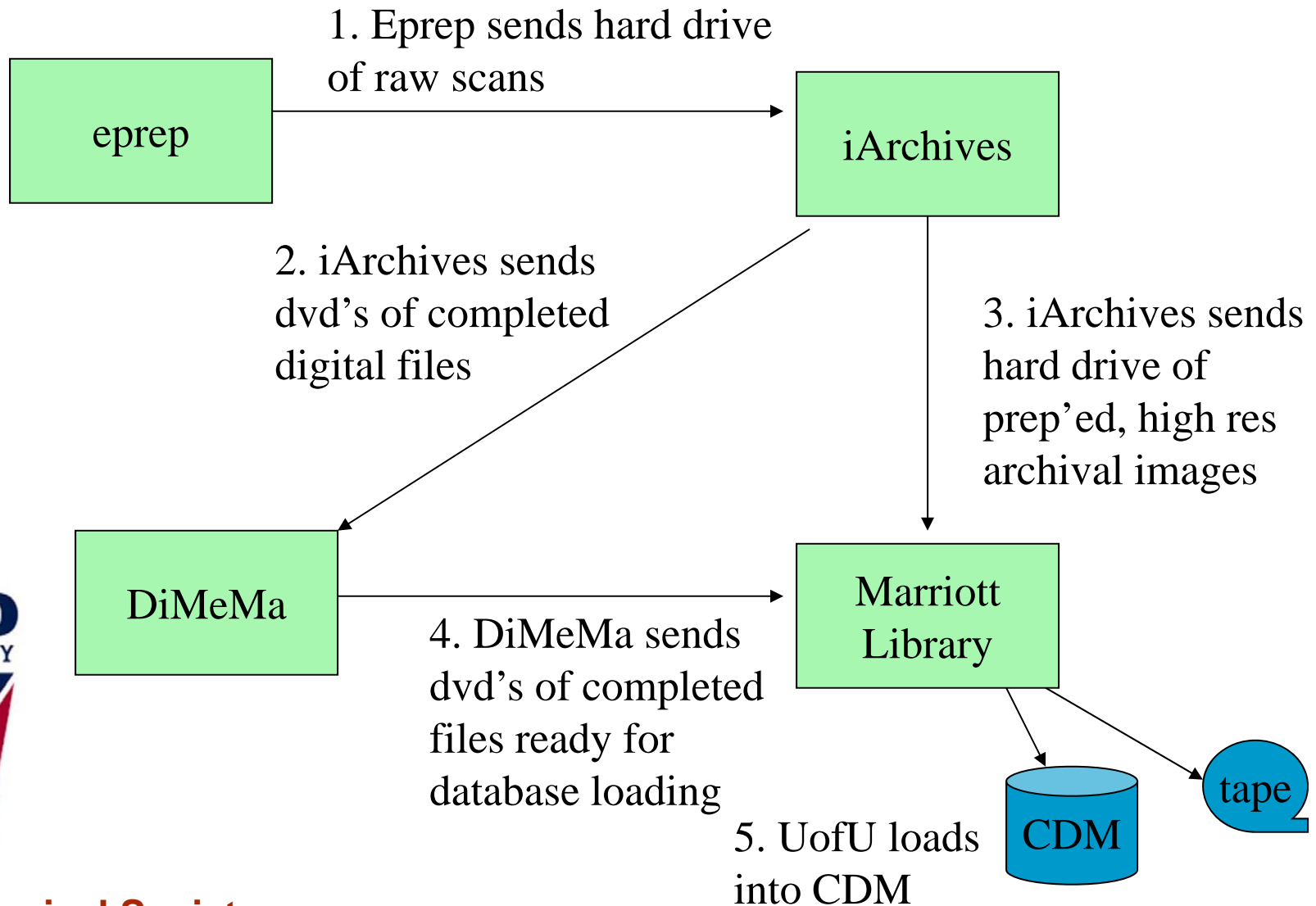
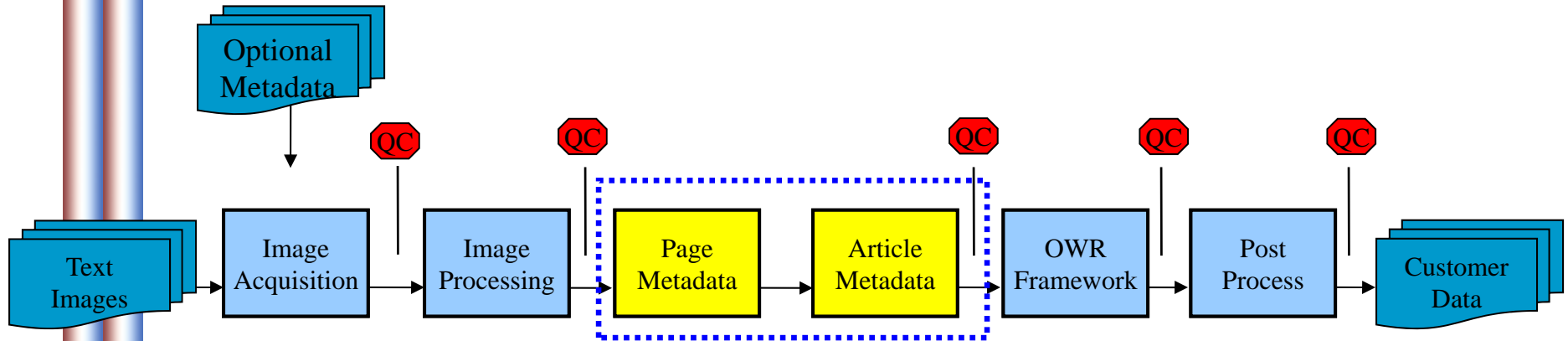
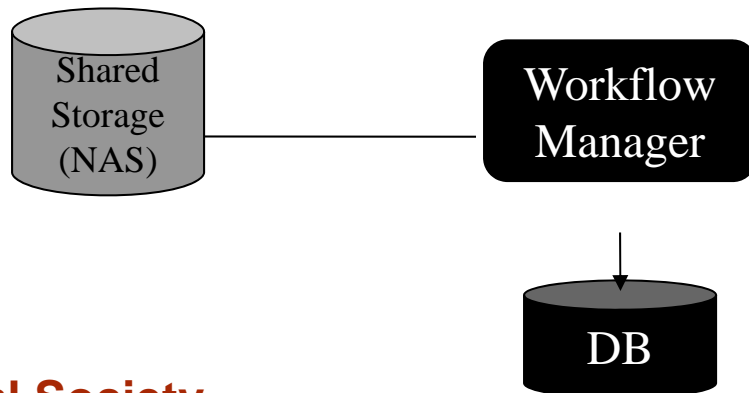


Image processing workflow

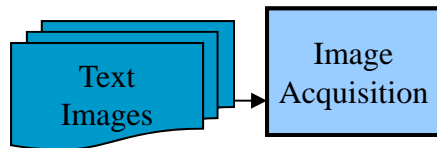


Ohio Historical Society



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

Image acquisition

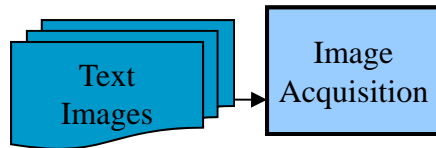


The quality of the original image has direct effect on every other aspect of the newspaper digitization process.

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



Image Acquisition

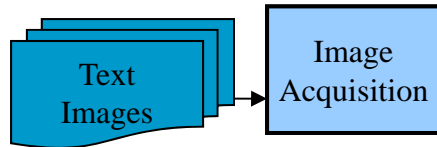


Selecting the **image type** is the first decision to be made when considering the viewing experience.

- 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)



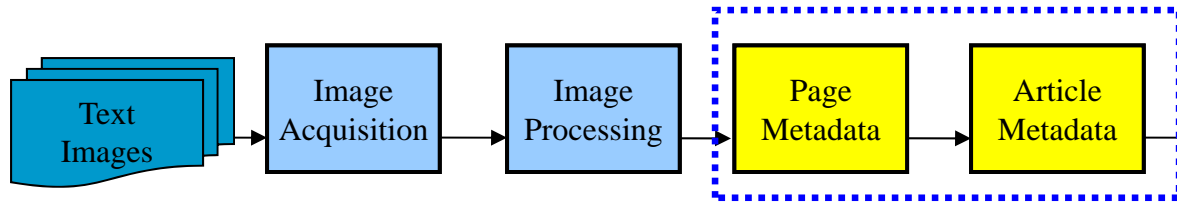
Image Processing



- 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



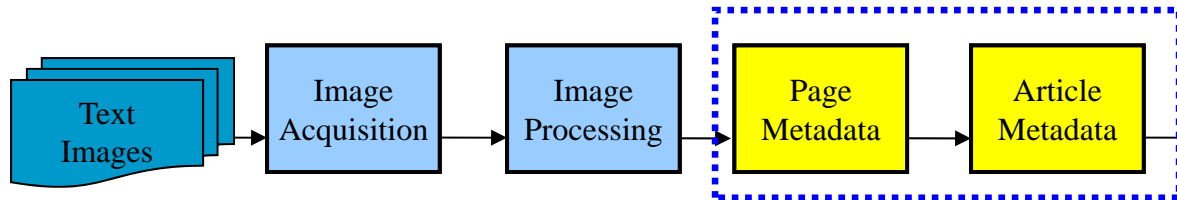
Metadata



- 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



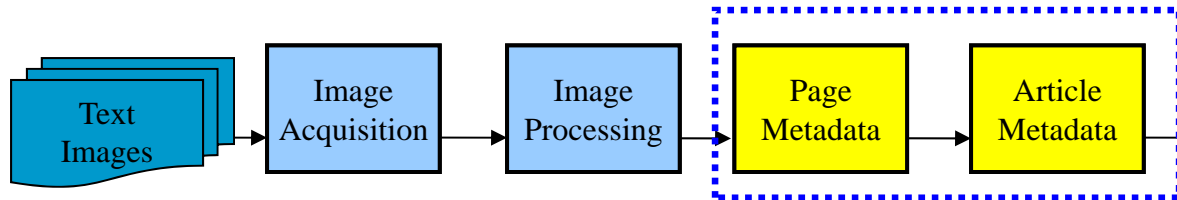
Metadata



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



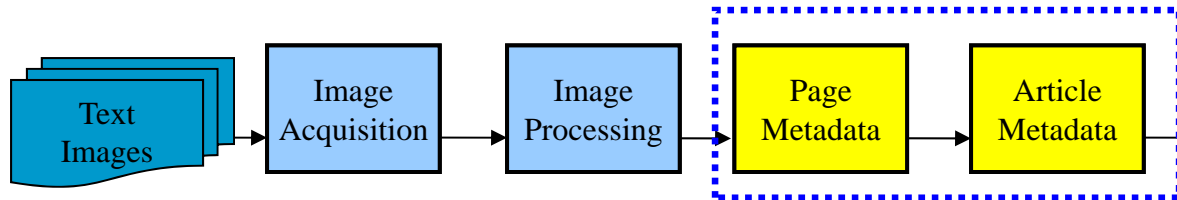
Metadata



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



Metadata



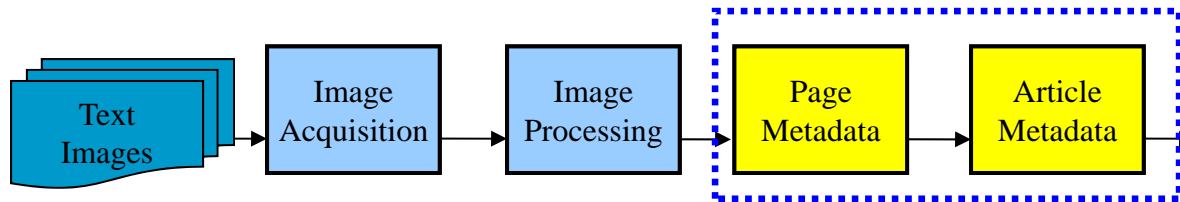
- Page level view

- 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.

The screenshot displays the HeaderMan application window. The title bar reads: "HeaderMan - /root/projects/production/newspaper/UofU/2004_2005/DeseretNews/iac_EMP34_36-18870119-18880111". The menu bar includes "File", "View", "Edit", "Tools", and "Help".

Image	Vol	Number	Date	Section	Page
p001n_va_q071_d45_v36	XXXXVI	1	19-Jan-1887, Wednesday		1
p002_834_va_q071_d45_v36	XXXXV	?	19-Jan-1887, Wednesday		2
p003_835_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		3
p004_836_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		4
p005_837_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		5
p006_838_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		6
p007_839_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		7
p008_840_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		8
p009n_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		9
p010_843_va_q071_d45_v36	XXXXVI	1	19-Jan-1887, Wednesday		10
p011_842_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		11
p011n_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		11a
p012_844_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		12
p013_845_va_q071_d45_v36	XXXXVI	?	19-Jan-1887, Wednesday		13
p014_846_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		14
p016_847_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		16
p016_848_va_q071_d45_v36	XXXXW	?	19-Jan-1887, Wednesday		16
p017n_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		17
p018_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		18
p019_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		19
p020_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		20
p021_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		21
p022_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		22
p023_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		23
p024_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		24
p025_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		25
p026_27_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		26
p027_28_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		27
p028_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		28
p029_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		29
p030_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		30
p031_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		31
p032_va_q071_d45_v36	XXXXW	2	26-Jan-1887, Wednesday		32

Below the table is a preview of a newspaper page. The masthead reads "THE DESERET NEWS" with the motto "TRUTH AND LIBERTY." below it. The page is identified as "No 1", "Salt Lake City, Wednesday, Jan. 19, 1887.", and "Vol. XXXVI". The text on the page is partially legible, mentioning "ESTABLISHED 1856. DESERET NEWS: WEEKLY. PUBLISHED EVERY WEDNESDAY." and "during the night, committed upon young Jenkins an unmentionable and beastly crime. The names of the reprobates are: Willie Padlock, Arthur Curtis, John Legadford, Richard Bubbles and Dan Henry. It seems that carp question, with an enclosure. He says: Please give place to the enclosed letter from Joseph Adams, Esq., of Meadows, Millard Co., in which he ment. The company assembled at 7 o'clock in the afternoon and remained still near 10 in the evening. They were regaled with plenty of good things, and with social conversation, songs, recitations, and briefs. The believer and the infidel cannot be indicted or convicted on account of religious beliefs; nor can you deprive an AMERICAN CITIZEN."

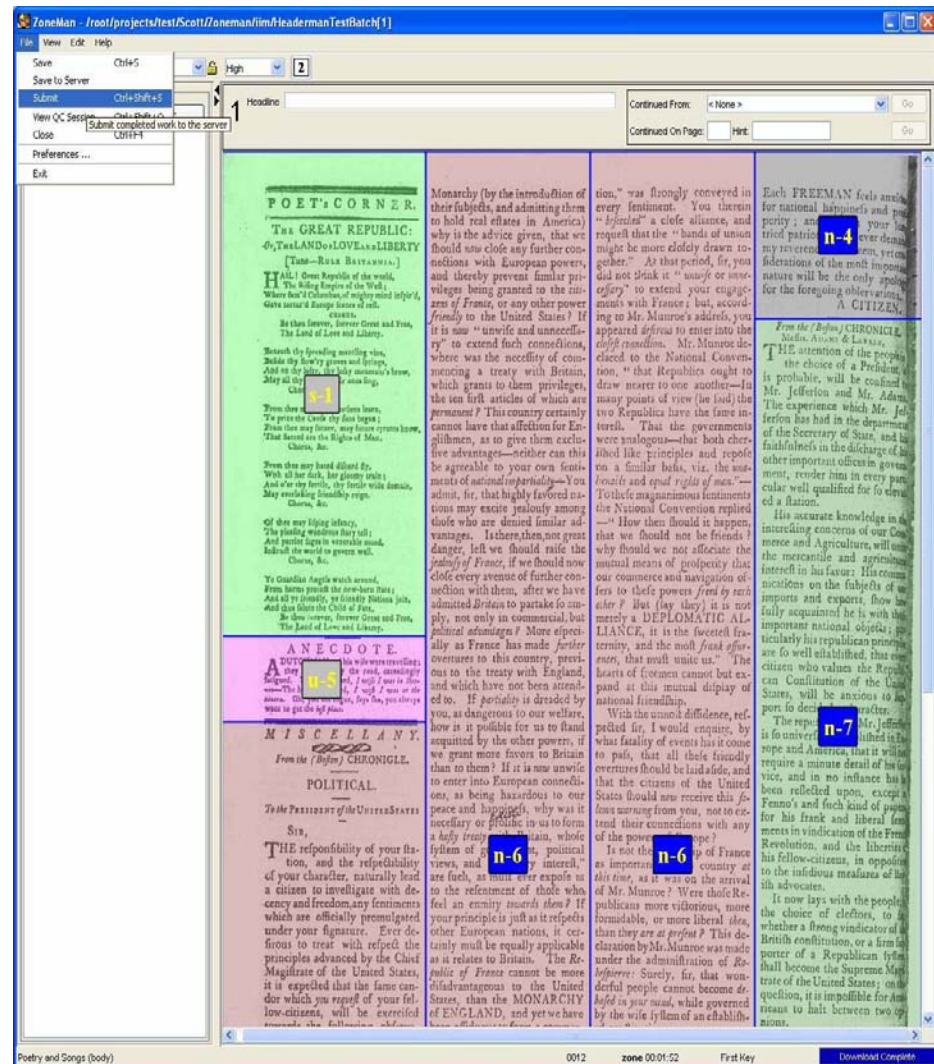


Tool Set - ZoneMan

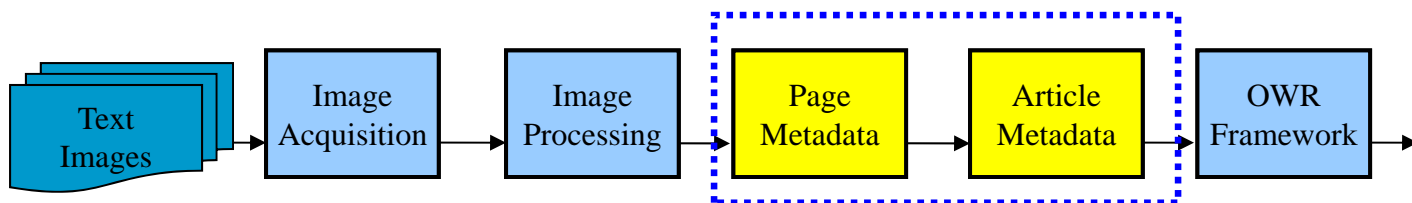
ZoneMan is a flexible, efficient means to collect customized article-level meta data.



Ohio Historical Society



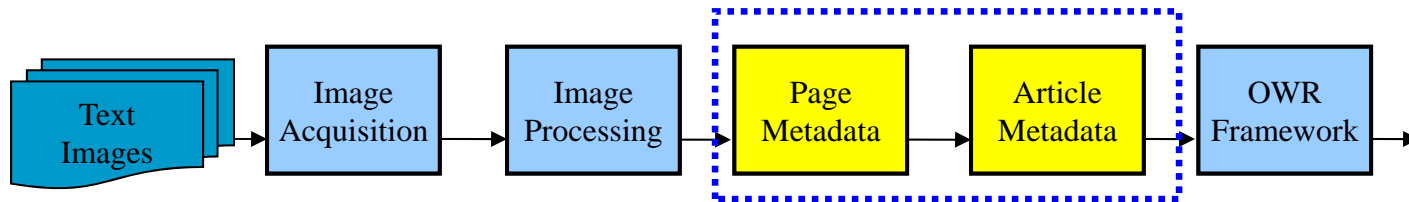
OCR and OWR



- 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**

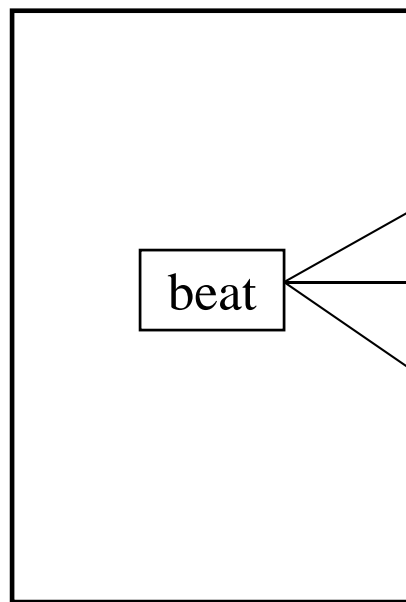


OCR and OWR



Text image

word (predicted accuracy)



OCR Engine 1

boat (73%)

OCR Engine 2

beat (90%)

OCR Engine 3

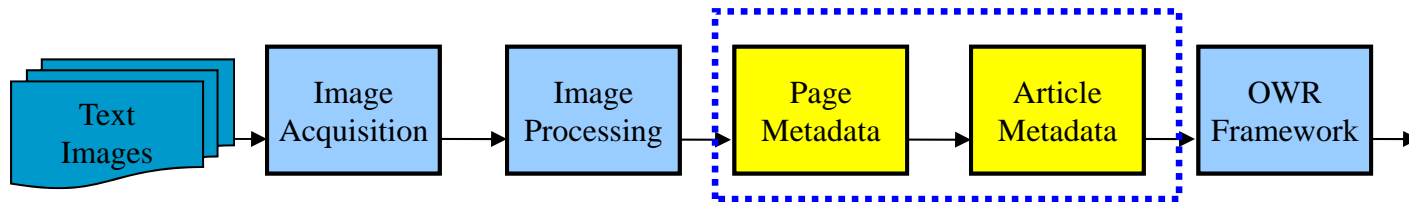
beet (86%)

Index

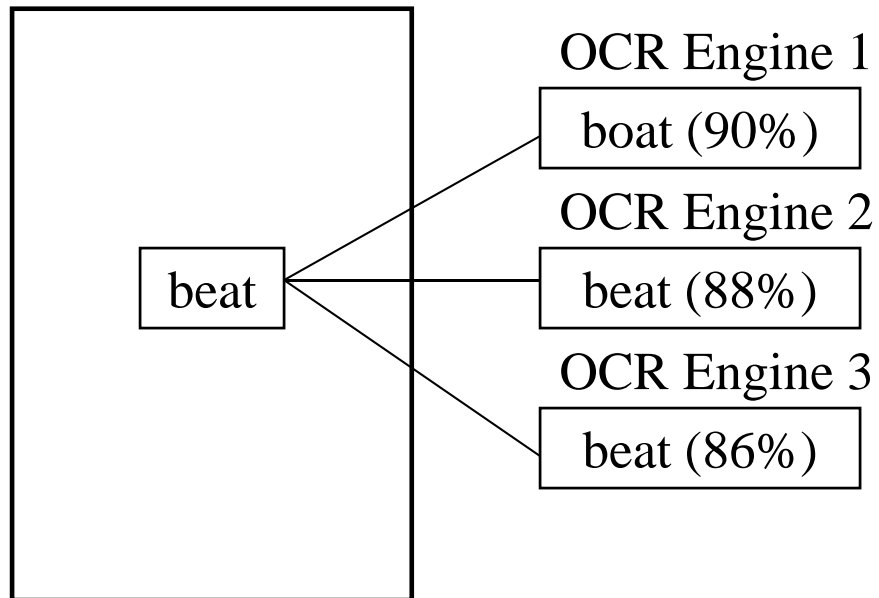
apple
beat
cat
dog
east
frog
grape
[etc...]



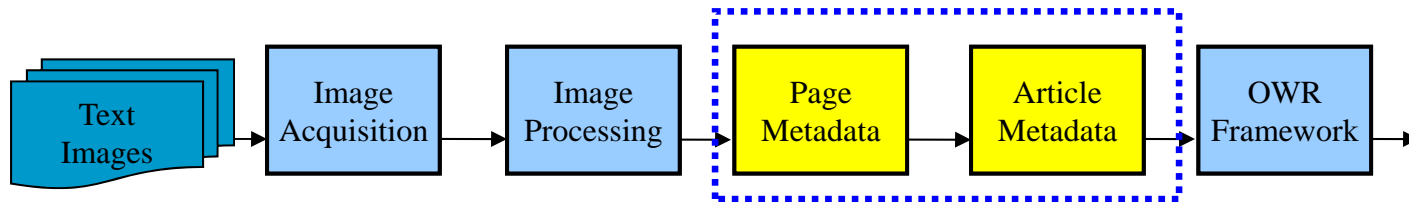
OCR and OWR



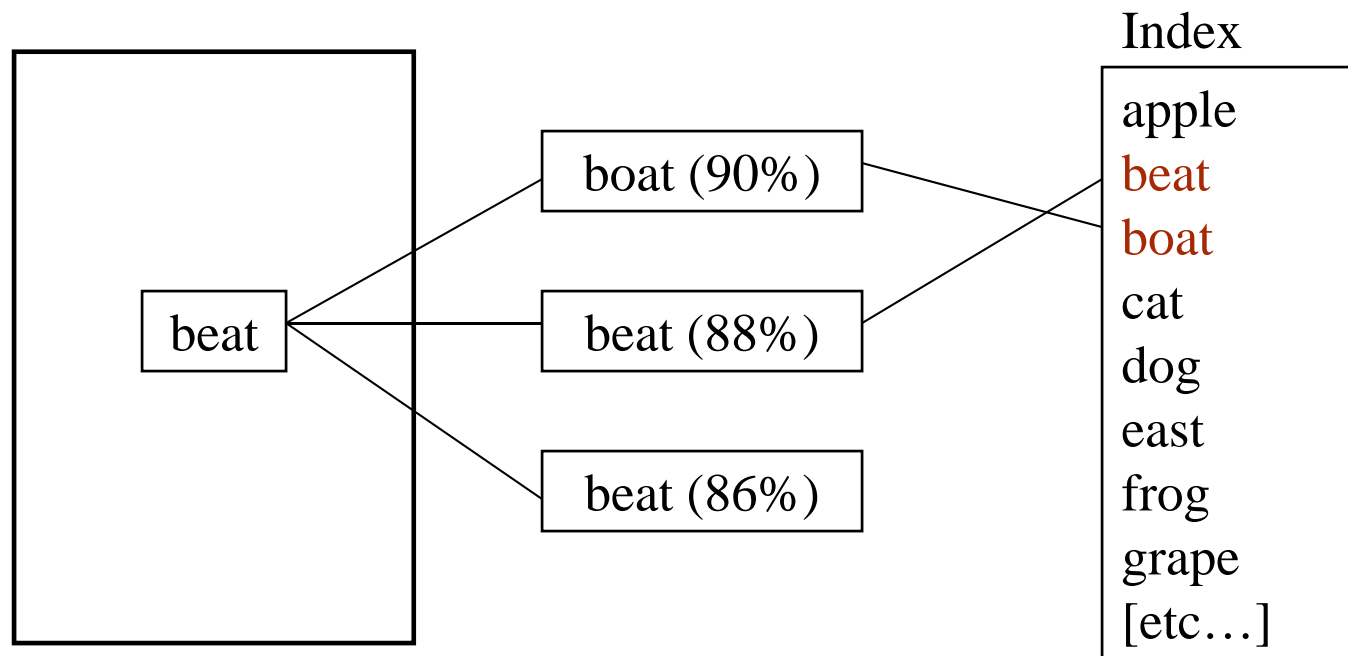
What if there are OWR ties?



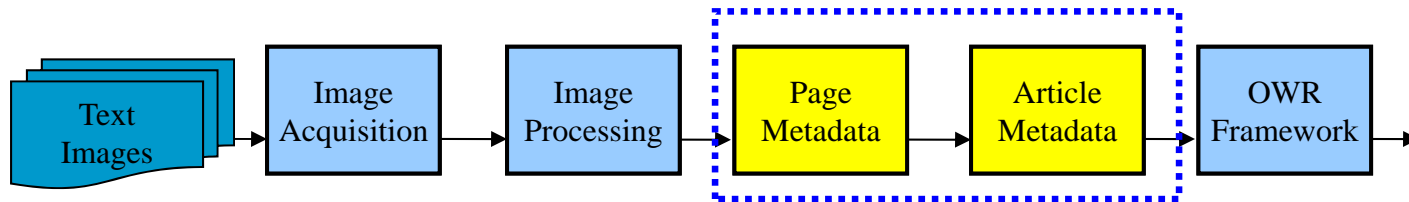
OCR and OWR



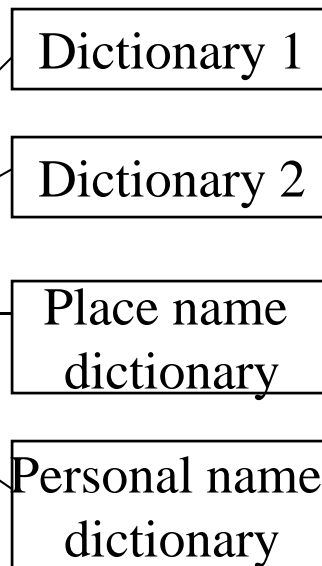
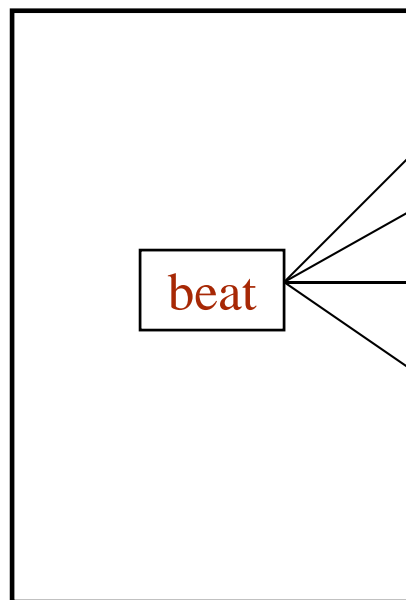
Can put two (or more) words in the searchable index



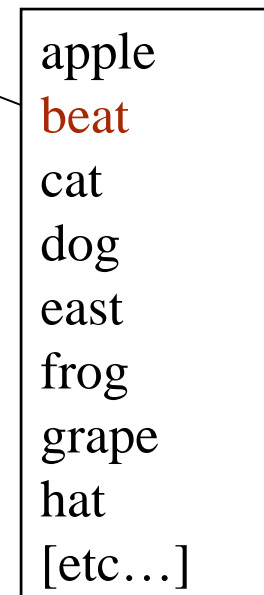
OCR and OWR



Text image



Searchable Index



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 STYLEREF="ID3" HEIGHT="104.0" WIDTH="364.0" HPOS="4096.0" VPOS="10556.0" CONTENT="duped"
    WC="0.95238096" />
  <String STYLEREF="ID5" HEIGHT="72.0" WIDTH="156.0" HPOS="4096.0" VPOS="10584.0" CONTENT="ell"
    WC="0.95238096" />
  <SP WIDTH="-364.0" HPOS="4460.0" VPOS="10556.0" />
  <String STYLEREF="ID5" HEIGHT="100.0" WIDTH="448.0" HPOS="4324.0" VPOS="10556.0" CONTENT="cl1Into"
    WC="0.8095238" />
  <SP WIDTH="72.0" HPOS="4252.0" VPOS="10584.0" />
  <String STYLEREF="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 STYLEREF="ID3" HEIGHT="144.0" WIDTH="588.0" HPOS="4832.0" VPOS="10528.0" CONTENT="Hospitals"
  WC="0.95238096">
  <ALTERNATIVE>UOSlltals</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 STYLEREF="ID5" HEIGHT="100.0" WIDTH="124.0" HPOS="3740.0" VPOS="10740.0" CONTENT="In"
    WC="0.96825397" />
  <String STYLEREF="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 STYLEREF="ID5" HEIGHT="128.0" WIDTH="592.0" HPOS="4248.0" VPOS="10724.0" CONTENT="operating"
  WC="0.95238096">
  <ALTERNATIVE>operatlns</ALTERNATIVE>
  <ALTERNATIVE>operations</ALTERNATIVE>
</String>
<SP WIDTH="92.0" HPOS="4156.0" VPOS="10732.0" />
<String STYLEREF="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 STYLEREF="ID5" HEIGHT="100.0" WIDTH="128.0" HPOS="5388.0" VPOS="10720.0" CONTENT="ot"
  WC="0.82539684">
  <ALTERNATIVE>of</ALTERNATIVE>
```



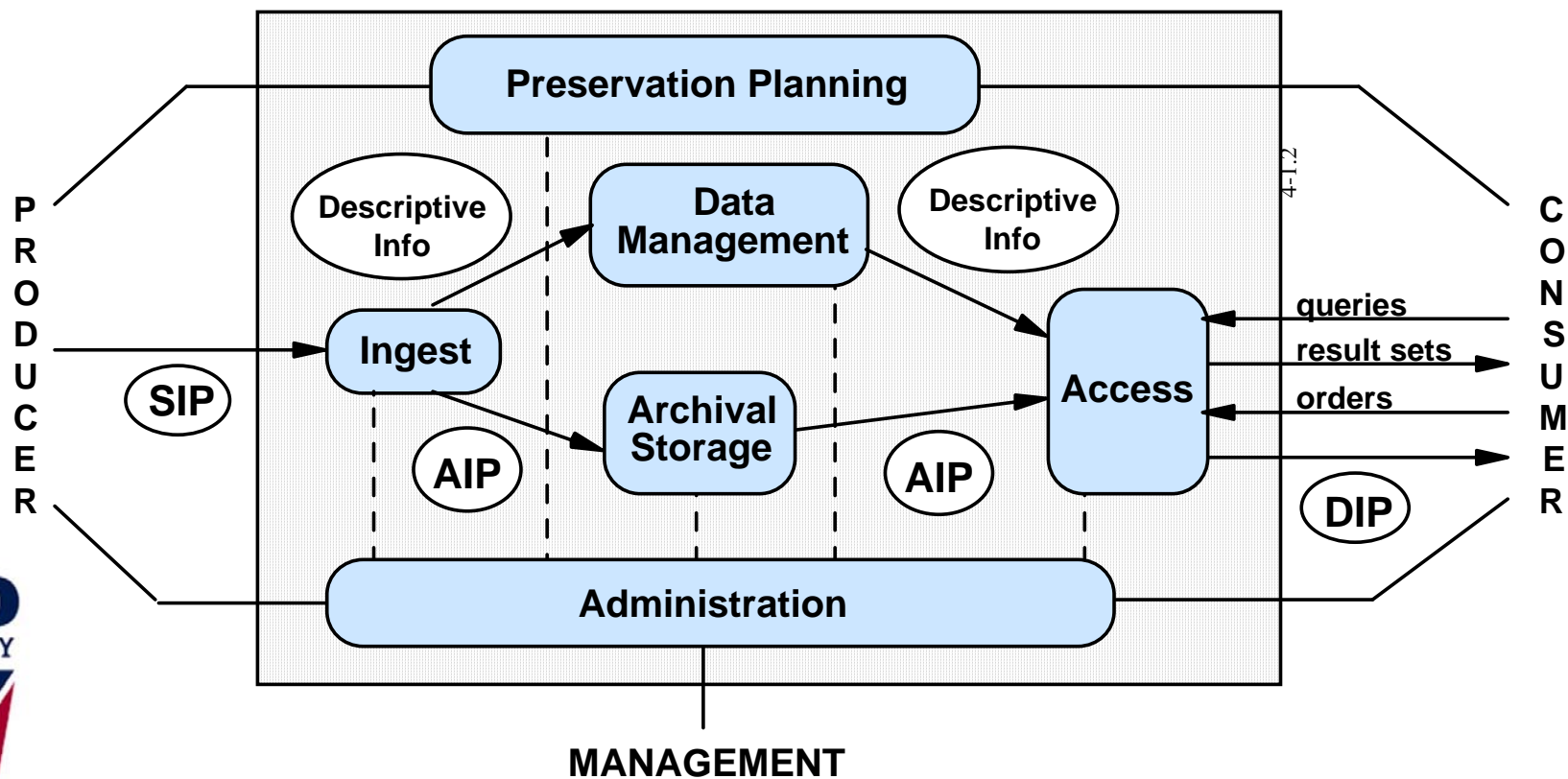
Digital Library (NDNP)

Address  F:\sn82015056\100493196\1910052801

Folders	Name	Size	Type
 My Music	Files Currently on the CD		
 My Pictures	 0013.JP2	4,014 KB	JP2 File
 OhioLink Uploads	 0013.PDF	2,184 KB	Adobe Acrobat Doc...
 Om Archive CD	 0013.TIF	32,101 KB	TIF Image
 Projects	 0013.XML	1,261 KB	XML Document
 PrRecOCR	 0014.JP2	4,039 KB	JP2 File
 temp	 0014.PDF	2,354 KB	Adobe Acrobat Doc...
 My Computer	 0014.TIF	32,296 KB	TIF Image
 3½ Floppy (A:)	 0014.XML	1,796 KB	XML Document
 Local Disk (C:)	 0015.JP2	4,162 KB	JP2 File
 Removable Disk (D:)	 0015.PDF	2,241 KB	Adobe Acrobat Doc...
 CD Drive (E:)	 0015.TIF	33,281 KB	TIF Image
 NDNPSample_1205 (F:)	 0015.XML	1,622 KB	XML Document
 20001931	 0016.JP2	4,138 KB	JP2 File
 12345678a	 0016.PDF	2,437 KB	Adobe Acrobat Doc...
 1918083001	 0016.TIF	33,090 KB	TIF Image
 sn82015056	 0016.XML	1,309 KB	XML Document
 100493196	 1910052801.xml	13 KB	XML Document
 1910052801	 1910052801_1.xml	40 KB	XML Document
 1910060401			
 sn84026749			



Digital Library Architecture





Digital Libraries

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



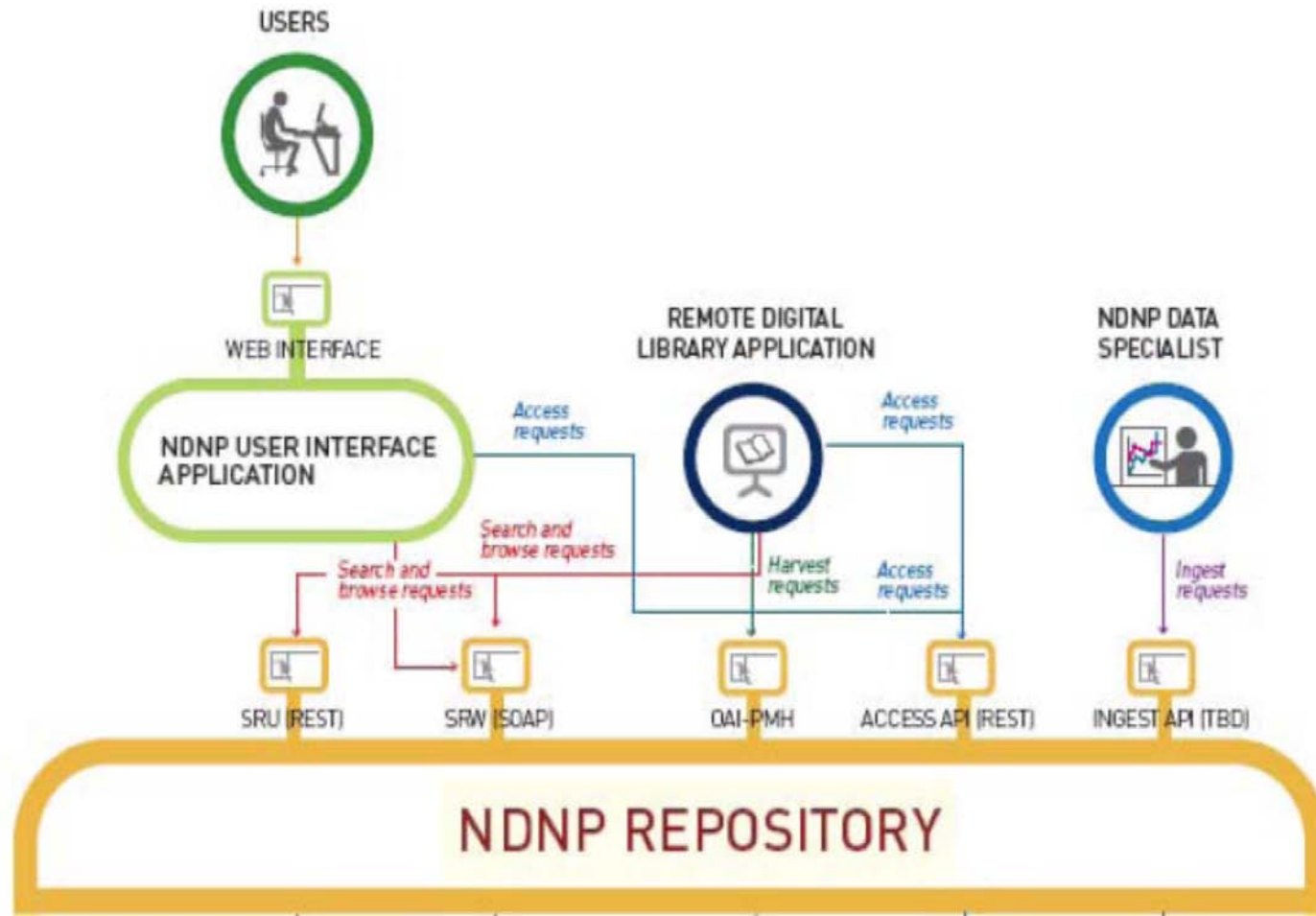


NDNP Digital Library (LoC)

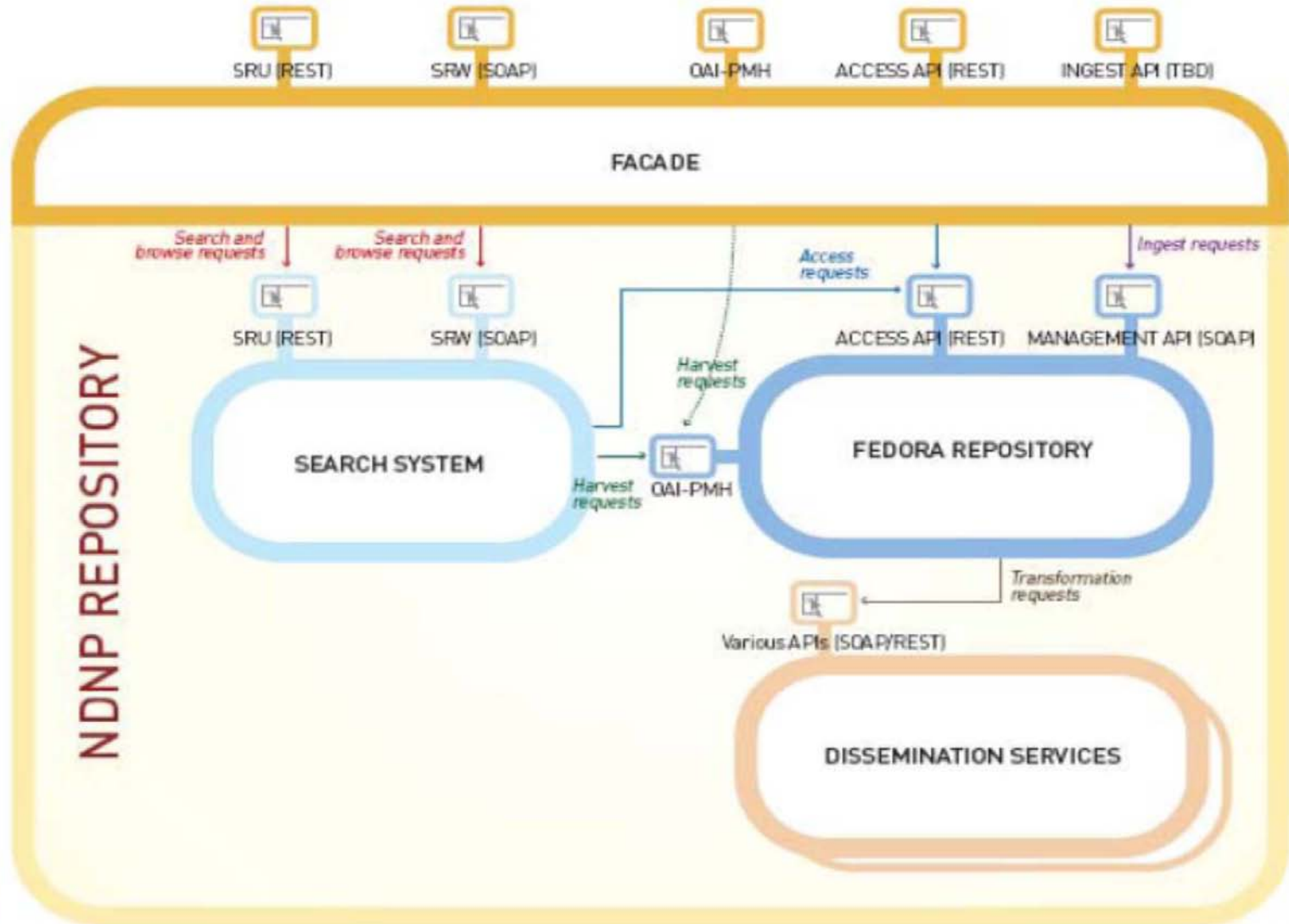
- Gentoo Linux (operating system)
- Fedora (digital repository)
- Tomcat (servlet container)
- Cocoon (web application framework)
- Lucene (index/search engine)
- MySQL (database)
- Apache (web server)
- JDK 1.5 (java environment)



NDNP Digital Library



NDNP Digital Library



The End

Phil Sager
OHS



Ohio Historical Society