Digital Newspapers Project Handbook

Table of Contents

Table of Contents

Contact Information

- Kenning Arlitsch kenning.arlitsch@library.utah.edu
- John Herbert john.herbert@library.utah.edu
- Jeff Jonsson jeff.jonsson@library.utah.edu
- Randy Silverman randy.silverman@library.utah.edu
- Scott Christensen schristensen@iarchives.com
- Dave Morrison dave.morrison@library.utah.edu
- Jill Koelling jill.koelling@du.edu
- Frederick Zarndt frederick.zarndt@iarchives.com

Early History of the UDN

Brief Chronology

- 1951 – U of U microfilming Utah newspapers
- 1983 – USNP grant award from NEH
- 2000 – Established Digitization Center
  - Several successful digital projects
    - Maps, photographs, documents, books
- 2001 – LSTA R&D grant for newspapers
- 2003 – Second LSTA grant
- 2003 – IMLS grant

First LSTA Grant

- Proposal:
  - R&D newspaper digitization
  - Benefit to entire state
  - Digitize three weekly titles – ~30 yrs each
    - Vernal Express
    - Grand Valley Times/Times Independent
    - Wasatch Wave
  - 30,000 pages total
- Awarded $93K in fall 2001

Vernal Express Digitization

- Scanning
  - Microfilm scanned by FutureVision Technologies Inc.
    - Returned 1-bit TIFF files
- U of U Processing
  - TIFF files cropped, converted to MrSID®
  - Loaded into CONTENTdm®
  - Uintah County index imported
- Results
  - Searching by index terms (no full text)
  - Initial JPEG full page display
  - Download of full-page SID file allowed zooming

Wasatch Wave/GVT Digitization

- iArchives Inc.
  - Scanned microfilm
  - Segmented articles
  - OCR'd full text
  - Keyed headlines, article types
  - Created PDF for display, TIFF for archive
  - Generated XML metadata tagging
- DiMeMa Inc.
  - Loaded XML files and PDFs into CONTENTdm
  - Delivered finished collections to UU
- University of Utah
  - Copied collections to CONTENTdm server
  - Created Web pages
  - Generated search and browse queries

Lessons Learned

- Some microfilm was terrible
  - Bad lighting, focus, cropping, smudges,etc.
- Some master microfilm in El Paso
  - Result of sale of one company to another
- Vernal Express method not scalable
  - Too labor-intensive
- Needed more funding, project manager
- People loved this project!

Second LSTA grant

- Proposal
  - Improve digitization process
  - 100,000 more pages
  - Break new ground by scanning from paper
  - Hire full-time project manager
  - Apply for federal funding
  - Begin publicity campaign
- Unprecedented community support
  - $100K matching funds raised:
    - Utah Academic Library Consortium
    - Public libraries
- Awarded $278,000 in November 2002

Baptism by Fire - 2003

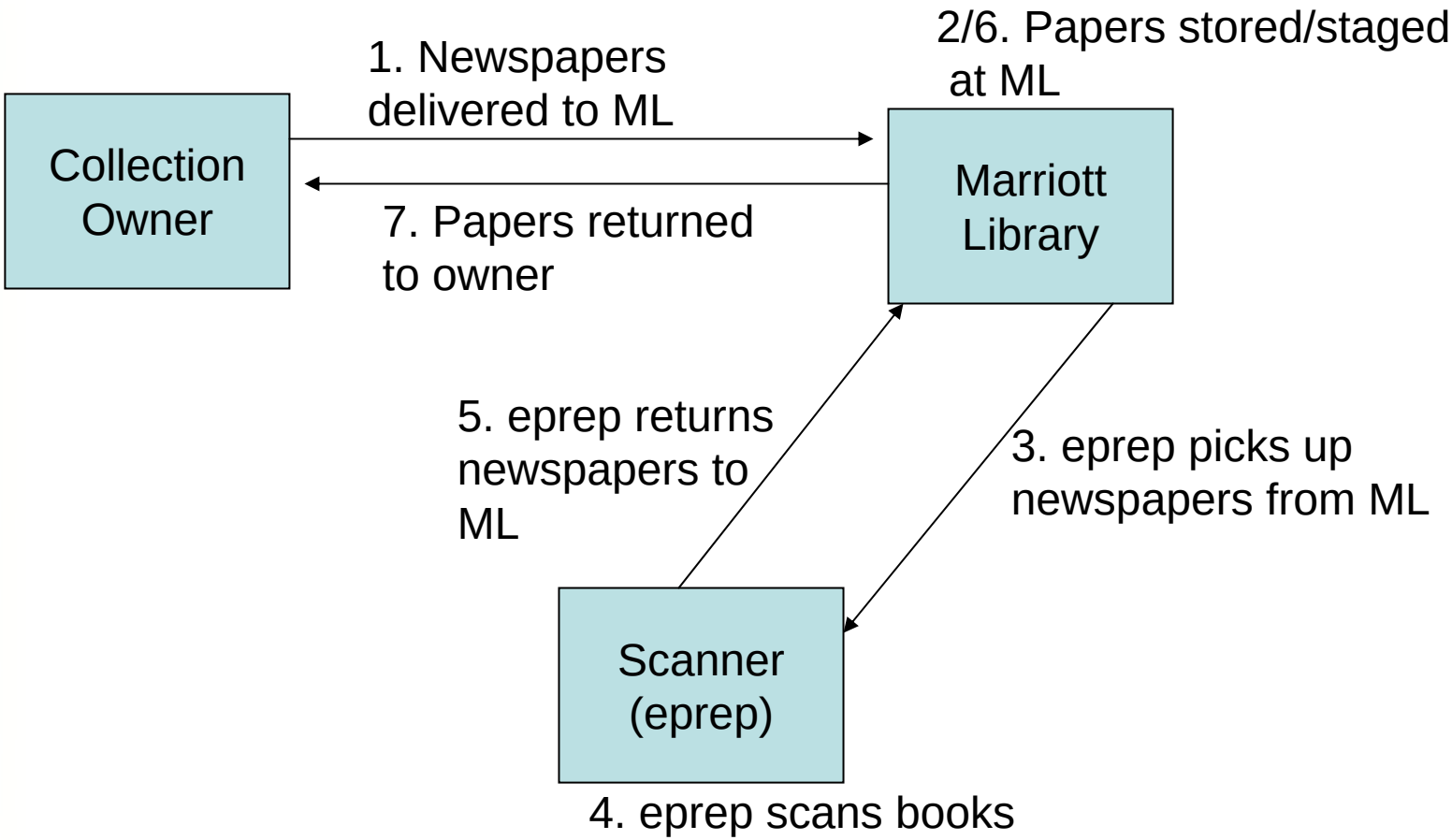
- January 5 – John Herbert started as PM
- February 1 – IMLS grant proposal
- John assumed management
- Advisory Board formed
- Website redesigned
- 65% newspapers scanned from paper
- September 23 – IMLS awarded

UDN Overview

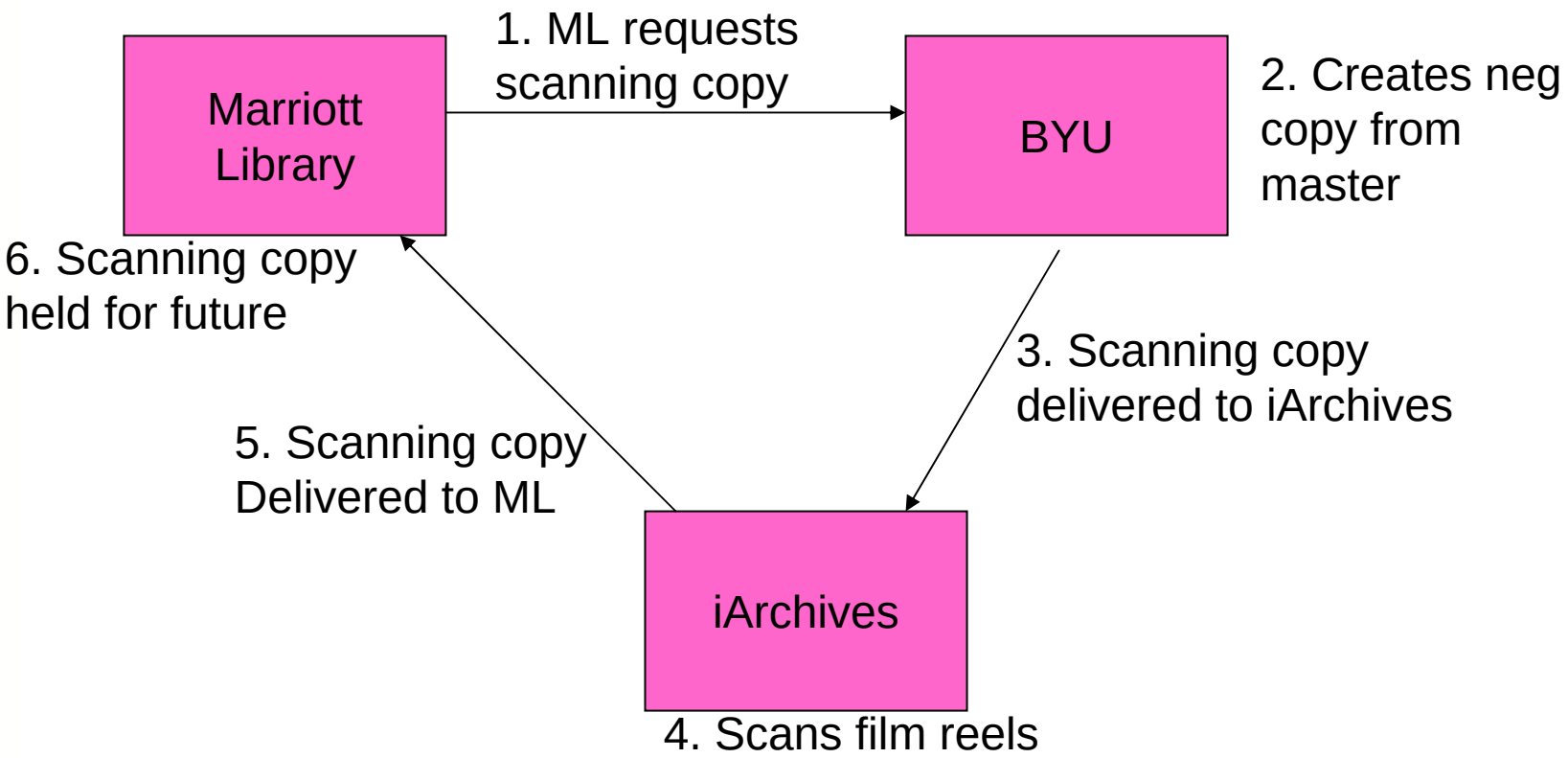
UDN High-Level Process

• 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	UofU

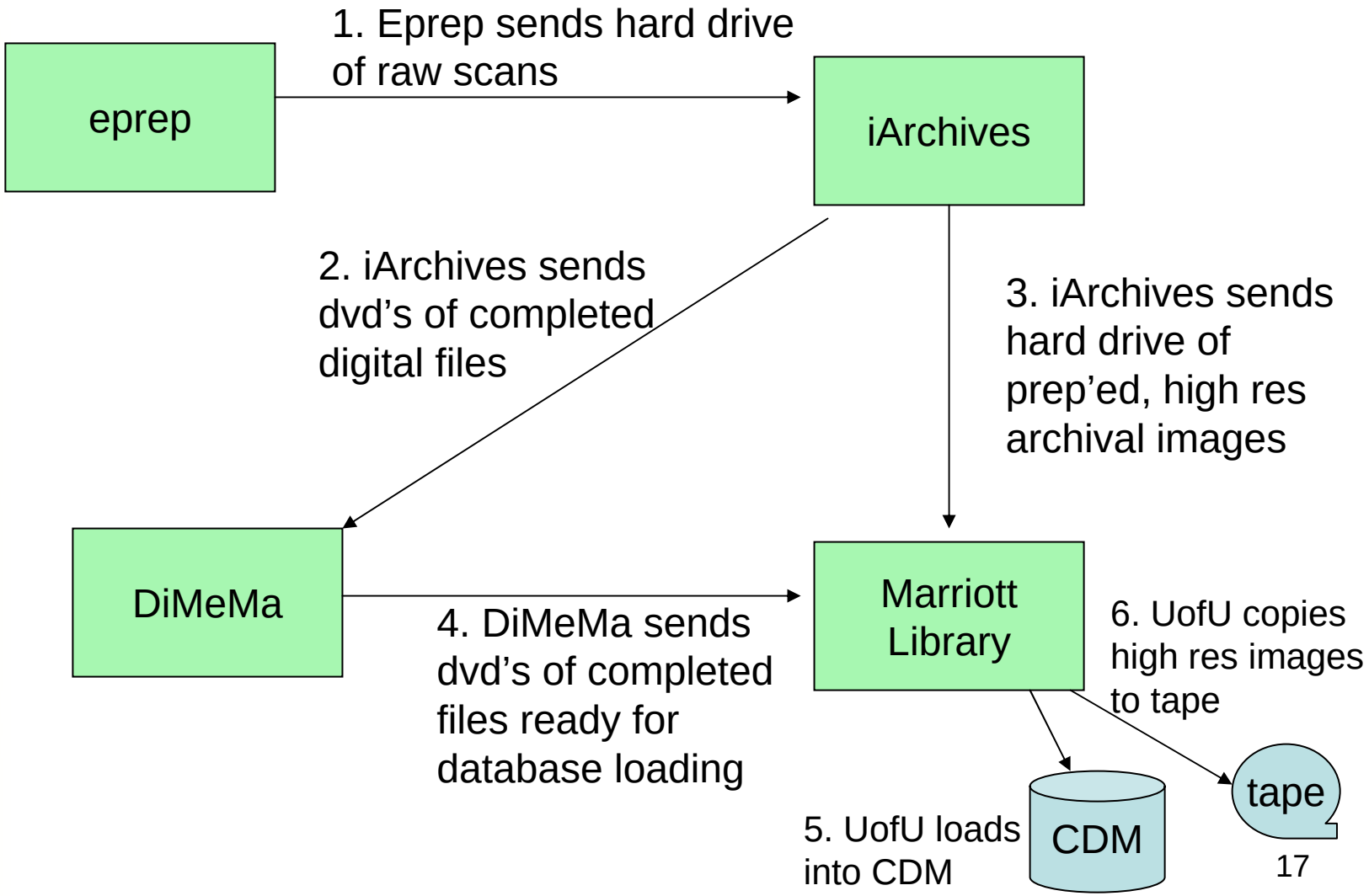
Flow of Original Papers



Flow of Microfilm



Flow of Data


eprep
1. Eprep sends hard drive of raw scans
iArchives
2. iArchives sends dvd's of completed digital files
3. iArchives sends hard drive of prep'ed, high res archival images
DiMeMa
4. DiMeMa sends dvd's of completed files ready for database loading
Marriott Library
6. UofU copies high res images to tape
tape
5. UofU loads into CDM
CDM

Project Setup

Getting Organized

- Gather interested, impacted parties
  - This could be a long list of people
  - Get buy-in up front from as many as possible
- Outline high-level objectives, strategies, roles
  - There can be support roles for institutions not in lead role
- Determine lead organization
  - Hopefully an obvious choice emerges
  - Will need to provide/apply resources to the project
- Lead institution
  - Write, submit grant proposal(s)
  - Secure matching funds
  - Form Advisory Board

Project Staffing

- Project Manager
  - Set direction and priorities
  - Run day-to-day activities
    - Manage work flow (and vendors)
    - Make financial decisions according to grant guidelines
  - Build/establish relationships
    - Publishers, libraries, historical societies, funding agencies
  - Liaise with Advisory Board and granting agencies
  - Networking and collaboration skills
  - Project and vendor management
  - The bigger your project, the stronger the manager required
  - UDN has full-time project manager

Project Staffing

- Web Developer
  - Design and develop website
  - UDN has full-time developer
- Systems Administrator
  - Manage system resources
    - CONTENTdm software and servers
  - UDN has 20% of systems admin
- Support Services
  - Accounting, Purchasing, Accts Payable, HR, Development, Public Relations, Sponsored Projects, Office of Research
  - Used as needed
  - No direct charge to project; allocated in F&A
- Vendors
  - Scanning, Zoning, Keying, OCR, Database indexing

Advisory Board

- Provide high-level (policy) direction and support
  - Content selection
  - Website
  - Funding opportunities
  - Publicity
- Aim high
  - You want influential, respected people
  - Use them to expand your network
- Diverse points of view
  - Avoid “group think”
  - Don’t just select from your own institution
- Geographic coverage
  - Especially if you have a state-wide initiative

Advisory Board

- Librarians – director, asst. director
  - Academic
  - Public
  - State Library
- Historians/Historical Society
  - Academic
  - Public
- Preservationists
- Newspaper Industry
  - Press Association
  - Publishers
- Other “non-voting” advisors
  - Digital Technologies
  - Special Collections
  - Collection Development

Selecting Content

Selecting Content

- Priorities
  - It's important to set your priorities up front
    - There is so much potential content, it's easy to get distracted
      - Utah 1850-1950: estimate 8 million pages
    - You will have to tell some people they have to wait
  - Best choice: titles with broad appeal but without broad access
  - Remember: crawl, walk, run
- Rural vs. metro
  - Rural will reinforce a statewide initiative
  - Generally there is good availability of metro papers via film
  - Metro papers will have more readers
- Weeklies vs. dailies
  - Weeklies cover a longer time period
  - Dailies provide more news coverage

Selecting Content

- Original materials vs. microfilm
  - More on this later
- Size
  - Fewer/bigger collections vs. more/smaller collections
  - We generally have 8-12K pages in a collection
- Dates
  - Start with the earliest date available and move forward in time
  - Minimize gaps; have continuous runs
- Preservation
  - Once digitized, source materials will rarely need to be handled
  - High use/poor condition documents could be higher priority
- Build around a theme or historical topic
- Titles that may not get done via NDNP
  - NDNP will eventually fund every state
  - Don’t initially do titles that NDNP might pay for later

Digital Newspapers Technical Issues

System Administration

- University of Utah uses CONTENTdm to manage all of our digital collections
  - Windows 2000 Server platform
  - Features to be discussed in later meeting
- Servers:
  - Dell 2650 PowerEdge Servers
    - Dual 3.2 GHz Xeon Processors w/ 2GB RAM
    - IIS 5 for web services
- 100Gb network connection to Univ. Backbone

Challenges

- Data “Wrangling”
  - Huge amounts of newspaper “raw” scans
  - Huge amounts of database records
- Digital Archiving
  - U of U archives high res, “raw” scans of full pages
  - Approx. 28MB each
  - 300K scans currently
    - 8TB’s
  - Will grow to 500K scans by end of year
    - 14 TB’s
  - Backed up on 2 copies of Ultrium Tape
    - One copy on-site
    - One copy off-site
- File Size Optimization
  - 1/3 of our users are still using dial-up connections
  - Must load web files in a reasonable amount of time

Storage

- Store database data, text & images
  - UDN uses PDF for web images
- UDN has 300,000 newspaper pages online
  - Each page & article are separate records
- 3.4M database records.
  - 2.5TB online storage available
  - Currently using approximately 400GB
  - Averages to 1.3GB's of online storage for each 1,000 pages

Quality Assurance (QA) Lessons

- Readability
  - If a patron can't read the article, then neither can the Optical Character Recognition (OCR)
  - It becomes functionally useless
- PDF has good features, but is not perfect
  - Reader V.6 vs. V.7
    - V.6 text arranged by entry order
    - V.7 text arranged by coordinates

Scanned Images

- Resolution depends on scanning device
  - 2400 – 8000 DPI for microfilm
    - Mekel, Sunrise scanners.
  - 150 – 300 DPI for paper
    - Digital Camera back @ 22MP
  - File sizes
    - Dependent on pixel dimensions and bit depth
    - Archival size is ~ 28MB per newspaper page.

Resolution Examples

- Microfilm original is about 1x2 inches
  - Scanned at 6000dpi
  - 6000x12000 pixel final image
- Paper original is about 18x30 inches
  - 22MP digital camera
  - 4700x4700 pixel image
  - Extrapolates to 156 DPI
- Readability
  - Final Pixel dimensions
  - 4000-5000 pixels across page width is acceptable. For OCR, and final readability

Bit Depth

- Scanned & processed in 8-bit grayscale
  - 256 shades
  - Approx. 50MB files
- Archived in 4-bit grayscale
  - 16 shades
  - Still ~ 28MB per full page image.
  - 300,000 pages = 8.4TB of “archived” images
- Displayed at 1-bit (bitonal) TIFF in PDF
  - Some will be converted to 8-bit JPEG for readability
  - PDF full page image size ~ 120KB

File Sizes For Web

- 1/3 of users using dial-up connection
  - Full Page image
    - 5MB JPEG takes 15 minutes
    - 2MB JPEG takes 6 minutes
  - Web delivery
    - 200-500KB range
    - Most are ~ 300KB
    - Rarely > 500KB, but they do exist
  - Salt Lake Tribune
    - Extremely small font
      - Size 4
    - Decided to use JPEG
      - Much more readable , but larger file size
      - Approx. 1.5 MB's

Source Materials

Original Paper

- PRO's
  - Capture new, high quality digital image
    - Using 21st century technology
  - Image quality more controllable
    - You/your vendor create images according to your spec
  - Upshot: cleaner images mean higher search accuracy
    - This is the most important reason
- CON's
  - Harder to find originals
    - And they may need repairs (avg cost = 22 cents/pg.)
  - More expensive to scan
    - 30 cents vs. 22 cents per page
  - Harder to find scanner
    - Need overhead camera; flatbeds won't work
    - Probably has to be local vendor
  - NDNP isn't funding
    - Obviously an important restriction

Microfilm

- PRO's
  - Readily available
    - Almost every important title has been filmed
  - Cheaper, faster to scan
  - Scanning can be remote
    - Film boxes are easily transported
  - NDNP "compatible"
    - You will gain experience in the NDNP process
- CON's
  - Digital image quality dependent on film quality
    - Varies widely, even within a single reel
    - USNP standards set in 1980's
      - After a lot of newspaper film had been created
  - Images are decades old
  - Search accuracy lower (generally)
  - Our process: scan a page up to 3 times to get good image

Search Accuracy Test - 2003

Title		Issues	Searches	Hits	Pct
American Eagle	film	1	3	3	100.0%
Carbon Co. News	film	2	9	7	77.8%
Eastern Utah Advocate	film	6	46	36	78.3%
Emery Co. Progress	paper	3	14	8	57.1%
Eureka Reporter	paper	6	29	26	89.7%
Manti Messenger	film	8	62	34	54.8%
Millard Co. Chronicle	paper	3	17	15	88.2%
Murray Eagle	film	2	18	8	44.4%
News Advocate	film	5	34	21	61.8%
Ogden Standard	paper	17	151	115	76.2%
Park Record	paper	5	22	10	45.5%
Park Record	film	2	9	2	22.2%
Times Independent	paper	3	18	12	66.7%
Tooele Co. Chronicle	paper	1	16	15	93.8%
Vernal Express	film	4	38	31	81.6%
Washington Co. News	paper	5	27	17	63.0%

Accuracy Test Totals

	Issues	Searches	Hits	pct
Paper	43	294	218	74.1%
Film	30	219	142	64.8%
Combined Totals	73	513	360	70.2%

Disclaimer: This test was performed in late 2003, and while we attempted to construct random samples and sample sizes that would render supportable results, we recognize that this test probably isn't academically rigorous. Furthermore, our scanning techniques for both originals and microfilm have improved since then. It is our educated guess that if a similar test was performed in 2005, the results would be significantly improved.

Cleaning and Mending Newspapers for Digitization

Techniques

Dry Cleaning

- Working on a smooth work surface, gently and systematically rub grated eraser crumbs over the sheet using a circular motion. Work from the center outward toward the page edge, cleaning torn areas carefully to prevent damage. Thoroughly remove eraser crumbs using a drafting brush.

Techniques

Opening Uncut Sheets

- With the newspaper resting on a smooth, flat surface, use a letter opener or moderately sharp knife to “slit” rather than “saw” the sheets along the uncut edge. The free hand should be used to compress the leaf slightly to ensure the knife blade cuts only the fold.

Techniques

Heat Set Tissue Mending

- Score heat set tissue (using a dissecting needle and steel straight edge) and tear repair strips slightly less that 1/8 inch wide. Alternatively, cut heat set tissue strips using a scalpel and steel straight edge.
- Mending should be applied to the verso (back) of the sheet to minimize any visual distraction caused by the repair. While mending on one side of the sheet is typically adequate, severe damage may require both the recto the verso be repaired.

Heat Set Tissue (cont.)

- Align torn paper to ensure the scarfed edge is properly repositioned. Set the tacking iron temperature to avoid the burning paper (approximately 275° F.).
- Working from the apogee of the tear toward the edge of the page, “tack” the heat set tissue in place with the tacking iron following the contours of the tear.
- With the mend lightly attached, cover the heat set tissue with a sheet of silicone release paper and apply sufficient heat and dwell time with the tacking iron to completely attach the repair tissue to the newspaper.

Materials

Materials for Dry Cleaning

- Nutmeg Grater - KitchenAid or similar
  - source, Kitchen Supply store
- Silicone Release Paper
  - source, TALAS, item # TPB017001
  - alternative material, Baker's Parchment from a local grocery store
- Staedtler Mars Plastic Eraser
  - source, TALAS, item # TCD064001
  - alternative source; local art or architectural supply store
- Drafting Brush
  - source; local art or architectural supply store

Materials

- Advanced Technology Tacking Iron
  - source, TALAS, item #TTB065001
- Neschen Filmoplast R Heat Set Tissue
  - source, Gaylord, item # GH-R12233
- Self Healing Cutting Mat, green, 9" x 12"
  - source, TALAS, item #TTB060002
  - alternative source, local artist supply store
- Scalpel Handle, #5P
  - source, TALAS, item #TTB006025
- Scalpel Blades, #11 (disposable)
  - source, TALAS, item #TTB006003
- Dissecting Needle
  - source; local art or architectural supply store
- 18" Steel Ruler
  - source; local art or architectural supply store

Sources for Supplies

- TALAS

  20 West 20th Street, New York, NY 10011

  Tel: 212-219-0770

  Fax: 212-219-0735

  Email: info@talasonline.com; www.talasonline.com

- Gaylord Bros.

  P.O. Box 4910, Syracuse, NY 13221-4901

  Tel: 800-448-6160

  Fax: 800-272-3412

  Email: www.gaylord.com

Digitizing Historic Newspapers

“the University of Utah Way”

3 Keys - Introduction

1. Viewing – image quality
2. Browsing – meta data quality
3. Searching – word index quality

3 Keys - Viewing

"Garbage In, Garbage Out"

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

- Photo realism
- Text readability
- Meta data collection
- Searchable word accuracy
- **User Experience!**

3 Keys - Viewing

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

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

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

3 Keys - Viewing

University of Utah Project Examples:

1. Resolution – Salt Lake Tribune font size
    1. 50% scale bi-tonal tiff (original config)
    2. 100% scale bi-tonal tiff (better)
    3. 100% scale jpg at 40% quality (best)
2. Media type – multi-pass film scanning

3 Keys - Browsing

There are Two levels of meta data in newspapers:

1. Page-level (NDNP at present)
2. Article-level (Many others including U of U)

3 Keys - Browsing

Page-level meta data

- Inexpensive to capture
- High value for browsing
- Low value for searching
- Relatively small amount of data

3 Keys – Browsing

Article-level meta data

- More expensive to capture
- High value for browsing
- High value for searching
- Relatively large amount of data

3 Keys - Browsing

Approaches for capturing meta data:

- Automated
  - High initial (development) cost
  - Long lead time to start project
  - Low per-page cost
  - Medium data accuracy
- Manual
  - Low initial cost
  - Short lead time to start project
  - High per-page cost
  - High data accuracy (99.95%)

3 Keys - Browsing

University of Utah Project Examples:

1. Time span covered by U of U project allows for simple page-level meta data configuration – lower costs, faster production
2. Zoning articles increases OCR accuracy by minimizing bounding errors and adding conjoined words, phrase continuity, etc.

3 Keys - Searching

Keyword searching is one of the most powerful tools available for researchers. It allows the user to search on key points they are interested in such as names, places, concepts, etc.

Therefore, it is critical to have as much accuracy as possible in your keyword index.

3 Keys - Searching

Approaches for capturing full word index:

- Automated (OCR)
  - Low initial cost
  - Low per-page cost
  - Variable data accuracy
- Manual (Keyboarding)
  - Low initial cost
  - High per-page cost
  - High data accuracy
- Hybrid (OCR with manual correction)
  - Low initial cost
  - High per-page cost
  - High data accuracy
- Automated Optical Word Recognition (OWR)
  - Low initial cost
  - Low per-page cost
  - Higher (than OCR) data accuracy

3 Keys - Searching

University of Utah Project Examples:

1. OWR Word Options – Strike balance between higher accuracy and manageable index (2 words with filters)
2. OWR vs OCR test resulted in 8% higher word accuracy on historic newspaper sample

Methodology

The goal is to create the best **Viewing**, **Browsing**, and **Searching** experience for your end user.

This is accomplished by applying **automation** where sensible to lower costs and **manual operation** where data accuracy is critical.

Methodology

Examples where automation is most cost-effective:

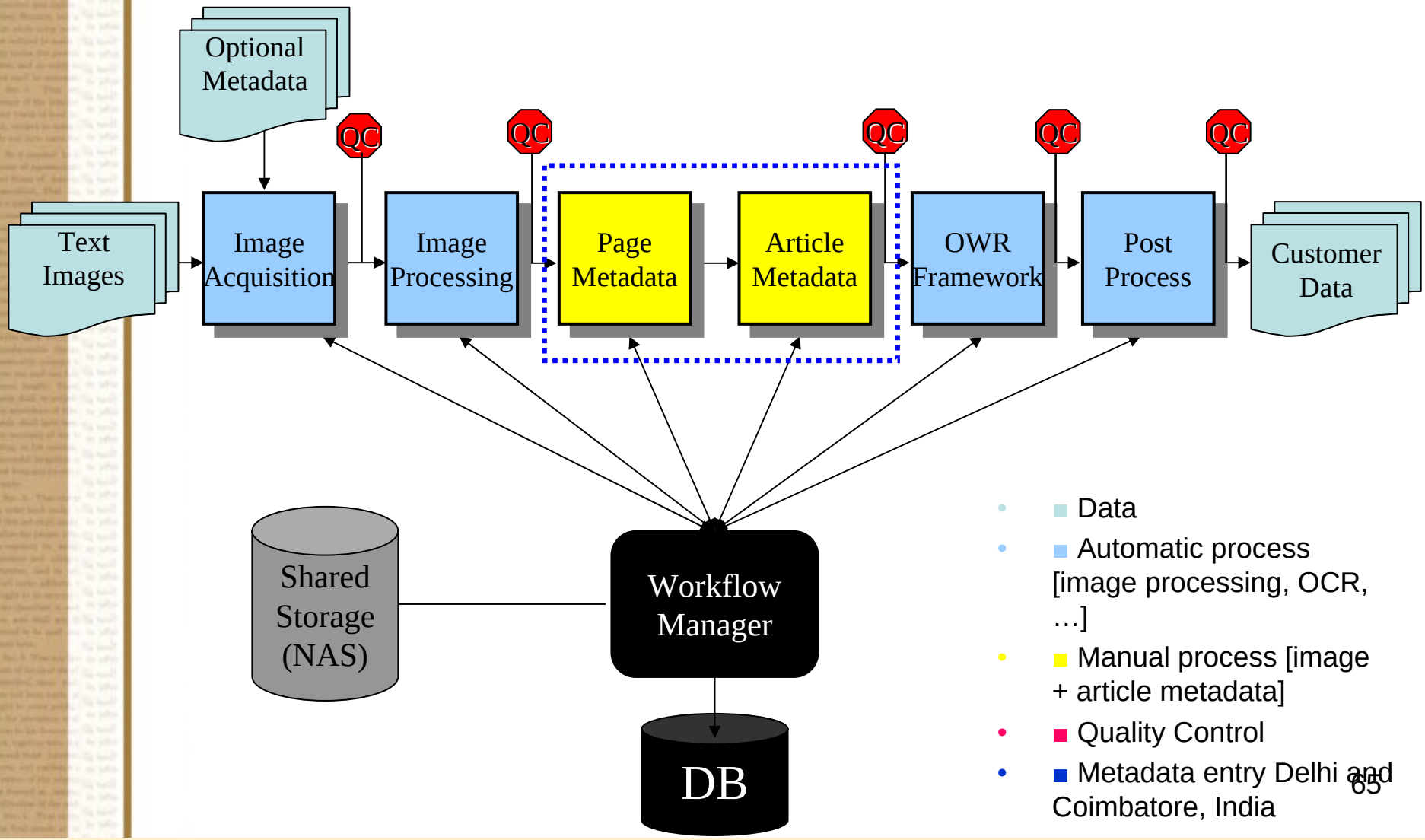
- Repetitive, programmable tasks such as
  - cropping images
  - replicating images
  - reading text from clean images
  - producing output files
  - moving data
  - simple quality checks
- Automated control over systems through workflow management

Methodology

Examples where human intervention is most cost-effective:

- Subjective tasks such as
  - classifying article type
  - linking article continuations
  - defining difficult article boundaries
  - complex quality control checks
- Mission-critical data entry such as
  - page-level meta data
  - article-level meta data

Methodology



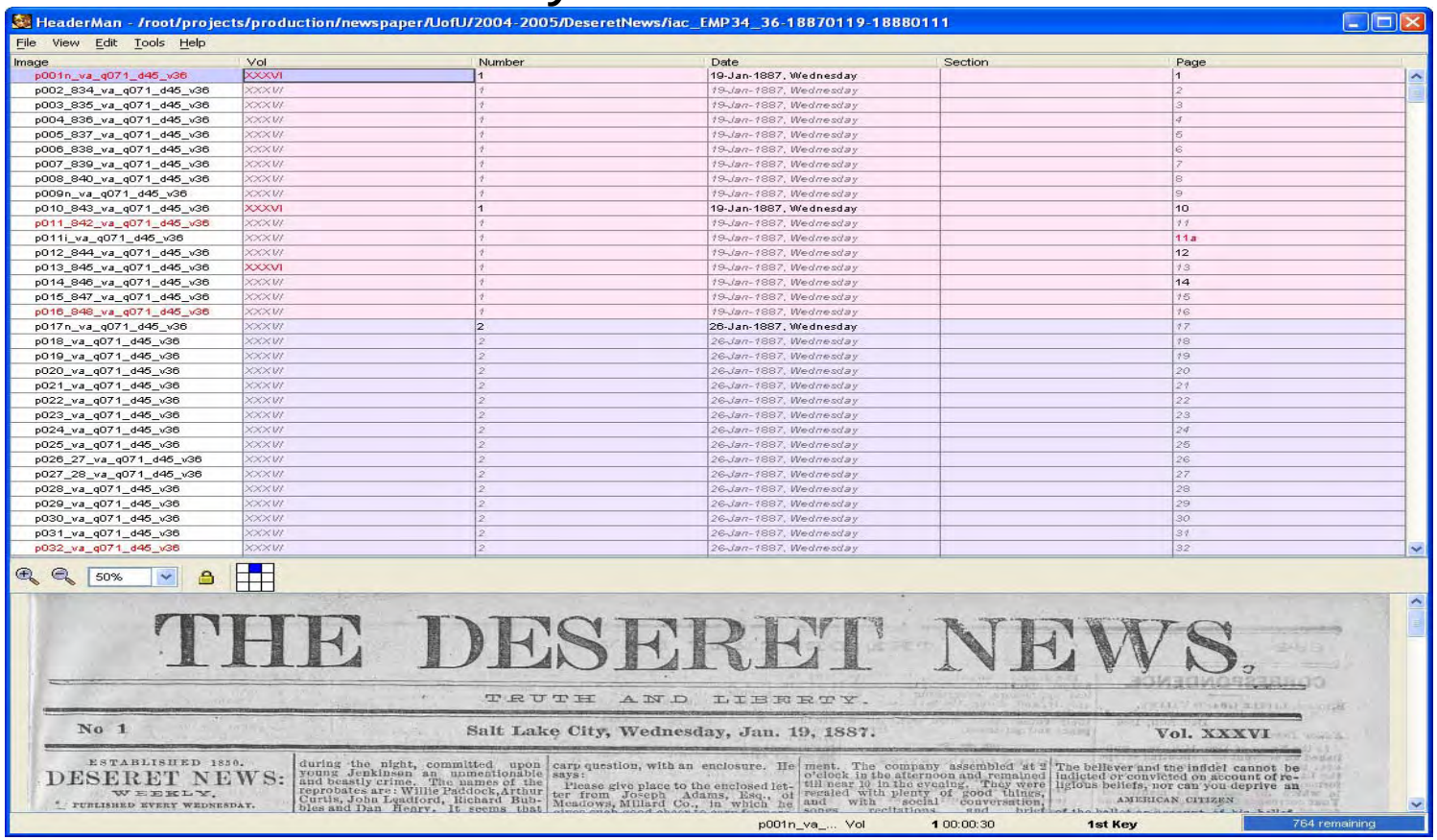
Tool Set

The tools used to allow the most accurate browsing and searching include:

- HeaderMan
- ZoneMan
- OWR™ Framework

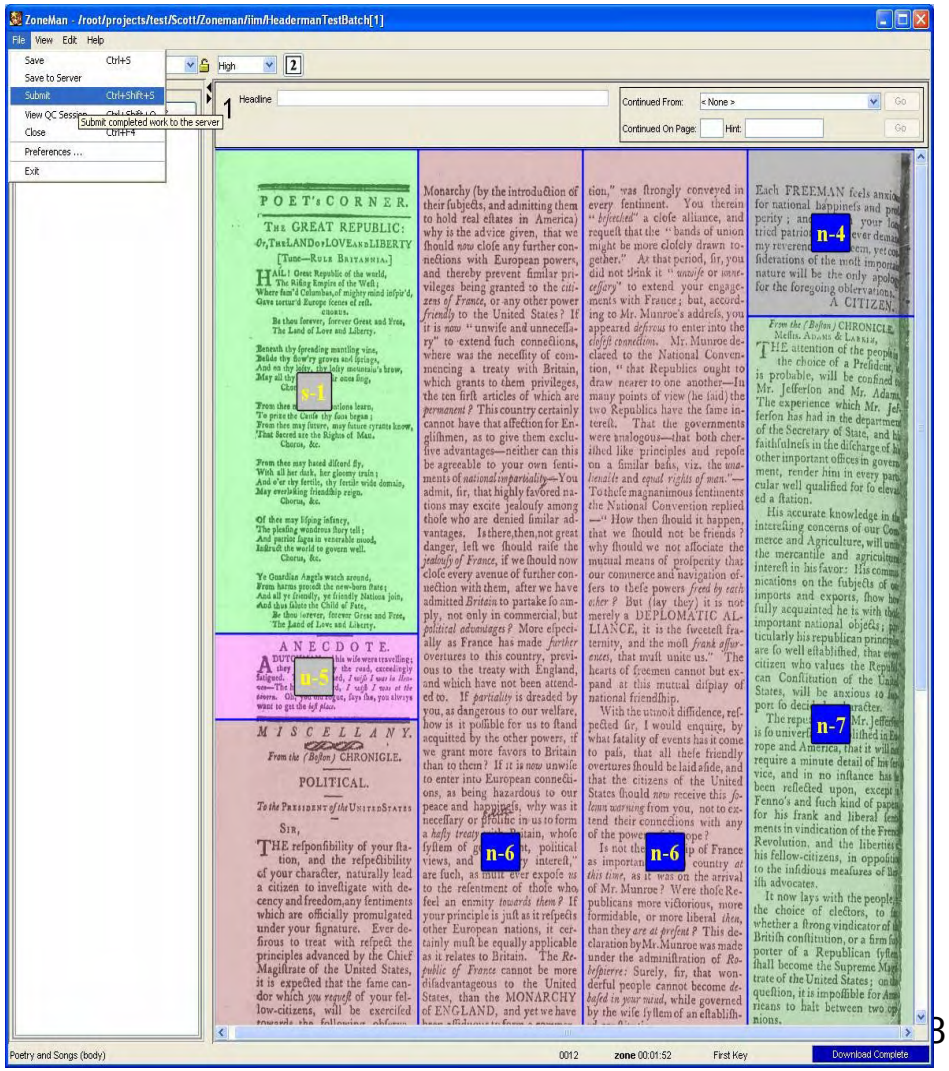
Tool Set - HeaderMan

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



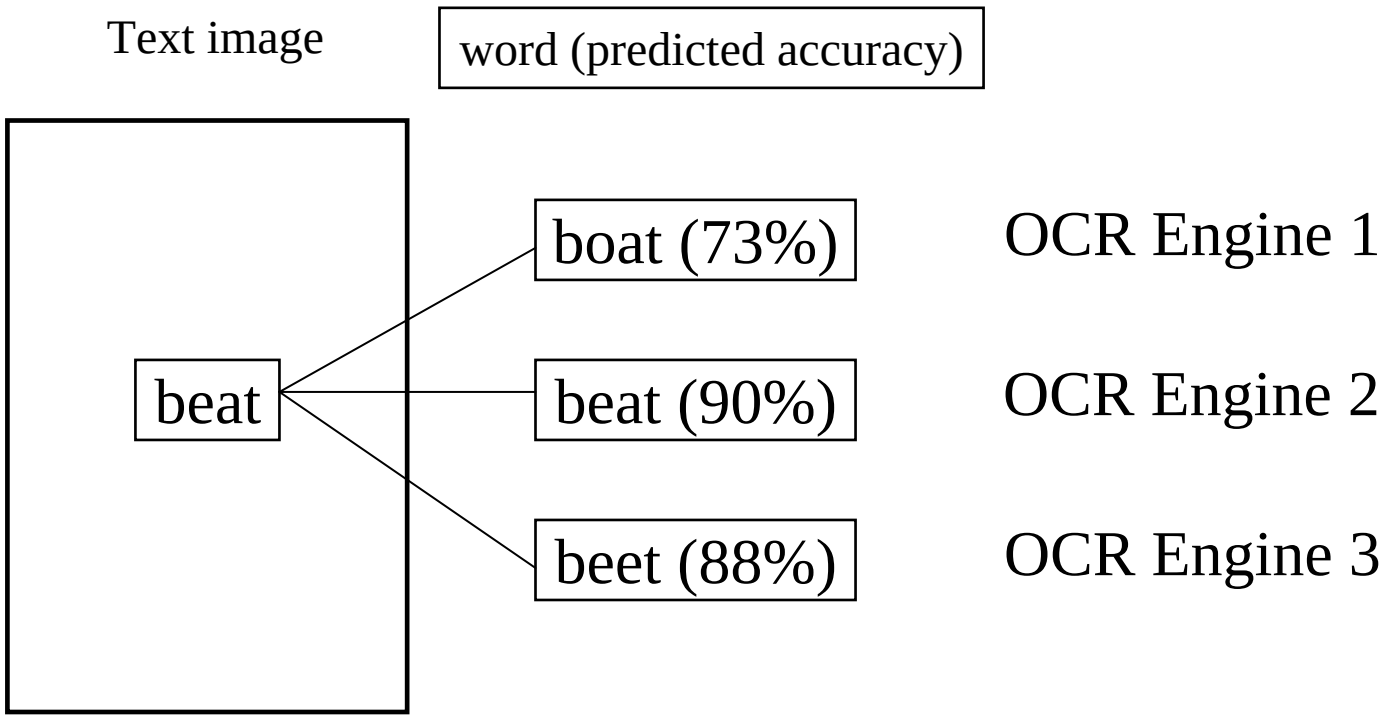
Tool Set - ZoneMan

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



Tool Set – OWR™

OWR™ stands for Optical Word Recognition. It is a state-of-the-art approach combining the best OCR engines to improve overall word search accuracy.



Copyright and Library Digitization

Copyright Periods

Works Published in the U.S. until 1977

Date of Pub.	Conditions	Copyright Terms
Before 1923	None	public domain
1923 - 1977	Published w/out notice	public domain
1923 – 1963	Published w/ notice but not renewed	public domain
1923 – 1963	Published w/ notice and renewed	95 years
1964 – 1977	Published w/ notice	95 years

- Historic newspapers published generally fall into the first and second categories, but you need to review each title on a case-by-case basis.

Metadata and Historic Newspapers

Presented by

What is Metadata?

- Definition
  - "data about data"
  - Information about the content, context and structure of information resources.

Metadata is:
- **Library catalog records**
- **Museum registration records**
- **Archival finding aids**

Types of Metadata

- **Descriptive metadata:** used for the indexing, discovery, and identification of the contents of a digital resource
- **Technical metadata:** information about the digital file itself, such as: format, size, encoding.
- **Administrative metadata:** management information for the digital object, which may include information needed to access and display the resource, as well as rights management information.

Types of Metadata

- **Structural metadata**: information used to display and navigate digital resources; also includes information on internal organization of the digital resource.
- **Preservation Metadata:** records information about the provenance of a digital object throughout it's lifetime.

7 Metadata Principles

1. Good metadata should be appropriate to the materials in the collection, users of the collection, and intended, current and likely use of the digital object.
2. Good metadata supports interoperability.
3. Good metadata uses standard controlled vocabularies to reflect the what, where, when and who of the content.
4. Good metadata includes a clear statement on the conditions and terms of use for the digital object.

7 Metadata Principles

5. Good metadata records are objects themselves and therefore should have the qualities of good objects, including achievability, persistence, unique identification, etc.
6. Good metadata should be authoritative and verifiable.
7. Good metadata supports the long-term management of objects in collections.

Interoperability Protocols

- Z39.50
  - Broadcasts user query to remote databases.
- Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)
  - "Harvests" metadata from registered metadata "providers." "Services" allow users to query harvested metadata.

Dublin Core Elements

- Title
- Creator
- Subject
- Description
- Publisher
- Contributor
- Date
- Type
- Format
- Identifier
- Source
- Language
- Relation
- Coverage
- Rights

DC Element: Title

- Definition: The name given to the resource.
  - Typically, a Title will be a name by which the resource is formally known.

Definitions are from the Dublin Core Web Site

http://purl.oclc.org/dc/

DC Element: Creator

- An entity primarily responsible for making the content of the resource.
  - Examples of a Creator include a person, an organization, or a service.

DC Element: Subject

- A topic of the content of the resource.
  - Typically, a Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource.
  - Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.

DC Element: Description

- An account of the content of the resource.
  - Description may include but is not limited to:
    - an abstract,
    - table of contents,
    - reference to a graphical representation of content,
    - or a free-text account of the content.

DC Element: Publisher

- An entity responsible for making the resource available.
  - Examples of a Publisher include
    - a person,
    - an organization,
    - or a service.

DC Element: Contributor

- An entity responsible for making contributions to the content of the resource.
  - Examples of a Contributor include
    - a person (photographer, translator, etc.)
    - an organization
    - or a service.

DC Element: Date

- A date of an event in the lifecycle of the resource.
  - Typically, Date will be associated with the creation or availability of the resource.
  - Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [W3C-DTF] and follows the YYYY-MM-DD format.

DC Element: Type

- The nature or genre of the content of the resource.
  - Type includes terms describing
    - general categories
    - functions
    - genres,
    - aggregation levels for content
  - Recommended best practice is to select a value from a controlled vocabulary
    - **For example, the Dublin Core Type Vocabulary.**
    - **To describe the physical or digital manifestation of the resource, use the FORMAT element.**

DC Element: Format

- Definition: The physical or digital manifestation of the resource.
  - Typically, Format may include the media-type or dimensions of the resource.
  - Format may be used to determine the software, hardware or other equipment needed to display or operate the resource.
  - Examples of dimensions include size and duration.
  - Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types [MIME] defining computer media formats).

DC Element: Identifier

- An unambiguous reference to the resource within a given context.
  - Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system.
  - Formal identification systems include the
    - Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL))
    - Digital Object Identifier (DOI)
    - International Standard Book Number (ISBN)

DC Element: Source

- A Reference to a resource from which the present resource is derived.
  - The present resource may be derived from the Source resource in whole or in part.
  - Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.

DC Element: Language

- A language of the intellectual content of the resource.
  - Recommended best practice is to use values from a controlled vocabulary standard
  - ISO639-2: Three letter language code
    - English = eng
    - Yiddish = yid
  - RFC 1766: Includes a two-letter Language Code (taken from the ISO 639-1 standard), followed optionally, by a two-letter Country Code (taken from the ISO 3166 standard)
    - English – United States = en-us
    - English – United Kingdom = en-uk

DC Element: Relation

- A reference to a related resource.
  - Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
  - Prescribed list of qualifiers are used in this element.

DC Element: Coverage

- The extent or scope of the content of the resource.
  - Coverage includes spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity).
  - Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names [TGN]) and that, where appropriate, use named places or time periods in preference to numeric identifiers such as sets of coordinates or date ranges.

DC Element: Rights

- Information about rights held in and over the resource.
  - Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information.
  - Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights.
  - If the Rights element is absent, no assumptions can be made about the status of these and other rights with respect to the resource.

Types of Dublin Core

- Simple Dublin Core 15 elements

- Qualified Dublin Core 15 elements with
  - **Refinements** make an element's meaning more specific without extending its meaning; "modifiers." (Title: Alternative; Relation: Is Part Of)
  - **Schemes** aid in the interpretation of an element value, including controlled vocabularies (LCSH) and formal notations (ISO639-2)

Audience Needs

- Who is your current audience?
- Who is your audience in a digital environment?
- Do your metadata practices meet defined audience needs?
- Will your metadata make sense in a shared environment?

Challenges in describing cultural heritage objects (newspapers)

- Level of description (collection vs. item)
- Focus of description (original vs. digital)
- What's a collection? (new digital collections vs. traditional physical collections)
- Metadata for complex objects (diaries, manuscripts, ephemera, digital audio, transcripts, etc.)

DC Metadata Resources

- Western States Dublin Core Metadata Best Practices
  - http://www.cdpheritage.org/resource/metadata/wsdcmbp/
- CDP Metadata Resources page
  - http://www.cdpheritage.org/resource/metadata/rsrc_metadata.html
- NISO:A Framework of Guidance for Building Good Digital Collections
  - http://www.niso.org/framework/Framework2.html

Digital Collection Management Software

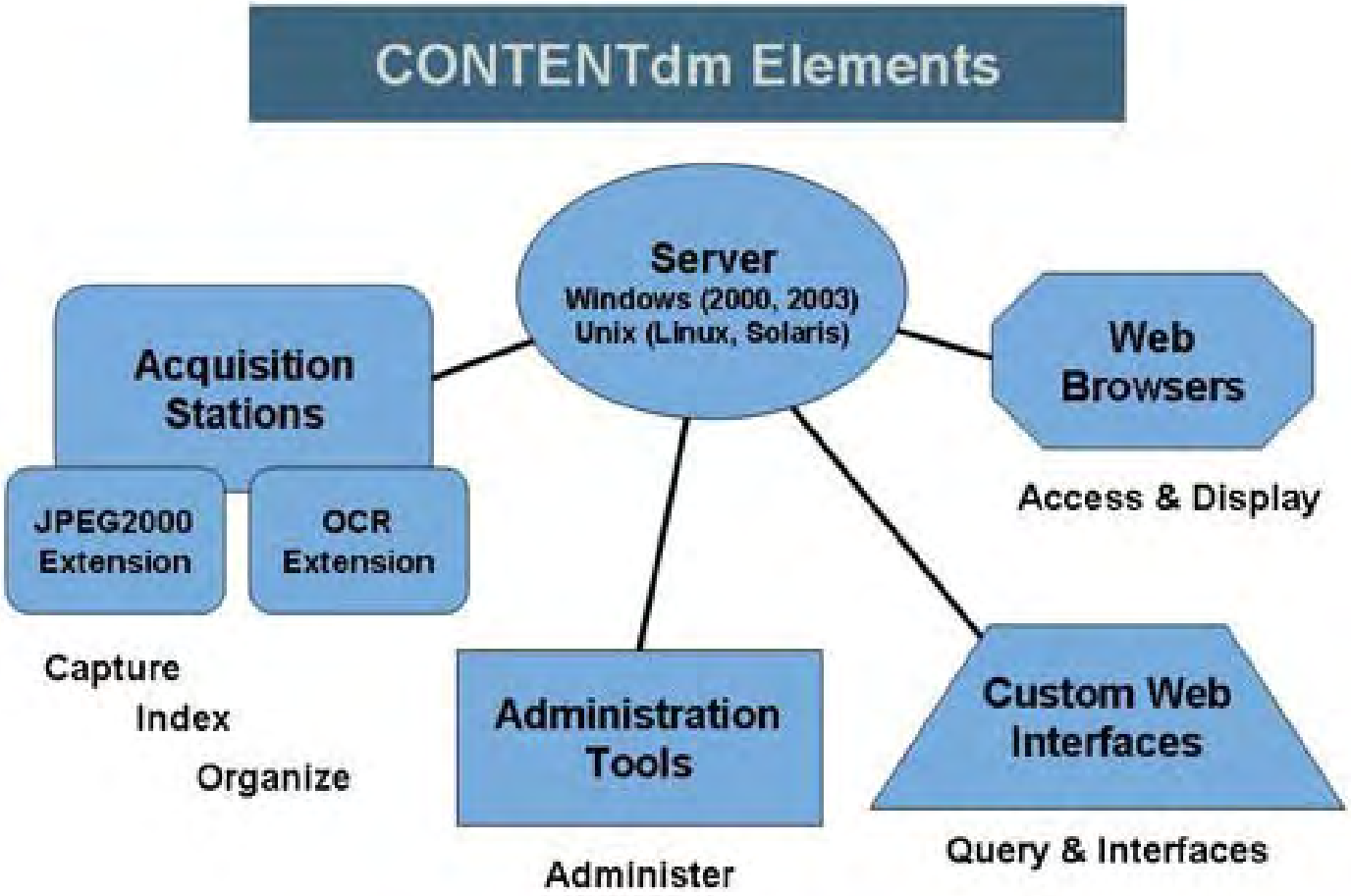
CONTENTdm history

- Developed at University of Washington
  - Center for Information Systems Optimization
    - Research group headed by Dr. Greg Zick, Professor of Electrical Engineering
    - Formed in late 1980s
  - CONTENT built collaboratively with UW libraries
- Spun off by UW in 2001
  - CISO formed DiMeMa Inc.
  - Marketing partnership with OCLC in 2002
- UU purchased Level 1 license in early 2000

Why we use CONTENTdm

- Search efficiency
- Scalability
  - Over 4 million objects and rising
- Web-based
- XML database
- Open standards
- Able to handle all media types
- Affordable
- Customizable
- Company is innovative, product continually evolving

CONTENTdm Architecture



CONTENTdm licenses

- Utah Digital Newspapers
  - CONTENTdm (unlimited) license
  - Stores 3.3+ million digital objects
- “Other” digital collections
  - CONTENTdm Level 6 license (256K objects)
  - JPEG2000 extension
  - 150,000+ digital objects

Dell Server hardware

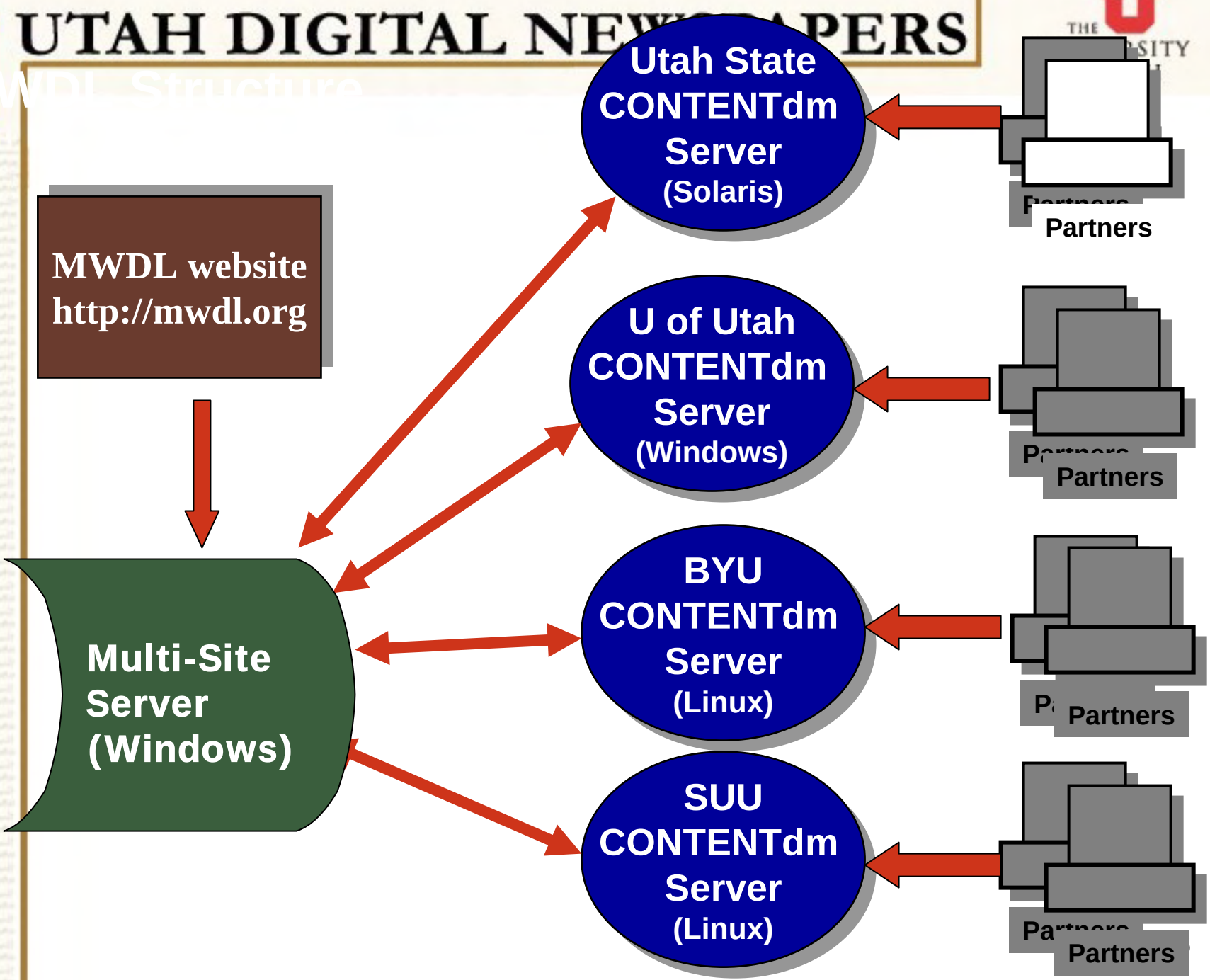
- Utah Digital Newspapers
  - Windows 2000 active failover cluster
    - Pentium Xeon 3.2GHz dual processors
    - 2 GB of RAM
    - Fibre Channel (FC) to 2TB SAN
- All “other” digital collections
  - Windows 2000 single server
    - Pentium Xeon 3.2GHz dual processors
    - 2 GB of RAM
    - SCSI direct attach to 1TB SAN

Digital Collection Examples

- www.lib.utah.edu/digital/Dard/index.html
- www.lib.utah.edu/digital/bodmer/index.html
- www.lib.utah.edu/digital/sanborn/index.html
- www.lib.utah.edu/digital/wright/index.html
- http://medstat.med.utah.edu/NOVEL/
- http://www.lib.utah.edu/digital/uupress/index.html
- http://www.lib.utah.edu/digital/mwdl/
- http://westernwaters.org/

UTAH DIGITAL NEWSPAPERS

MWDL Structure



Open Source Digital Libraries

Definition

digital library n.

Focused collections of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance.

Ian H. Witten and David Bainbridge, How to Build a Digital Library, Morgan Kaufmann Publishers, 2003.

Digital Library Evaluations and Guides

- A Guide to Institutional Repository Software, v 3.0 from the Budapest Open Access Initiative (http://www.soros.org/openaccess/software)
- Creating an Institutional Repository: LEADIRS Workbook from **LE**arning **A**bout **D**igital **I**nstitutional **R**epositories (http://www.dspace.org/implement/leadirs.pdf)
- Reference Model for an Open Archival Information System (OAIS) (http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html)

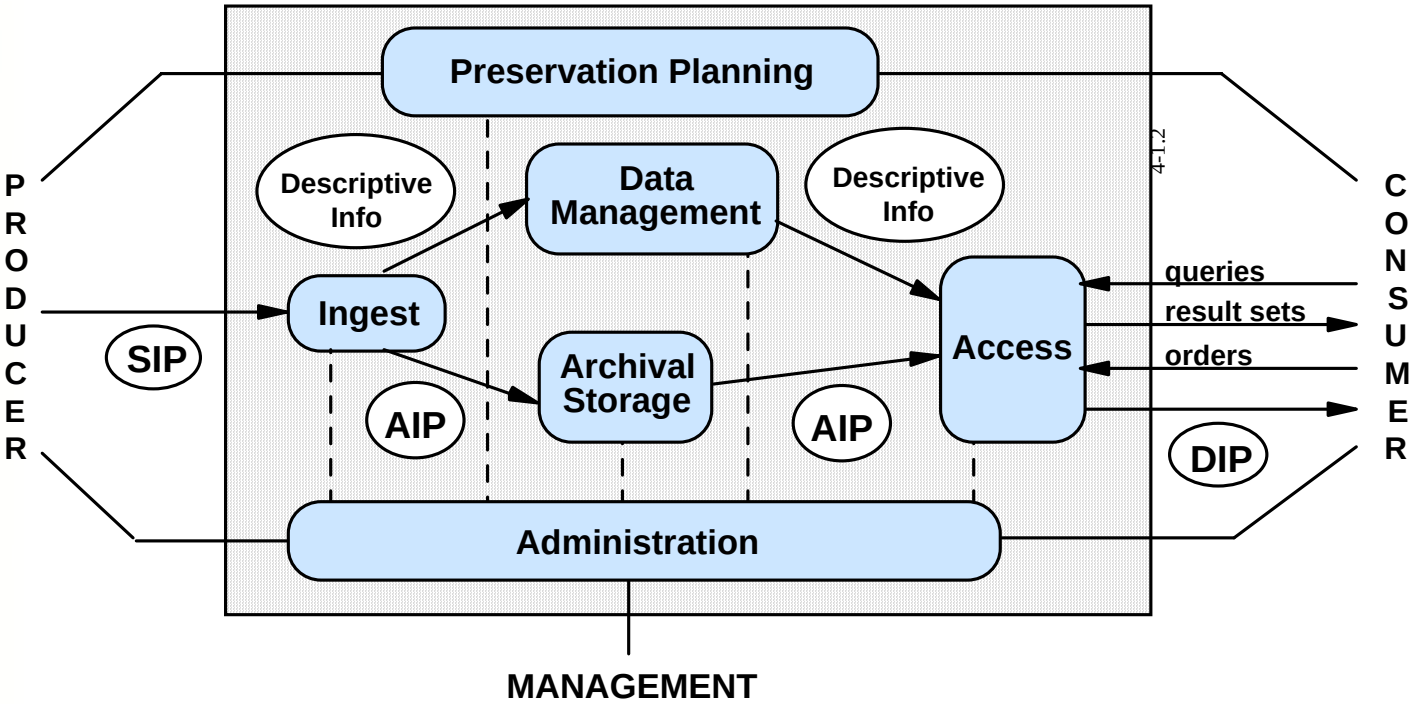
Suggested Evaluation Criteria

- **If the software is open source, what is the quality of the software?**
  - If all other factors are equal, open source software allows maximum flexibility to add new features or to customize the software for a client.
- **Does the software consume XML, TEI, METS, and other standard digital objects?**
  - Because XML data objects enable a great deal of flexibility in data delivery and utilization, the software must consume XML data. Software that ingests METS objects is preferred since METS appears to be an emerging standard in the library world.
- **Is it possible to extend the software?**
  - Is there an SDK available? For commercial software, how cooperative is the vendor?
- **If you want to customize or extend the software, what language(s) is it written in?**
  - Again a question of what local expertise is available.
- **Does the software support collection browsing?**
  - Even though it seems that this feature must be an integral part of any data delivery software, it sometimes is not.

Suggested Evaluation Criteria 2

- **What languages are supported?**
  - English, Spanish, French, Chinese, etc
- **What standards are supported?**
  - Standards that should be supported are Z39.50, OPAC, OAI, OAI-PMH, Dublin Core, MARC 21, etc.
- **What image/file types are supported?**
  - The software must support a wide variety of image and file types, for example, TIFF, JPEG, JPEG2000, PDF, etc.
- **What platforms – Windows, Linux, OS X, etc – does the software run on?**
  - If you have local support for Linux but not for Windows, this may be an important question.
- **What access methods / browsers are supported? Are browser plug-ins or extensions required?**
  - Generally you will want to make the user experience as easy and as simple as possible.
- **Also see *Digital Object Library Products* by William Lund**
  - RLG DigiNews, October 2001.

Digital Library Architecture



SIP = Submission Information Package
AIP = Archival Information Package
DIP = Dissemination Information Package

From Merrilee Proffitt, "*News from the Digital Library: The Metadata Encoding and Transmission Standard: The Open Archival Information System*", RLG, April 2003.

Digital Library Software

- aDORe
  http://public.lanl.gov/herbertv/papers/aDORe_20050128_submission.pdf
- Archimede http://archimede.bibl.ulaval.ca
- ARNO http://www.uba.uva.nl/arno
- Bepress http://www.bepress.com
- CDSware http://cdsware.cern.ch
- CONTENTdm http://www.contentdm.com
- DSpace http://www.dspace.org
- Eprints http://www.eprints.org
- Fedora http://www.fedora.info
- Greenstone http://www.greenstone.org
- i-Tor http://www.i-tor.org
- MyCoRe http://www.mycore.de
- Open Repository http://www.openrepository.com

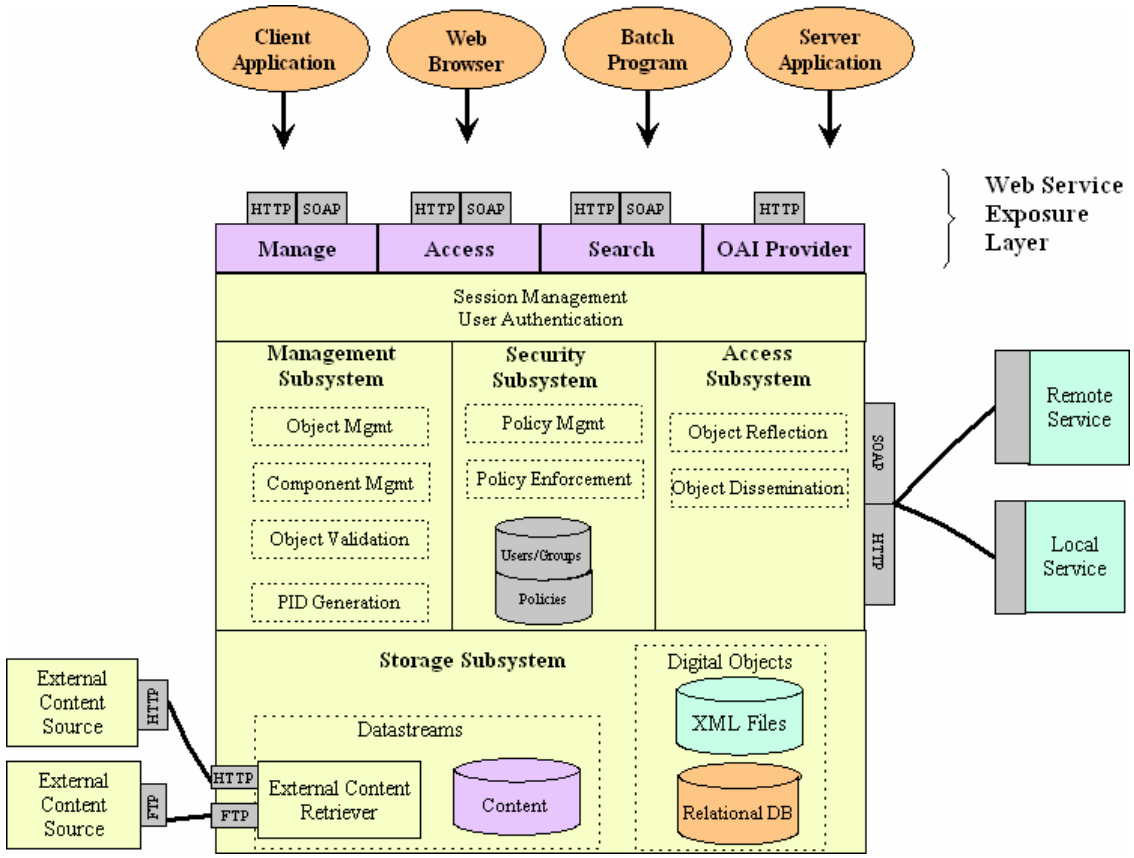
Digital Library Software
Fedora 1

- **F**lexible **E**xtensible **D**igital **O**bject and **R**epository **A**rchitecture is a general purpose repository service developed jointly by The University of Virginia Library and Cornell University.
- In 1998 Fedora began as a DARPA and NSF-funded research project of Carl Lagoze and Sandy Payette.
- Fedora is now funded by a grant from the Andrew W. Mellon Foundation.
- Current release version is 2.0. Version 2.1 is due summer 2005.
- Platforms supported include Linux and Windows.
- Sun's *Java Software Development Kit,* J2SDK 1.4 or better is required. MySQL, Oracle, or other databases are optional.
- Not to be confused with Fedora Linux project from Red Hat! (Red Hat's claim to the name came *after* the digital library's.*)*
- Mozilla Public License (MPL) (http://www.opensource.org/licenses/mozilla1.1.php)

Digital Library Software
Fedora 2

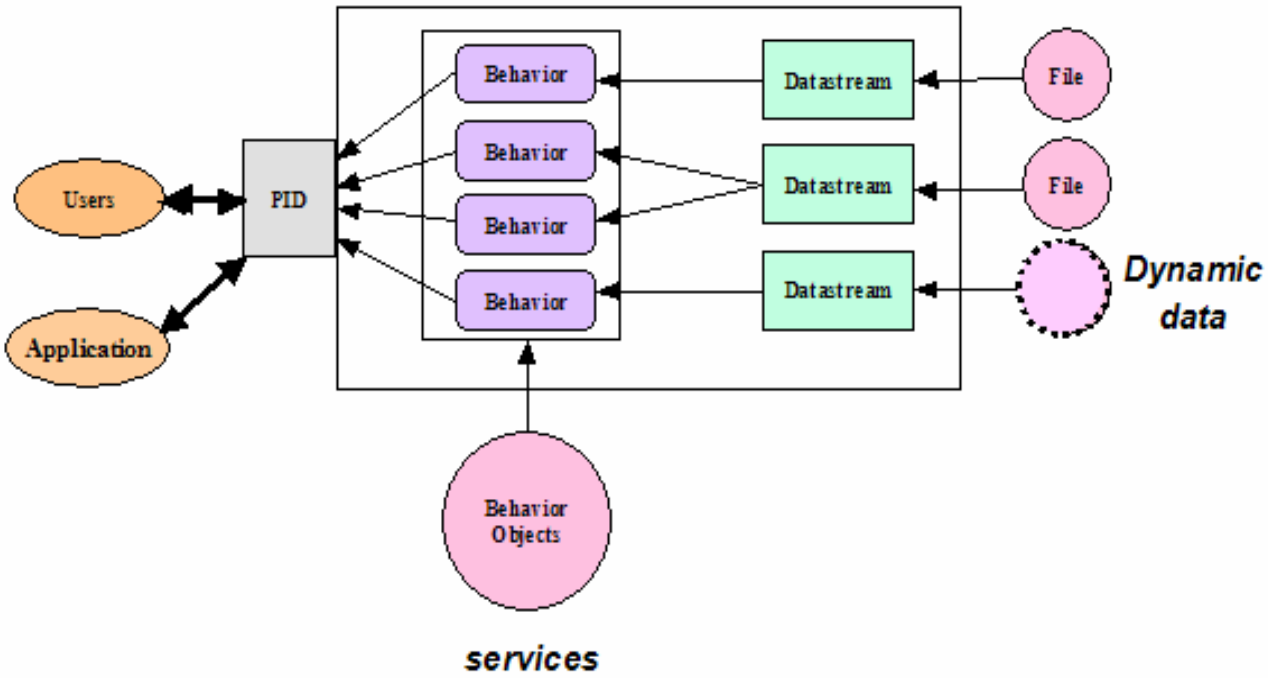
- Library of Congress will use Fedora for its NDNP data.
- LC will open source its NDNP disseminators (release date unknown).
- Fedora mailing list is
  - https://comm.nsdl.org/mailman/listinfo/fedora-users
- Commercial support for Fedora is available from VTLS, Inc. **V**isionary **T**echnology in **L**ibrary **S**olutions (http://www.vtls.com).
- Limited support is also available from iArchives.
- First Fedora Users Conference held in May, 2005.
- Fedora users include National Library of Australia (ARROW Project), University of Virginia, Cornell University, Tufts University, New York University, Northwestern University, Rutgers University, etc. (cf. http://fedora.info/about/deployment.shtml).

Digital Library Software
Fedora 3



Thornton Staples, Ross Wayland, and Sandra Payette, "*The Fedora Project: An Open-Source Digital Object Repository*". D-Lib Magazine, April 2003.

Digital Library Software
Fedora 4



Thornton Staples, Ross Wayland, and Sandra Payette, "*The Fedora Project: An Open-Source Digital Object Repository*". D-Lib Magazine, April 2003.

Digital Library Software
Greenstone 1

- Greenstone grew out of the New Zealand Digital Library Project at the University of Waikato (http://www.sadl.uleth.ca/nz/cgi-bin/library).
- Greenstone developed and distributed as an international cooperative effort (established Aug. 2000) between University of Waikato, UNESCO, and the Human Info NGO (Antwerp, Belgium).
- Current release version is 2.60. Version 3.0 is due in summer 2005.
- Platforms supported include Windows 3.1/3.11/95/98/Me/NT/2000, most distributions of Linux, Mac OS X, Solaris, and FreeBSD.
- Required software (Linux): Apache and Perl.
- Required software (Windows): No additional software required for the *Local Library* version; for the *Web Library* version, Apache or IIS is required.
- Greenstone is internationalized – many languages are supported.
- Gnu General Public License (GPL) (http://www.opensource.org/licenses/gpl-license.php)

Digital Library Software
Greenstone 2

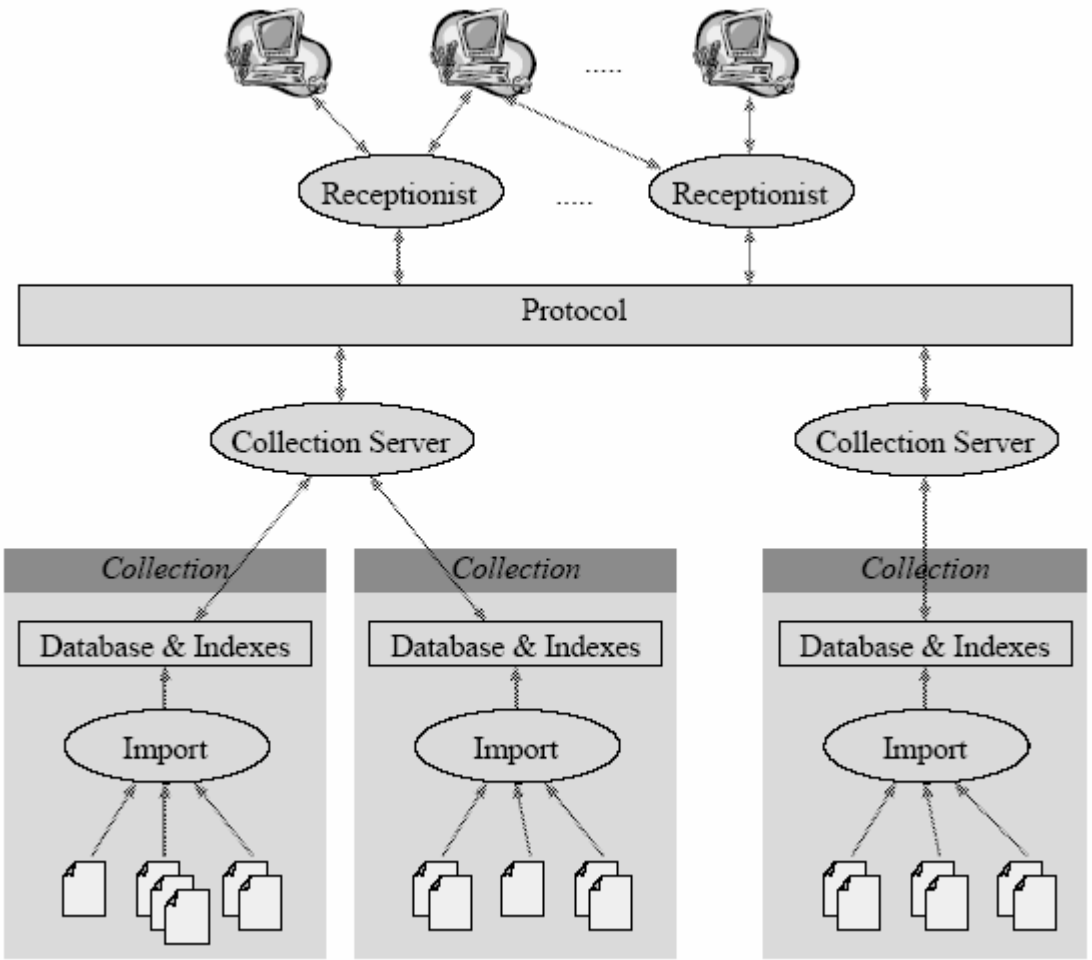
- Greenstone is used by a very wide variety of institutions and organization throughout the world, for example, Auburn University, Wesleyan University, Indian Institute of Management, Peking University, Project Gutenberg, etc. See http://www.greenstone.org/cgi-bin/library?e=p-en-support-utfZz-8&a=p&p=examples for more examples.
- Greenstone has a user's (https://list.scms.waikato.ac.nz/mailman/listinfo/greenstone-users) and a developer's mailing list (https://list.scms.waikato.ac.nz/mailman/listinfo/greenstone-develop).
- Training for Greenstone is readily available through UNESCO, self-study courses, digital library conferences (JCDL, ECDL, ICDL), and elsewhere.
- Commercial support is available from DL Consulting in Hamilton, New Zealand (http://www.dlconsulting.co.nz), founded by one of Greenstone's principal developers, Stefan Boddie.
- Support is also available from iArchives.

Digital Library Software
Greenstone 3

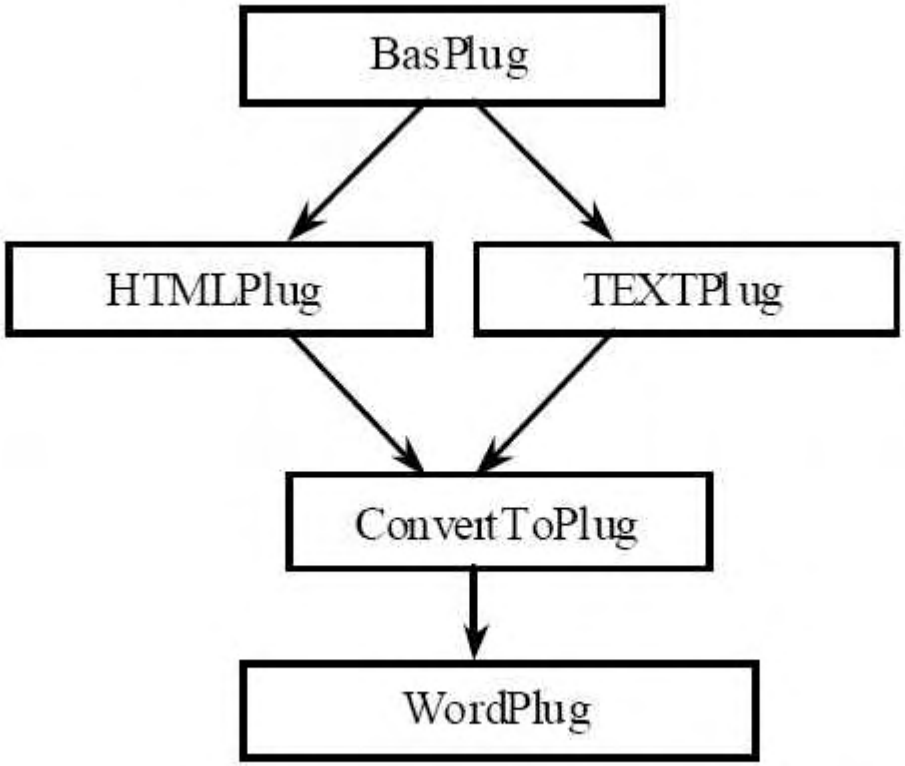
Greenstone Newspaper Collections

- The Argus Digital Collection at the Ames Library, Illinois Wesleyan University
  - http://www.iwu.edu/library/services/argus1.htm
- Ulukau: The Hawaiian Electronic Library
  - http://ulukau.org
- Niupepa: Maori Newspapers
  - http://www.nzdl.org/cgi-bin/niupepalibrary?a=p&p=about&c=niuepa
- The Tundra Times at Ilisagvik College
  - http://ttip.tuzzy.org/cgi-bin/ttimes.exe
- University of Florida (future).

Digital Library Software
Greenstone 4

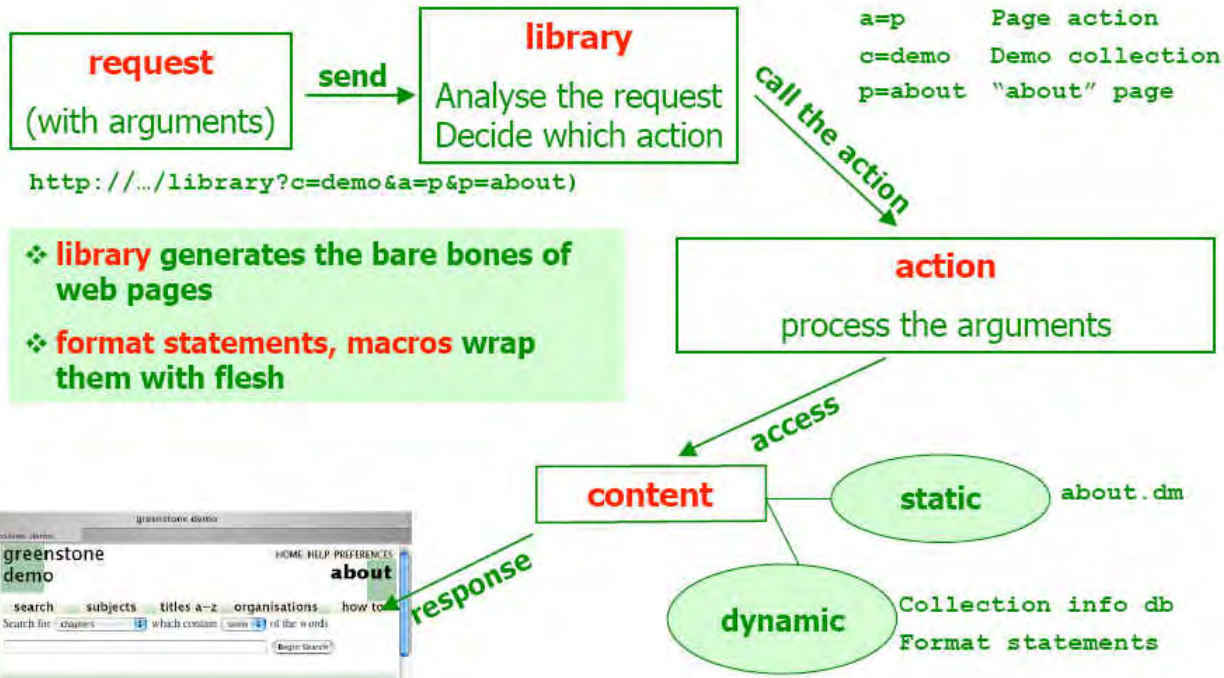


Digital Library Software
Greenstone 5



Digital Library Software
Greenstone 6

Generating web pages



Fund Raising

Fund Raising

- Newspapers have low unit costs
  - Processing (per page) paper: $2.00 film: $1.72
    - Preservation 0.20 0.00
    - Scanning 0.30 0.22
    - Processing 1.15 1.15
    - Database 0.15 0.15
    - Hardware 0.20 0.20
  - Internal staffing costs and general overhead
    - Approx. another $1/page
- But, they are <u>voluminous</u>
  - 8-page weekly 416 pages/year
    - 25 years 10,400 pages/ $20,800
  - 10-page daily 3120 pages/year
    - 10 years 31,200 pages/ $62,400
- So, this is an expensive process
- Rough numbers: 500K pages cost $1.5 million

Federal Funding Is NDNP

- NDNP: National Digital Newspapers Program
  - NEH and LC program
- What NDNP will fund
  - 80% of the costs to digitize newspapers up to 1922
  - Create and send in files according to LC specs
- What it won't (important)
  - Article-level segmentation
  - Local website implementation
    - Hardware, website development
  - Incorporation into other database structures
    - e.g., indexing for CONTENTdm
  - All the volume that you will want to do
- So…..you will have to find additional funding

NDNP

- Guidelines for selecting content
  - Titles that reflect your political, economic, cultural history
  - “Papers of record”
  - Complete runs from high quality film
  - Statewide or multi-county coverage
  - Preference to titles no longer being published
- Can submit “legacy” images
  - Digital content already created and re-purposed for NDNP
  - Work you do before NDNP can still get in
- Phases
  - Phase 1 is underway
  - Phase 2 within a year
    - Not everyone will get in
    - You will need to have a competitive proposal
- Helpful to establish a track record first
  - Will make your NDNP proposal more compelling

Matching Funds

- NDNP requires 20% match
- NDNP accepts F&A/indirect costs
- If your institution's F&A rates are greater than 20%
  - You come out ahead

- Example
  - Total direct costs $300K
  - Indirect costs (30%) $90K
  - Total budget $390K
  - NEH funding (80%) $312K
  - Match (20%) $78K

  - "Receive" 90K from NEH and "contribute" 78K in match
  - Difference is 12K "net income"

- UofU has Office of VP of Research
  - Role is to provide matching funds for university proposals

State and Local Funding

- LSTA
  - This is how UDN got its start
  - It helps to be well networked at the State Library and/or LSTA review board
- Other local funding sources:
  - Public libraries
    - Bigger ones will have more discretionary funds available
  - Newspapers
    - They will have an interest in seeing their early issues on the web
  - Municipalities
  - Foundations
  - Memberships
- Can form groups of local institutions for match
  - e.g., city gov't and local library
  - Selling point is that their contributions are tripled/quadrupled when they're used as a match

Vendor Selection / RFP

Vendor Selection

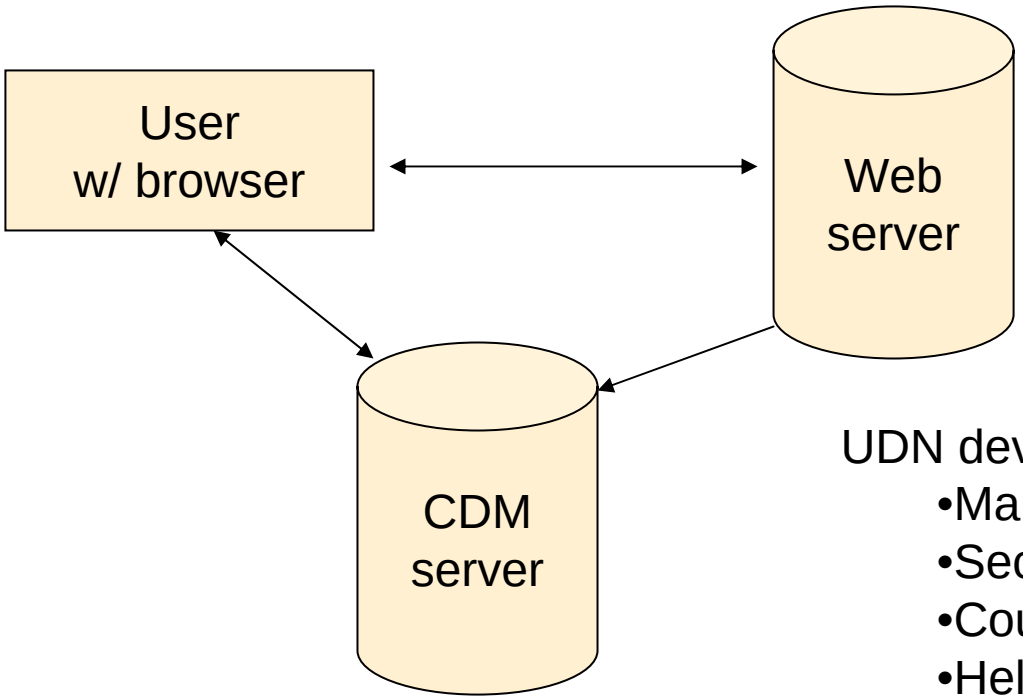
- Which processes are you going to do in-house?
  - What are your internal capabilities?
    - There are a lot of different things to do
    - Assess not only skill level but also staff availability
    - Very important: be honest in your evaluation
  - Everything you don't do will need to be out-sourced
- Easier tasks
  - Title selection/ Obtaining source materials/ Database loading
- Medium tasks
  - Web design and development, Database indexing
- More difficult tasks
  - Preparing source materials/ Scanning/ Processing

Vendor Selection

- What spec's do you want delivered?
  - More detail is generally better
    - But don't get bogged down in them
  - NDNP specs are online at neh.gov
    - But they may not cover everything you want
    - Most vendors should know them well
- Scanning originals will need to be local
  - Old newspapers don't travel well
  - Expensive to ship
- Scanning film could be anywhere
  - Shipping microfilm boxes quite easy
- Data easily sent to/from different locations
  - ftp, mailing disks or hard drives

Web Design

Web Servers



UDN developed pages
- Main page
- Secondary pages
- County map
- Help

Digital files indexed for CDM
Configurable templates
- Search Results
- Advanced Search

Website Design

- Main webpage: Getting into the content
  - Search: across all collections
  - Browse: select a single title
  - Map: browse/select by county
- Secondary pages
  - One for each title
  - Browse
    - Issue date
    - Genealogical
  - Search
    - Headline: nearly 100% accurate
    - Keyword/text: approx. 70% accurate
    - Genealogical
  - Note: classifying articles is not completely accurate
- Provide robust Help
- We can provide our webpage source code

User Feedback

User Feedback

- Newspapers have very popular appeal
  - You're bringing your state's history to life for the general public
  - A "populist" project may be a departure from a more traditional academic approach
- Put the interests of users/customers first
  - Do your best to understand their needs
    - Who they are
    - What they're doing
    - What they think of you
  - Stay in touch with them
    - The environment is always changing

User Feedback

- Create a user survey to receive input
  - Zoomerang, Survey Monkey, Hosted Survey, Web Surveyor, plus many others
- Ask online and/or by email
- Keep it simple to get more responses
  - Fewer questions are better
  - Demographics may not be very important initially
  - Ask about outcomes
  - Ask if you can contact later
    - This can become focus group for more in-depth discovery
- Be open to the feedback
  - Don’t be in denial of negative ratings/comments

User Survey Results

- Frequency of visit
  - 33% first time
  - 54% at least monthly
- Purpose of visit
  - 52% genealogy
  - 33% history, research
- 89% give overall rating of good or excellent
- 79% will return soon
- 78% will tell others about website
- 75% rate search accuracy as good or excellent

User Survey Results

- 67% found new sources of info
- 56% more knowledgeable of family history
- 39% outside Utah
- 32% have dial-up connection
- Comments
  - Add more content (by far the most frequent)
  - Tribune images are poor quality