So You Want to Digitize?: Creating a Digitization Workflow and Maximizing Limited Resources

Whitney Alexander
Director of Technical Services

David Brian Holt
Electronic Services Librarian

Why should your library have a digitization service?

- Document delivery
- Inter-library loan services
- Preserving archival materials
- Improving access for patrons who are visually-impaired
- Improving discoverability

The Future is Digitization-on-Demand

Several law libraries have already begun digitization-on-demand services for materials in the public domain

Why is digitization-on-demand so exciting for libraries?

- Helps to "triage" what materials should be digitized first based on patron demand
- Recognizes that libraries simply do not have the staff nor budget to digitize everything

Vendors are already responding to digitization-on-demand

Outsource or buy your own equipment?

If a library has only a small collection of materials to scan, it may be unwise to purchase digitization equipment as the return on investment will be too low. There are a number of companies that provide digitization services, some of them at a surprisingly low cost.

HEINONLINE

If you want to purchase equipment

It is comparatively inexpensive to purchase back issues of your student journals from Hein. What if you can't afford that or what if your journal isn't on Hein?

You can purchase digitization equipment if you have the budget for this. Major vendors include ATIZ, BetterLight, Digital Library Systems Group, i2S, Indus, Kirtas, Konica/Minolta, Microbox, Phase One, SMA, Tarsia, Treventus, ZBE and Zeutschel. You can also try building your own bookscanner. This is MUCH cheaper and uses standard SLR digital cameras that are easily replaceable. There are lots of designs available at http://www.diybookscanner.org.

The Evolution of Book Scanning

In the beginning, there was the flatbed scanner. This technology has proven to be poorly suited for book scanning because it is: slow, difficult to use, destructive to the book binding, and produces poor images particularly near the area of the book binding. Pages must be scanned individually by hand.

The Planetary Book Scanner

The second phase of book scanning technology is the planetary book scanner. This is the first scanner truly designed to scan bound materials. Disadvantages of this technology include: single CCD to capture both pages, need for page curvature correction software, and margin crawl (the center of the book moves as the user turns the pages). One advantage however is that these machines are typically easier to use (because it is one unit) and may be used as a kiosk scanning station.

The V-Shaped Book Scanner

Some institutional and library scanning of bound materials is done using a v-shaped book scanner. This eliminates the need for page curvature correction software as the pages lay flat. It is also easier on delicate bindings as the book doesn't need to be opened 180 degrees. It also has a separate CCD for each page.

Advantages of a v-shape book scanner

- Doesn't require page curvature software
- Can be easily upgraded by replacing the cameras (which may be standard SLRs that are widely available)
- There is no margin crawl as the book is held in place in the cradle
- The cost is roughly comparable to a single-CCD planetary scanner but should produce better quality images and higher OCR accuracy

What about robotic page turning?

Robotic page turning is still very expensive. Even entry-level book scanners with robotic page turning start at $60K. These machines are capable of scanning around 2500 pages an hour. For about 1/6th the cost, you can acquire a non-robotic v-shaped book scanner that can produce around 700 pages an hour.

SANTA CLARA UNIVERSITY
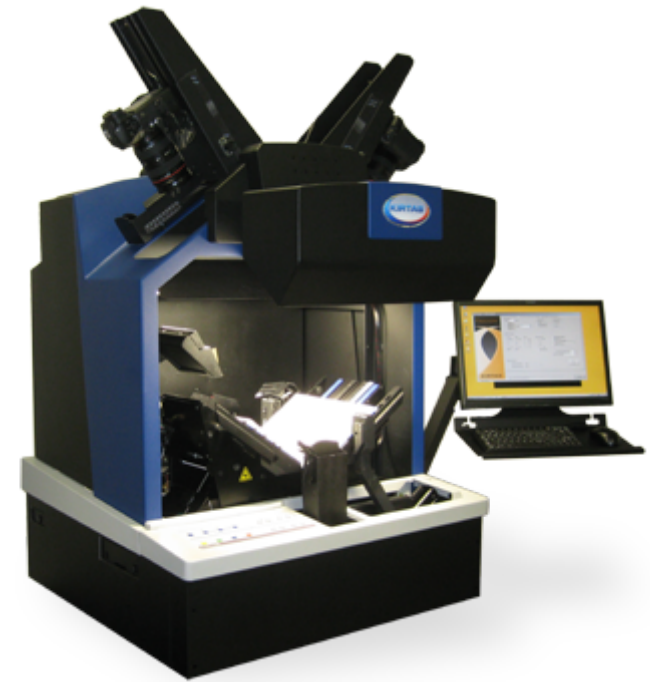
Issues to consider

- Scanners based on standard digital cameras would be much easier to upgrade
- A digital camera that can produce a DPI high enough for OCR software is fairly expensive (with the model that comes included with the Atiz Bookdrive, we'd be unable to produce newspaper size images that would be OCR compatible)
- A platen must be moved after scanning each page, this may cause repetitive movement injuries among staff
- A platen also, however, negates the need for page curvature correction software

| Name:	Scanner Type:	Capture software Included:	Post-processing/OCR software:	Warranty:	Advantages/Disadvantages:	Maximum scan size:	Price:
Ristech Kiosk Book Scanner	Planetary with single CCD	Proprietary embedded system with a touch-screen monitor	Outputs to PDF, unknown if OCR included	Three years	Kiosk-style scanner that could be used by students after digitization project winds down. Can use SmartPrint system.	18" x 24"; spines up to 12cm	$35,800 for system; $3,095 for software. **Grand total of $38,895**
Kirtas i2S e-Scan Kiosk	Planetary with single CCD	Proprietary embedded system with a touch-screen monitor	Outputs to PDF, unknown if OCR included	One year	Kiosk-style scanner that could be used by students after digitization project winds down. Can use SmartPrint system.	14" x 20.5"; spines up to 4"	$11,650 for system; $1,970 for book cradle. **Grand total of $13,620**
Kirtas Skyview 3525	Planetary with single CCD (DSLR camera)	Proprietary Windows-based system (includes computer hardware)	Unknown	90 days	Can scan maps, large documents, newspapers, etc.	25" x 35"; unknown for book spines (probably 4")	**$68,000**
Atiz BookDrive Pro	V-shaped scanner with two CCDs (DSLR cameras)	Proprietary Windows-based system (hardware not included)	Outputs to PDF, no OCR software	90 days	Can easily be upgraded. Major vendor. Being used by Google, Stanford, UCLA, etc. Includes auto capture switch so machine can be used without pressing buttons.	16.5" x 24.2"; spines up to 11cm	**$17,020**
Atiz BookDrive Mini	V-shaped scanner with two CCDs (DSLR cameras)	Proprietary Windows-based system (hardware not included)	Outputs to PDF, no OCR software	90 days	Can easily be upgraded. Major vendor. Being used by Google, Stanford, UCLA, etc. Does not include auto capture switch. User must press button for each scan.	10" x 15"; spines up to 5cm	**$8,895**
Book2Net Spirit Scanner	Planetary with single CCD	Proprietary embedded system with a touch-screen monitor	Outputs to PDF, unknown if OCR included	One year	Kiosk-style scanner that could be used by students after digitization project winds down. Can use SmartPrint system. Rave reviews from other libraries.	13.82" x 19.21"; spines unknown	Reportedly around **$9000**
Kirtas CopiBook BW	Planetary with single CCD	Proprietary embedded system with a touch-screen monitor	Outputs to PDF, unknown if OCR included	One year	Kiosk-style, easy to use. Black and white only.	16.5" x 25.2"; spines up to 3.9"	**$22,000**
Zeustchel Zeus	Planetary with single CCD	Proprietary Windows-based systems	Outputs to PDF, unknown if OCR included	90 days	Very easy to use; kiosk-style, touch screen.	18.1" x 25"	**$13,650**

Digitization Equipment Vendors

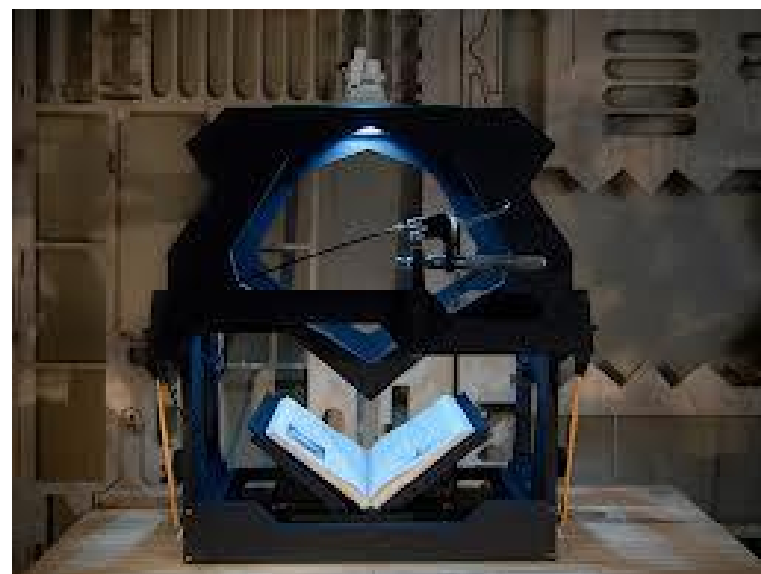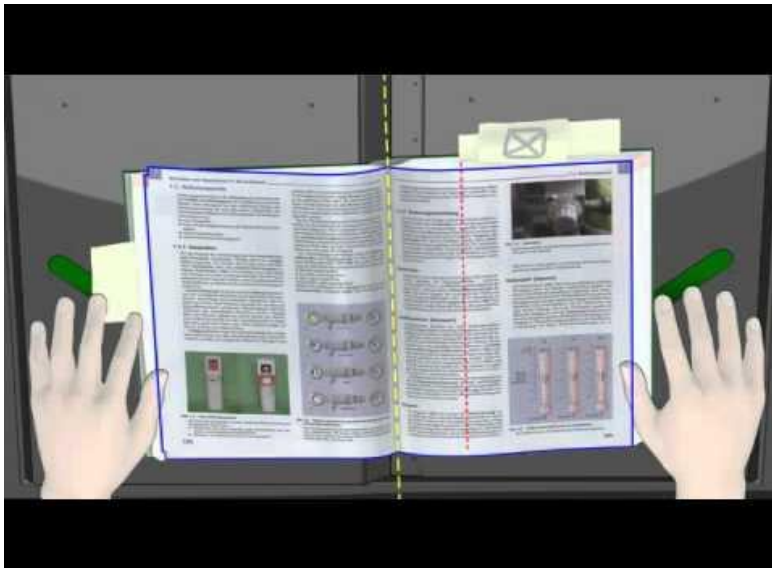# The Crowley Company

Kirtas Technologies
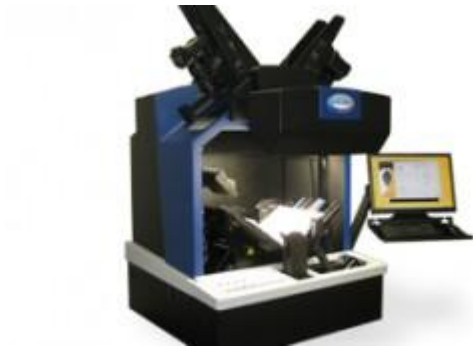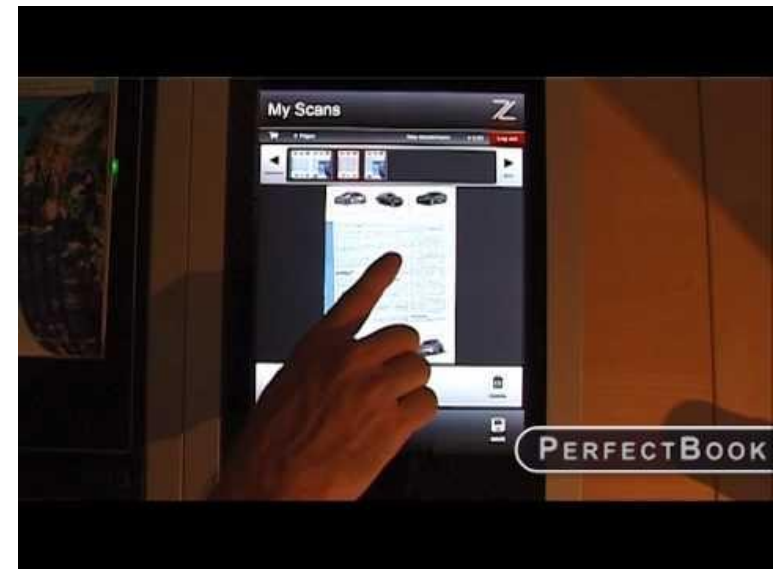
Atiz

Do it yourself!

What we purchased at Santa Clara Law

Zeutschel OS 12000C

Zeutschel Zeta

- Can be easily upgraded
- Software is being constantly updated
- Large enough for legal newspapers
- Good price (~$18,000)

- No platen to move
- Excellent price ($10,000)
- Easy to use touch screen
- Can be used by students after the digitization project has slowed down

Review of overhead scanners

Jody L. DeRidder, Overhead scanners: reports from the field. 29 Library Hi Tech 9 (2011).

Workflow Management

Distributing workflows economically

- Cross training across departments
- Work with technical services and circulation
- Give individual staff members responsibility for a project
- Work with library science interns! These projects make GREAT virtual internships (check out http://slisweb.sjsu.edu/current-students/courses/internships/virtual-internships)

A Few Examples .... and the metadata

I will cover these topics:

- Very brief overview of considerations for digital projects
- Examples of problems (even in a straightforward project)
- A small project from start to (almost) finish
- A few words about metadata
- Summation

Initial project considerations

- Copyright considerations
- Where will digital files be stored?
  - Local database
  - Commercial database (ex: digital commons)
- What resolution (300 dpi, 600 dpi, less)?
- What format (PDF, TIFF, both, other)?
- How will you handle graphics?
- How will you handle analog content?
- How will you handle video and audio content?

Initial project considerations - Discoverability

- Run OCR software over scanned documents? If so, do OCR cleanup or leave it as raw text?
- Create indexed text files?
- If not, what about visually impaired users?
- And what about submission to discovery service platform?

Some digitization projects are relatively straightforward ...

- ... and if you believe that I have some land west of San Francisco ...
- even straightforward projects aren't straightforward
- case in point: the Watergate Hearings

The Watergate Project

- Cong. Don Edwards' annotated papers from the Watergate hearings
- Typed, one-sided leaves in binders (70 binders to be exact)
- http://digitalcommons.law.scu.edu/watergate/

WATERGATE HEARINGS

This series is a collection of papers produced during the Watergate Hearings that were donated by Congressman Don Edwards to the Heafey Law Library. Don Edwards (born January 6, 1915) is an American politician of the Democratic Party, formerly a member of the United States House of Representatives from California. Edwards was elected as a Democrat to the 88th from the 10th Congressional District

Watergate Hearings, a few problems

- Yuck!!
- What do you do? Especially when there are hundreds of pages in the collection look like this?
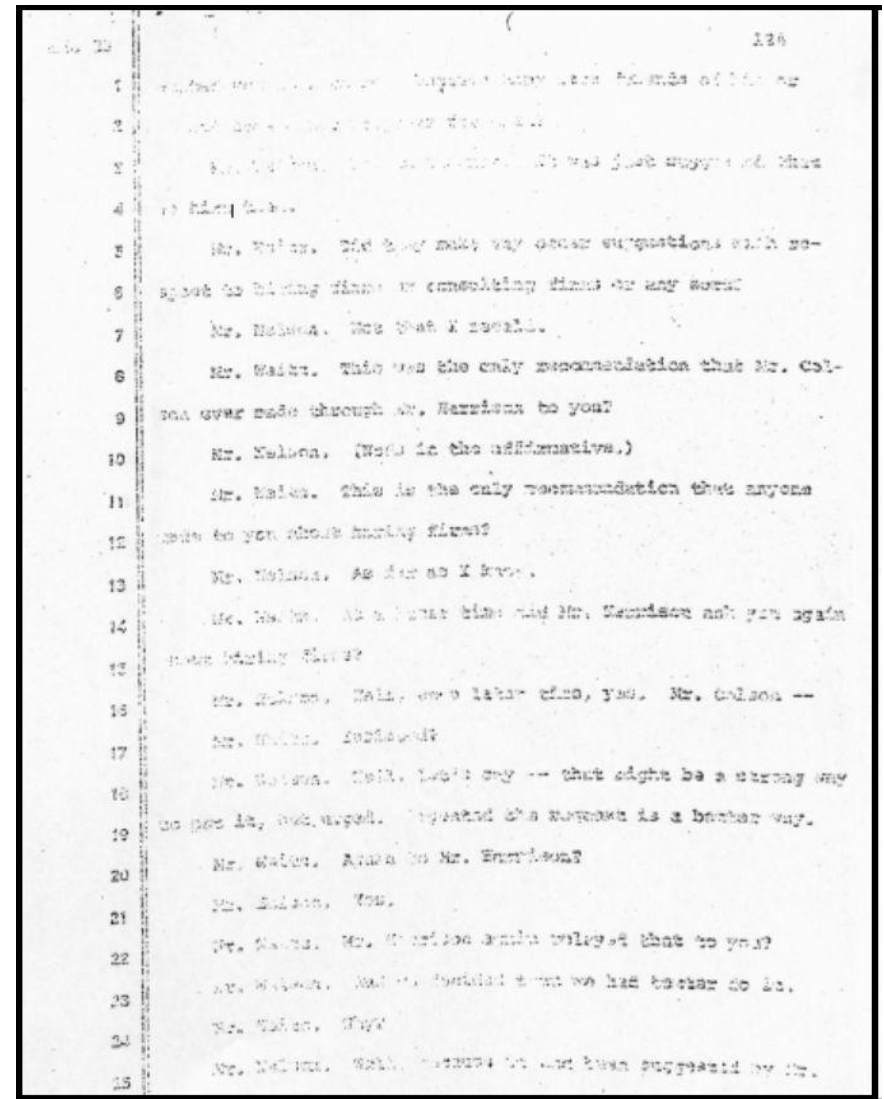
Watergate Hearings, a few problems

- The set has many annotations made by Rep. Edwards
- What should you do with them? Transcribe as OCR readable data?
- What about the annotations that are hard to read, should they be transcribed, if possible?

```
K   Yep. It was not improper.
E   Right.
K   And there was nothing illegal about it.
E   See, he's the house lawyer.
K   Yep, exactly and I just couldn't believe that you and Bob and the President
    just too good friends to ever put me in the position where I'd be putting my
    family on the line.
K   And it's just unbelievable, unthinkable. Now shall I just—I'll just if I'm
    asked by Silver I'll just lay it out just exactly that way.
E   Yeah, I wouldn't haul the President into it if you can help it.
K   Oh, no, I will not.
E   But I think the point that which I will make in the future if I'm given the
    chance that you were not under our control in any sort of a slavery sense
    but that we had agreed that you would not be at the beck and call of the
    committee.
K   And, of course, too, that I act only on orders and, you know, on direction
```

K implicates H & the Pres, staff says is only implication of P's involvment.

Watergate Hearings, a few problems

- Other issues:
  - A single PDF file per binder would be much too large to download from the Digital Commons
  - How to split up the binder into manageable size files?
  - This is a very large project. How can it be divided among multiple people without any overlap
    - How to maintain scanning and metadata quality across all participants
    - Communication is paramount

Digitizing ... its not just scanning

- Collections may have non-print material in addition to print
- In addition to scanning print items, we digitize analog video and audio materials
- We have also started a fiche digitization project
- Once the word gets out that you can digitize audio and video, you'll find that **everyone** has analog materials they need digitized

A project from the beginning

- Maria, our main scanner
- This is the story of the Bench & Bar Historical Society of Santa Clara County papers.
- Sounds boring, doesn't it?
- I thought so too, at first …

Bench & Bar Hist. Soc.

- A collection of annual mock trials and lectures held by the Society
- Each mock trial or lecture has a video on CD as well as accompanying print documents
- The first consideration was how to scan the documents
  - each page as a separate file
  - or create one document per trial/lecture
- We chose one doc. per trial/lecture
- What format for the video? mp4
- Obtained permission from donor include collection in our digital commons
- As example, we'll look at SJ v. Paris

Bench & Bar Hist. Soc.- SJ v Paris

The not-so-original? Eiffel Tower.

File photo

- A copyright infringement case concerning SJ Light Tower (1881) and Eiffel tower (1889)
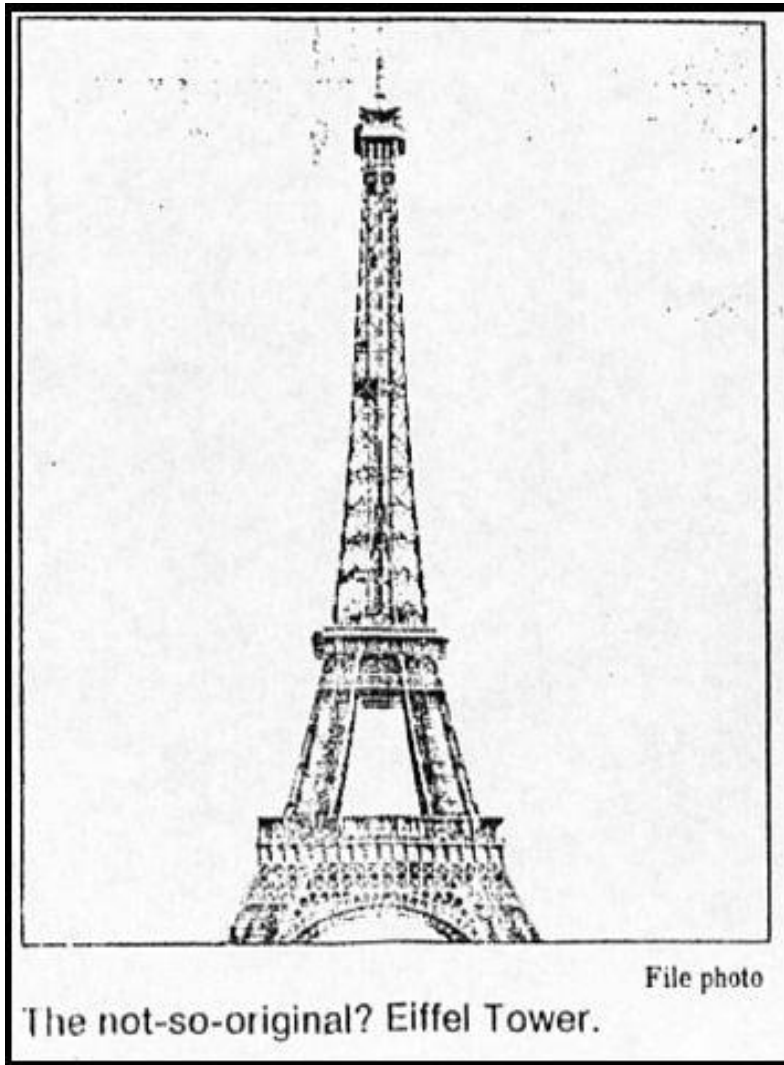
Bench & Bar Hist. Soc.- SJ v Paris

City of San Jose vs. City of Paris

San Jose Bar Society
January 25, 1990

The Bench and Bar Historical Society
of Santa Clara County

2001 Gateway Place, Suite 220 West   •   San Jose, CA 95110-1013   •   (408) 288-8840   (408) 299-4727
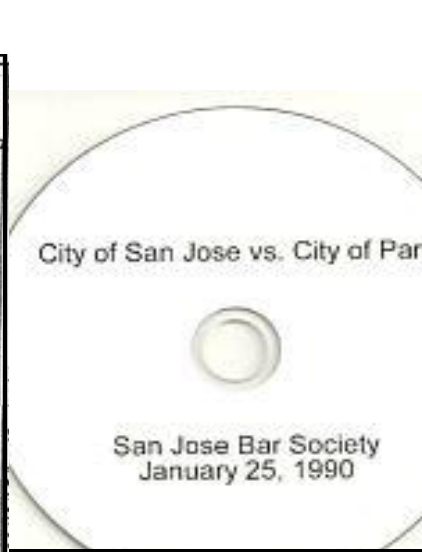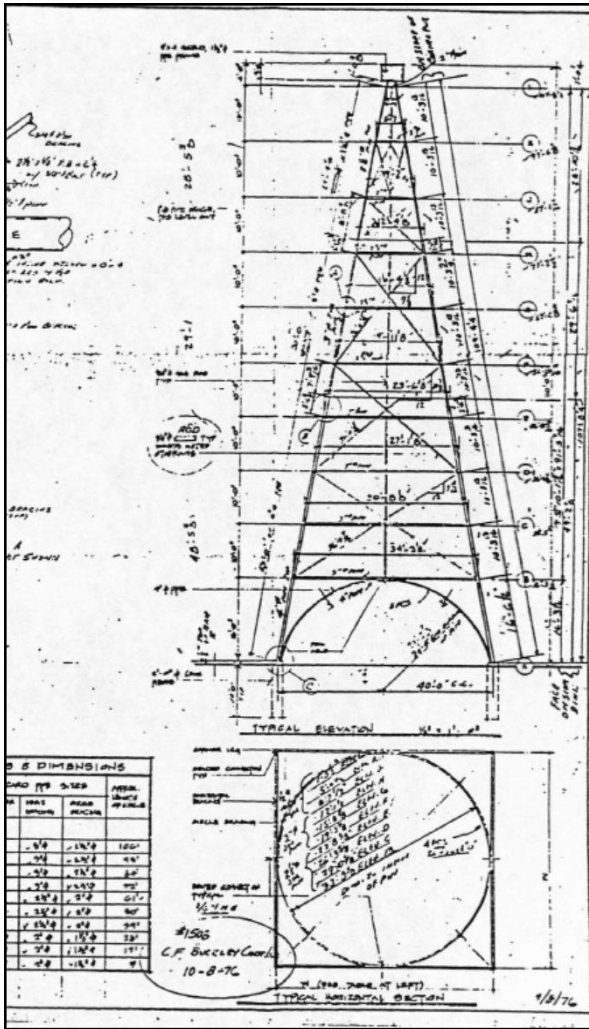
BOARD OF DIRECTORS

Geoffrey J. Wright
  President
Allan T. Hikoyeda
  Secretary
Judy A. Fuller
  Treasurer
Hon. Robert P. Aguilar
Victor H. Beauzay
Hon. Robert Beresford (Ret.)
Paul R. Bernal
Hon. Robert M. Foley
Philip L. Hammer
Richard G. Mansfield
Hon. Marilyn Morgan
Louis Oneal
Willys I. Peck
Philip H. Pennypacker
Hon. Eugene M. Premo
Donald B. Richardson, Jr.
Hon. Conrad L. Rushing
Hon. Mark E. Thomas, Jr.
Gerald F. Uelmen

December 19, 1989

Jacques Chirac, Mayor
The Office of the Mayor of Paris
The City of Paris
Paris, FRANCE

Dear Mssr. Chirac:

    I represent the Bench and Bar Historical Society
of Santa Clara County, located in California on the

THE CITY OF SAN JOSE,   )
CALIFORNIA,   )
    Plaintiff,   )
)
vs.   )
)
THE CITY OF PARIS, FRANCE;)
  FRANCE; THE ESTATE   )
  OF MSSR. GUSTAVE   )
  EIFFEL; DOES 1   )
  THROUGH 50,   )
    Defendants.   )
)

COMPLAINT FOR INFRINGEMENT
OF COPYRIGHT AND UNFAIR
COMPETITION, DESIGN PATENT
INFRINGEMENT, COPYRIGHT
INFRINGEMENT OF WORK
ART

1.   The above named Plaintiff, THE CITY OF SAN JOSE

does hereby file this SPURIOUS complaint against the

SAN JOSE DAILY MERC

NEW SERIES.   SAN JOSE, CALIFORNIA. SUNDAY MORNING. DECEMBER 25, 1881.

Bench & Bar Hist. Soc.- Metadata

- This is a PDF list we received with the collection and a text file derived from it

Bench & Bar Hist. Soc.-- Metadata

- Metadata was included in collection as a PDF file (probably originally a MS Word file)
- PDF was exported as a text file
- The text file was run through a python script for cleanup and formatting
- The new text file was then ready for import into a Digital Commons batch load Excel spreadsheet
- PDF and video files are uploaded to public file on Dropbox for harvesting by DC

Bench & Bar Hist. Soc.-- Metadata

- Formatted text file
- Excel file ready for batch upload

A little more on metadata

- At CALI last year I presented a method for batch loading metadata for the backfiles of our three student law reviews
- Method involved gathering metadata from several online sources and combining it into an Excel spreadsheet
- Used Excel functions to parse the metadata into the correct form
- Information available at: http://digitalcommons.law.scu.edu/librarian/8

Other projects - video conversion

- We are currently digitizing all of the analog videos in our collection
  - many of our videotapes are in bad condition
  - digital format is easier for faculty to use for classroom instruction
  - purchase equipment necessary for digital conversion
  - identify which are available for purchase in digital format
  - videos are burned onto DVD-ROMS and some are uploaded to YouTube

Other projects - audio conversion

- We are currently digitizing selected audio tapes that are not available for purchase digitally
  - digital format is easier for faculty to use for classroom instruction
  - purchase equipment necessary for digital conversion, available at any electronics store

Summing it all up...

- Examine collection contents and decide how best to digitize the items
  - what format are they in
  - how well will the collection scan
  - do you want print documents to be OCR readable for full text access
  - should you transcribe audio and video files
  - where will the files be stored
- Metadata (when available) comes from many different sources and in all shapes and sizes
  - metadata in electronic format can be manipulated to fit your needs

Thank you!

Whitney Alexander
Director of Technical Services
walexander@scu.edu

David Brian Holt
Electronic Services Librarian
dholt@scu.edu

Presentation available at: http://bit.ly/11CuAxT