Digital Libraries on a Shoestring

Walter Nelson

RAND Corporation

walternelson.com

Introduction

- Digital libraries – not so tough
- Core functions can be performed by standard or easily acquired tools
- High end KM (Stellent etc.) systems are not necessarily geared towards library needs
- Usually WAY too expensive

The Shoestring takes many forms

- Limited funds
- Limited labor
- Lack of technical expertise
- Lack of management interest/support
- Institutional resistance to new or complex applications

My Context

- This is an composite of multiple projects for multiple organizations
- There is no one place where I have done all of these
- Some of my sites are publicly available, most are not

I don’t know what I don’t know

- This is from my personal experiences
- There are many interesting things with which I have no direct experience
- Mysteries include Google books & Content DM
- When I get Punditish and Existential it’s because I have had to confront these issues

Where things are headed

- Digital content is being absorbed into the information cloud (Google books etc.)
- You should consider the value of adding your stuff to the cloud
- If your stuff is already part of the cloud, you should consider the costs/benefits of cataloging it

If it isn’t on Google, does it exist?

- If you have content available to the public, do people need to search your interface to find it?
- If so, you may want to rethink that
- If it’s hard to find, will your audience have the patience to find it?

Essential Elements of a Digital Library

- Digital objects
- A way of finding them
- A way of delivering them
- You can refine your tools, but that is the essence

Get out of that box!

- The digital networked world has changed everything
- Be prepared to embrace “non-traditional” tools
- If it gets the job done, it’s the right tool

Digital Archive (not)

- Digital libraries and archives are not the same thing
- Archival storage may:
  - Retain native format
  - Use high resolution imaging
  - Have multiple storage locations
  - Have “bit rot” and version update plan
- Archives are more about preservation than user experience
- Archive can be source for library

Part 1 – The Digital Object

- “Digital object” comes in many forms
- Self contained document (PDF, DOC, XLS, PPT etc.)
- Media (image, video, audio)
- Web page (requires additional files)
- Application or database requiring special software
- Anything that has a distinct identity and can be linked to

COPYRIGHT!!

This all assumes you have the rights to post the content in question.

'Nuff said

Where Does it Live?

- Exists on your servers
- Exists on the big internet in “the cloud”
- Provided by vendor (e-books)
- Managed by vendor (Serials Solutions etc.)
- I will focus on things you own
- If it is vendor provided – create bib record and link out – if you decide to catalog it at all

Taming Your Digital Object

- Item can be opened by your target audience
- Item is only as large as it has to be
- Item is findable by relevant tools
- This will require your intervention to do it right

PDF is King of Multipage Documents

- When ever possible, go with PDF for documents
- Broad acceptance
- Most users can open them
- Industry standard
- Likely to be supported for a while
- Use JPEG for images

“Born Digital” PDF

- Digital documents converted to “vector” PDF
- MS Word, Powerpoint etc.
- HTML to PDF captures images – format will be distorted
- Only keep “native” format if there is some special feature not supported by PDF

Word to PDF

- Uncheck “show markup”
- Check for attachments
- Go for it
- Prevents editing, removes annoying spell and grammar check marks
- Avoids MS Office compatibility issues

Scan to PDF

- Print to digital migration
- Creates “Bitmap PDF” (based on TIFF standard)
- Easiest with documents you can un-bind.
- Books require complex machines

Flatbed Scanner

- Slow
- For images, not text pages
- Not a viable choice for documents

Desktop Scanner

- Canon DR-5010C
- $2-$3 K

Book Scanner

- More complex and expensive
- $10,000 and up
- Perhaps you can McGyver it

Multi-Function Devices

- Copy machine/printer/fax/scanner
- Not "best of breed"

Outsource

- Good if you have the money but not the labor
- Price and quality vary
- Much work is done offshore
- The more you scan, the cheaper it gets
- You might be in a position to partner with Google

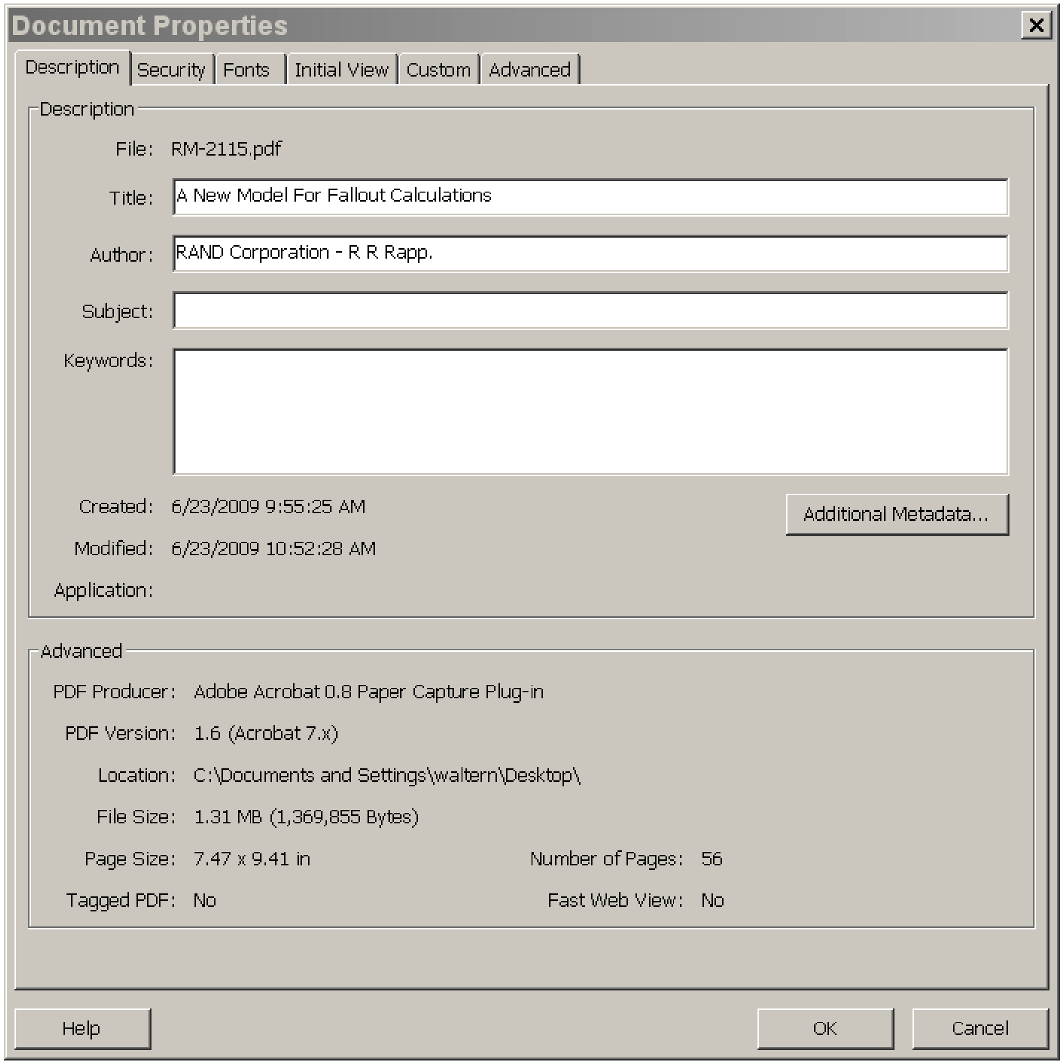
Making Your PDF Web-Friendly

- Best scan at 300 dpi/Black & White
- OCR (creates invisible ASCI text overlay and deskews pages)
- OCR on B/W can make file smaller
- Color and grayscale are larger, and with Adobe Acrobat, grow in size if you OCR
- “Reduce File Size”

Make Your PDF Findable

- Go to “File>Properties”
- Under “Description” enter the title and author
- “Born-digital” files will have dumb titles
- Scanned files will have none
- “Title” is what appears in a Google Search

PDF With Proper Title



Description

File: RM-2115.pdf

Title: A New Model For Fallout Calculations

Author: RAND Corporation - R R Rapp.

Subject:

Keywords:

Created: 6/23/2009 9:55:25 AM

Modified: 6/23/2009 10:52:28 AM

Application:

Advanced

PDF Producer: Adobe Acrobat 0.8 Paper Capture Plug-in

PDF Version: 1.6 (Acrobat 7.x)

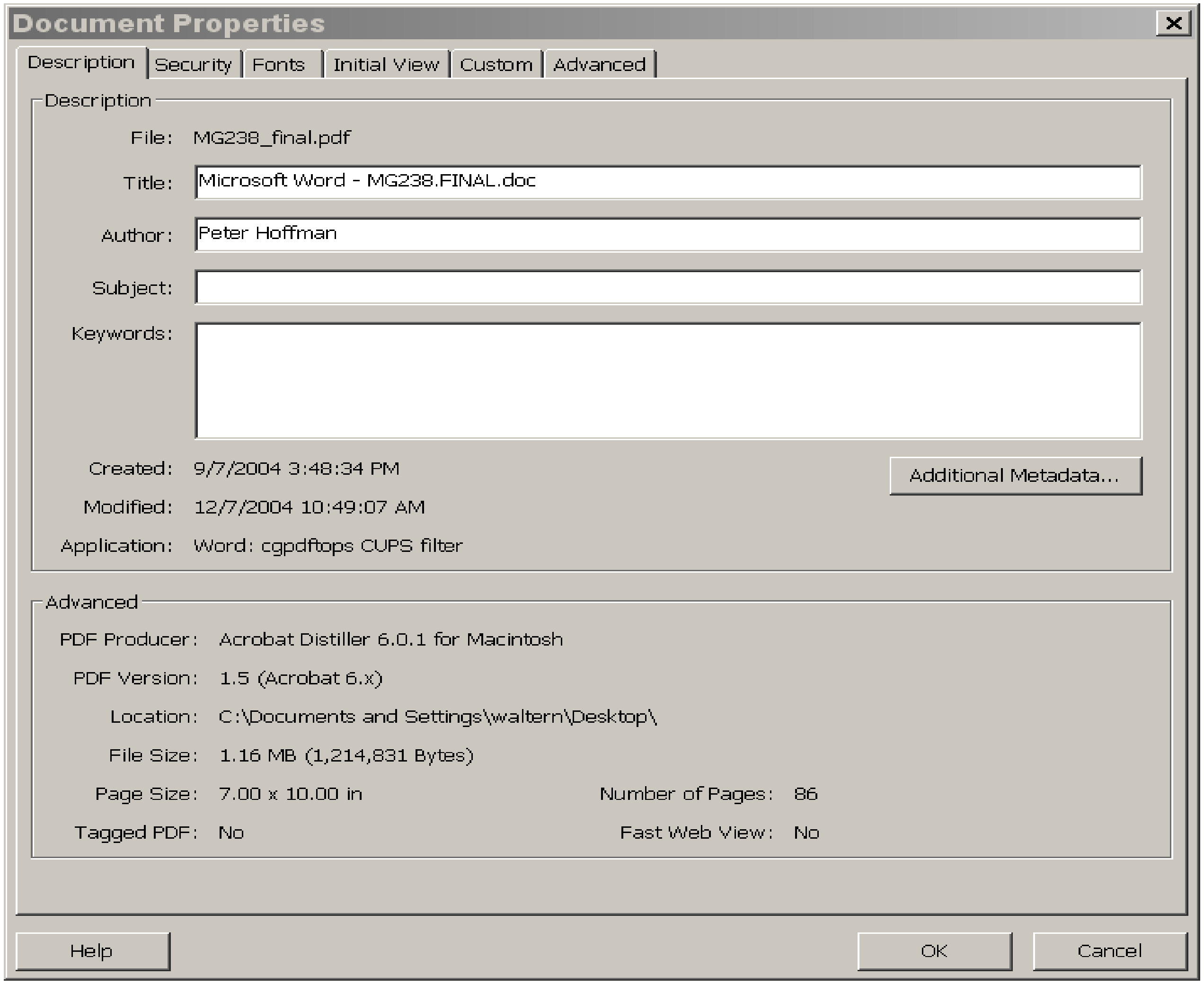
Location: C:\Documents and Settings\waltern\Desktop\

File Size: 1.31 MB (1,369,855 Bytes)

Page Size: 7.47 x 9.41 in   Number of Pages: 56

Tagged PDF: No   Fast Web View: No

Typical Title/Author



Description

File: MG238_final.pdf

Title: Microsoft Word - MG238.FINAL.doc

Author: Peter Hoffman

Subject:

Keywords:

Created: 9/7/2004 3:48:34 PM

Modified: 12/7/2004 10:49:07 AM

Application: Word: cgpdftops CUPS filter

Advanced

PDF Producer: Acrobat Distiller 6.0.1 for Macintosh

PDF Version: 1.5 (Acrobat 6.x)

Location: C:\Documents and Settings\waltern\Desktop\

File Size: 1.16 MB (1,214,831 Bytes)

Page Size: 7.00 x 10.00 in

Number of Pages: 86

Tagged PDF: No

Fast Web View: No

The Housebroken PDF

- Modest file size (1 – 8 Mb)
- Text that can be spidered by search engines and searched by readers
- A title that describes the contents
- Security settings can have unintended consequences

Software

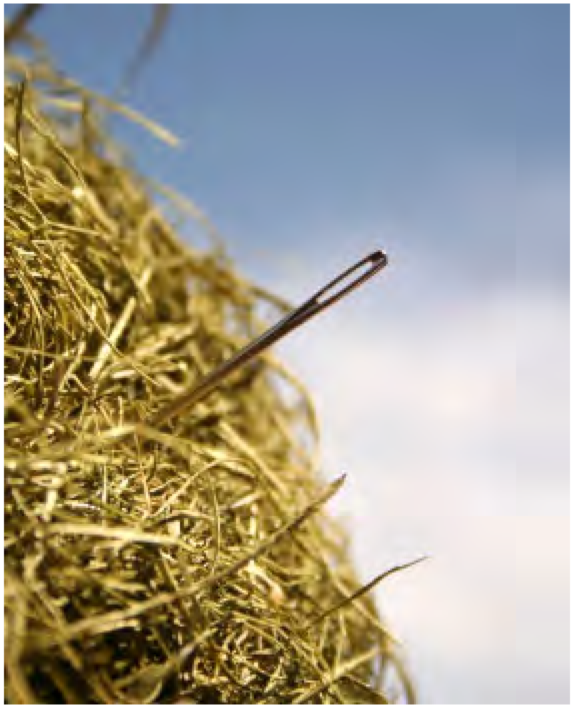
- Adobe Acrobat Pro does OCR and optimization pretty well
- 3rd party software can do specialized tasks (OCR, Web optimization) a little better

E-Books and Readers

- Amazon Kindle, Sony Reader, iPhone apps etc, etc. etc.
- Some proprietary formats
- PDF Support?
  - Sony – yes
  - Nook – yes
  - Kindle – no
- Don’t embrace new formats unt it all settles down

Part 2: Finding the bloody thing

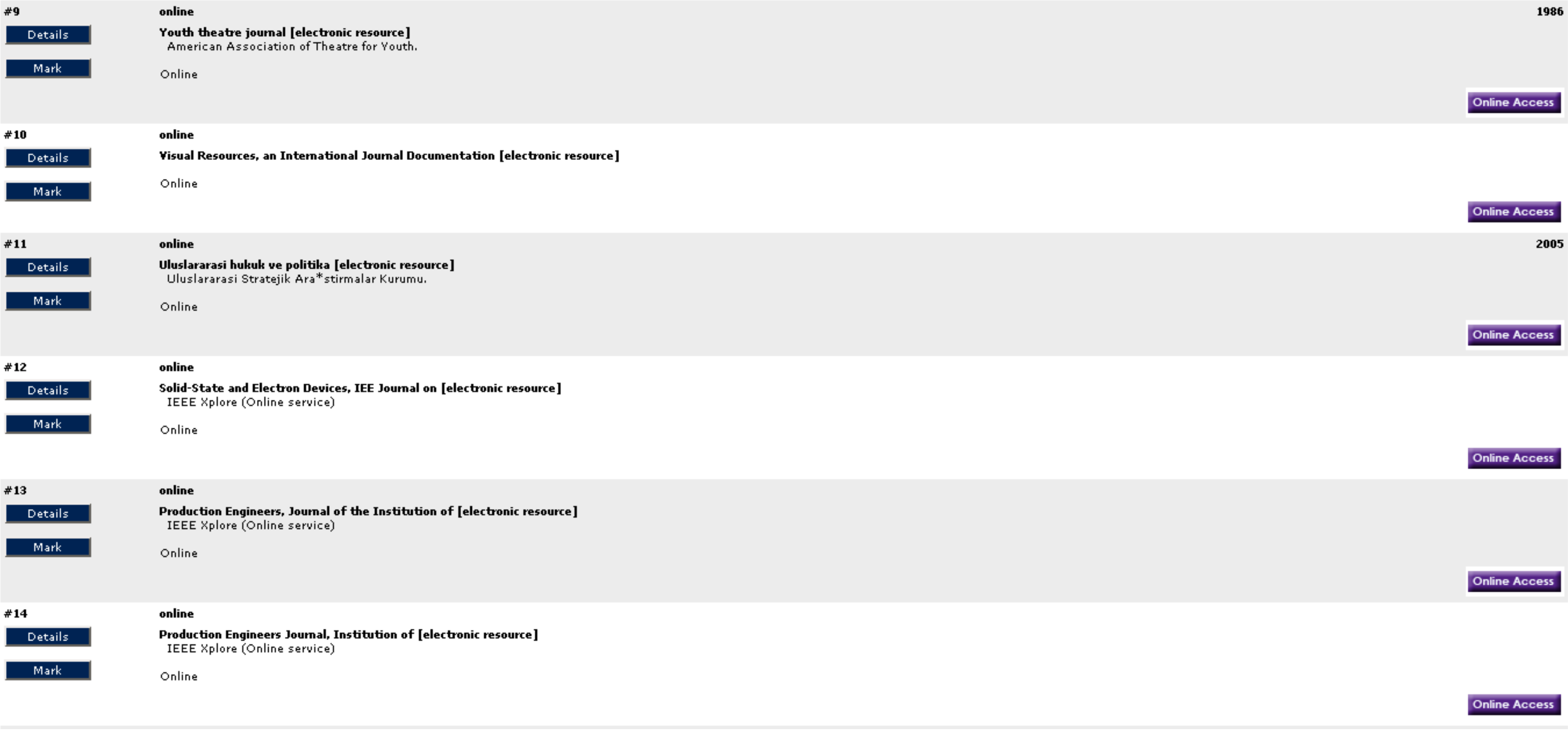
- Load the PDF on your server
- Create something to point to it
- I have often had to get pretty creative



Option 1: The OPAC

- Create a record on your online catalog
- Link to the file from the OPAC record
- MARC field 856
- Pretty basic

SIRSI/DYNIX OPAC



#9 online 1986

Youth theatre journal [electronic resource]
American Association of Theatre for Youth.

Online

#10 online

Visual Resources, an International Journal Documentation [electronic resource]

Online

#11 online 2005

Uluslararasi hukuk ve politika [electronic resource]
Uluslararasi Stratejik Ara*stirmalar Kurumu.

Online

#12 online

Solid-State and Electron Devices, IEE Journal on [electronic resource]
IEEE Xplore (Online service)

Online

#13 online

Production Engineers, Journal of the Institution of [electronic resource]
IEEE Xplore (Online service)

Online

#14 online

Production Engineers Journal, Institution of [electronic resource]
IEEE Xplore (Online service)

Online

OPAC Pros

- If you are integrating digital and paper content, the OPAC is the logical choice
- If your customers are already used to using it, it’s a good place to put new stuff
- Good for large collections

OPAC Cons

- If you don’t have one, getting one is a big deal
- Most catalogs not open to search engines
- OPAC Boolean search is obsolete
- OPAC technology is struggling to keep up

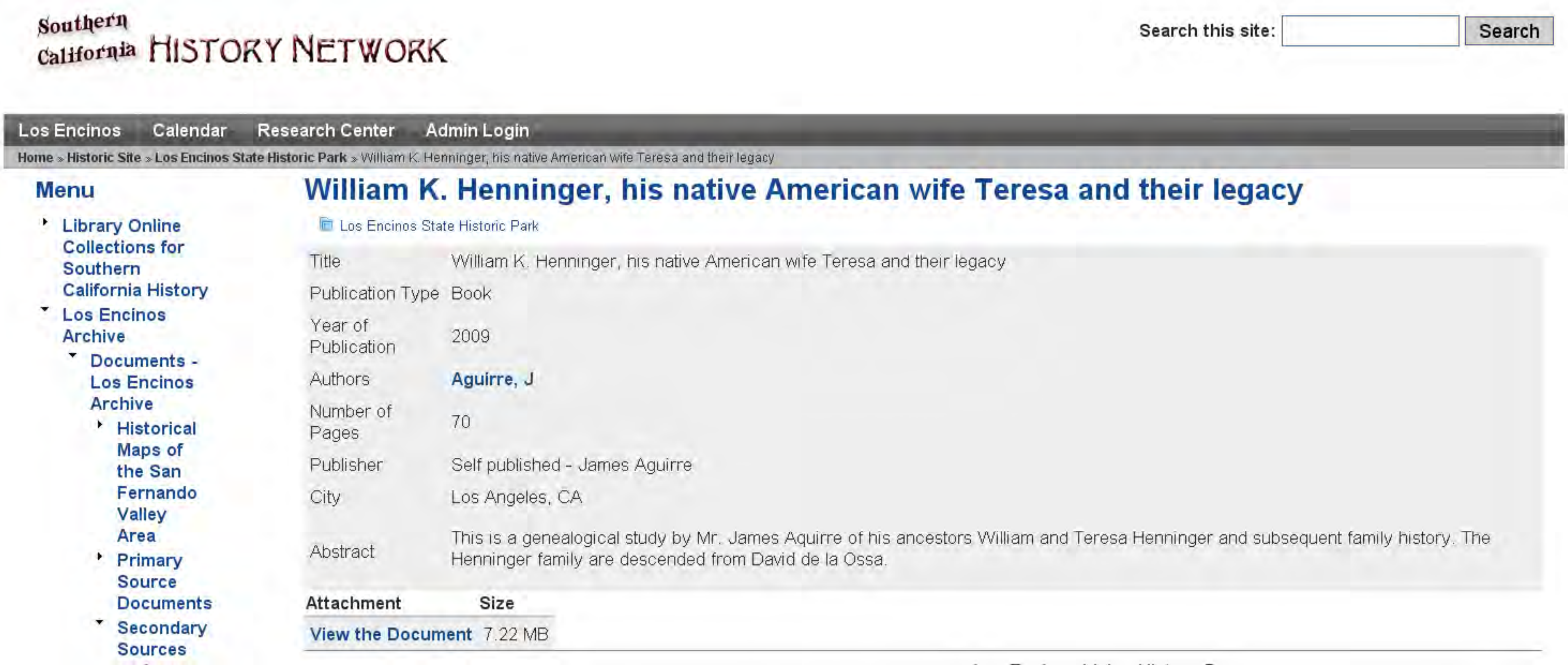
ILS Future

- The day may come when you don’t need to check out a book or check in a serial
- The book may be eternal, but the ILS may not be
- For special libraries, the day may have already arrived

Option 2: Drupal

- Drupal is open source
- Includes search engine
- Web based, multiple authors
- Includes “biblio” and “faceted search” modules

A Drupal Bib Record



Southern California HISTORY NETWORK
Search this site:
Search
Los Encinos
Calendar
Research Center
Admin Login
Home » Historic Site » Los Encinos State Historic Park » William K. Henninger, his native American wife Teresa and their legacy
Menu
Library Online Collections for Southern California History
Los Encinos Archive
Documents - Los Encinos Archive
Historical Maps of the San Fernando Valley Area
Primary Source Documents
Secondary Sources

William K. Henninger, his native American wife Teresa and their legacy

Los Encinos State Historic Park

Title	William K. Henninger, his native American wife Teresa and their legacy
Publication Type	Book
Year of Publication	2009
Authors	Aguirre, J
Number of Pages	70
Publisher	Self published - James Aguirre
City	Los Angeles, CA
Abstract	This is a genealogical study by Mr. James Aquirre of his ancestors William and Teresa Henninger and subsequent family history. The Henninger family are descended from David de la Ossa.

Attachment	Size
View the Document	7.22 MB

Drupal Pros

- Open Source (i.e. free)
- Easy to install, configure and use
- Manages multiple users, complex permissions
- Constantly being updated and expanded
- Creates pages open to Search engines
- Built in search engine

Dupal Cons

- Open Source – you have to become an expert
- Constant updates – can’t keep up
- Features may not be supported from one version to another
- Scale may be a problem for a very large collection

Option 3: Sharepoint

- Microsoft product
- Designed for collaboration
- Built in search tool
- Does all sorts of things (document library, wiki, calendar etc.)

Sharepoint Pros

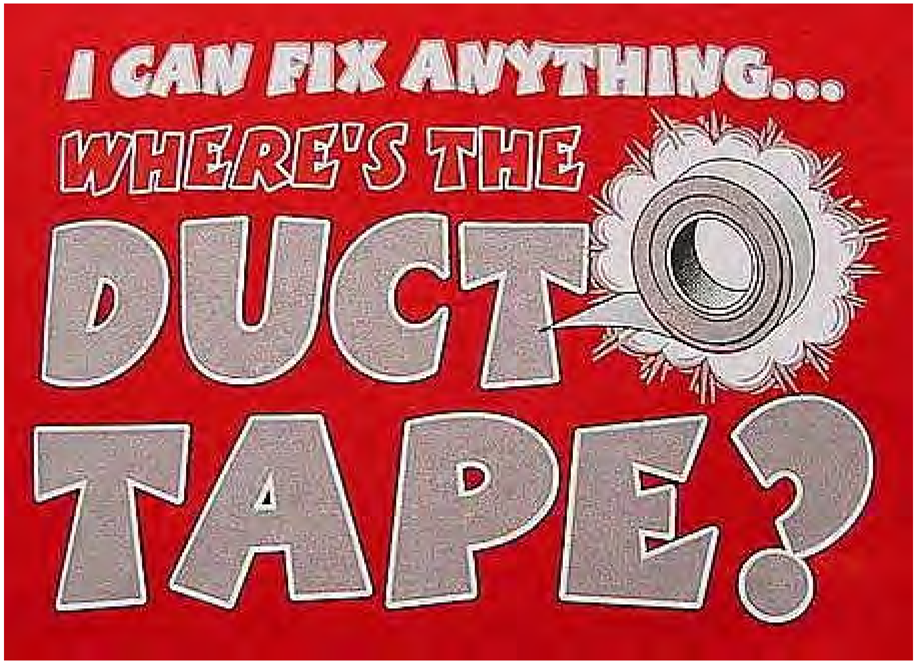
- If it’s all you have, maybe it’s better than nothing
- Good at controlling access to individual sensitive items
- Permits broad participation

Sharepoint Cons

- Requires MS backbone – can be expensive
- It does many things – but doesn’t do any of them well
- Search is BAD
- Confusing to use and administer
- Standards hard to maintain - chaos sets in
- Not really scalable

Kluges and Make-Dos

Sometimes you just have to do the best you can with the tools at hand



The Blog

- Use it for regularly updated content
- Each blog entry becomes a bibliographic record
- Generates RSS feed
- Categories and tags
- Search tool and search engine friendly
- Good for posting your organization’s content
- Not infinitely scalable

HTML + Search Engine

- Build HTML pages that serve as bibliographic records
- Records link to digital objects
- Search engine retrieves items with ordinary search
- Drupal does this sort of thing better

Full Text Search (not a big deal)

- Use standard search engines (Google, Vivisimo)
- Digital objects must have searchable text
- Born digital already there
- Scanned items must be OCRed.
- Will help if there are hyperlinks to each item to facilitate spidering

Free-Wheeling It!

- Do not create bib records
- Ensure that attached metadata are good (Title, author, key words etc.)
- Ensure that full text is searchable
- Use your search engine exclusively
- The future?

The Search Engine & You

- Your non OPAC content is being spidered
- People will find it if you make it findable
- People may stumble on your stuff without ever seeing your “home page”

When people come in the side door

- Have good navigation on all your web pages
- Consider a “cover sheet” on your PDFs, linking back to you and stating copyright terms

Work & Play Well With Others

- Consider how your content will interact with current search and KM systems
- Is your content in a form that will be compatible with future systems?
- I suggest XHTML and PDF
- OPAC alone is too insular

Final Thoughts

- It’s past time to attend to this
- If your library isn’t digital, it has a very limited future
- You don’t have to buy a $50,000 KM system to make this happen

Find me

Walter Nelson

http://walternelson.com

walter@walternelson.com

Look for me on Linkedin

http://www.linkedin.com/in/waltertnelson