Beyond “This material is unprocessed”

MINIMALLY DESCRIBING AND PROCESSING BORN DIGITAL COLLECTIONS

PRESENTED BY:
LAURA UGLEAN JACKSON, ASSISTANT UNIVERSITY ARCHIVIST, UC IRVINE

Description

+

Min. Processing

My presentation…

1) Experience minimally processing Hillis Miller born digital files
2) What I learned
3) Born Digital Processing Framework Group

J. Hillis Miller

- UCI faculty member
- Papers part of critical theory collections

The UCI Virtual Reading Room

UCI LIBRARIES
UNIVERSITY of CALIFORNIA • IRVINE

UCISPACE @ the LIBRARIES

UCIspace Home → Richard Rorty born digital files, 1988-2003 → Richard Rorty born digital administrative and prof... → Search

SHARE

Search This Subcollection

Go

Browse This Subcollection

Collections
Titles
[-] Creators
- Rorty, Richard (12)
[-] Subjects
- Rorty, Richard (12)
[-] Dates
- 1990 - 2000 (11)
- 1989 - 1989 (1)

About UCIspace

About This Project
Contact Us

Other Links

Special Collections & Archives
University Archives

Search

[-] **Search For**

Go

[-] **Search Filters**

Use filters to refine the search results.

Title | Add

Currently selected filters:

☑ Title: *

Update filters

[+] **Search Display Options**

Search Results for Subcollection: Richard Rorty born digital administrative and professional files, 1989-1999

Now showing items 1-10 of 12 — 1 2 Next

Thumbnail icon — "Freebies list for Achieving Our Country" list of scholars and their addresses
Creator: Rorty, Richard
Date Created: 1998-03-05

Thumbnail icon — "Freebies list for CIS" list of names of scholars and their spouses
Creator: Rorty, Richard
Date Created: 1989-02-08

Thumbnail icon — Application for residence in Bellagio, including detailed description of Rorty and His Critics project
Creator: Rorty, Richard
Date Created: 1996-09-01

IES
FORNIA • IRVINE

UCISPACE @ the LIBRARIES

al files, 1988-2003 → Richard Rorty born digital administrative and prof... → View Item

SHARE

"Freebies list for Achieving Our Country" list of scholars and their addresses

Date Created: 1998-03-05T07:11:44PST

Creator: Rorty, Richard

Date Available to Public: 2017-04-18 ⇐ **UCIspace account required for access to this file**

Permanent Link To This Item: http://hdl.handle.net/10575/28

Item Files

File Name: freebies.pdf
Size: 10.30Kb
Format: PDF

Additional Information

Title: "Freebies list for Achieving Our Country" list of scholars and their addresses

Subjects: Rorty, Richard

Type: Archives and Manuscripts

Language: en_US

Related Item: Is Part Of: Richard Rorty Papers. MS-C017. University of California, Irvine Libraries. Special Collections and Archives. Finding aid for entire collection available at: http://www.oac.cdlib.org/findaid/ark:/13030/kt9p3038mq/

Rights: This material is provided for private study, scholarship, or research. Transmission or reproduction of any material protected by copyright beyond that allowed by fair use requires the written permission of the copyright owners. The authors or their heirs retain their copyrights to the material. Contact the University of California, Irvine Libraries, Special Collections and Archives for more information (spcoll@uci.edu).

To access this item, interested researchers should submit an application online at http://special.lib.uci.edu/application-virtual-reading-room-ucispace. Access may be granted in less than 5 business days.

Provenance: Original file location on Richard Rorty's disk: MS-C017-FD044/plato97/FREEBIES

According to the documentation:

2011

Collection acquired. Consists of 400+ floppy disks and 1 hard drive

11 gigabytes total

2012

Disk images created and ingested to preservation repository

2014

Critical Theory Archivist processed physical components

Appraised born digital files, determined what to keep

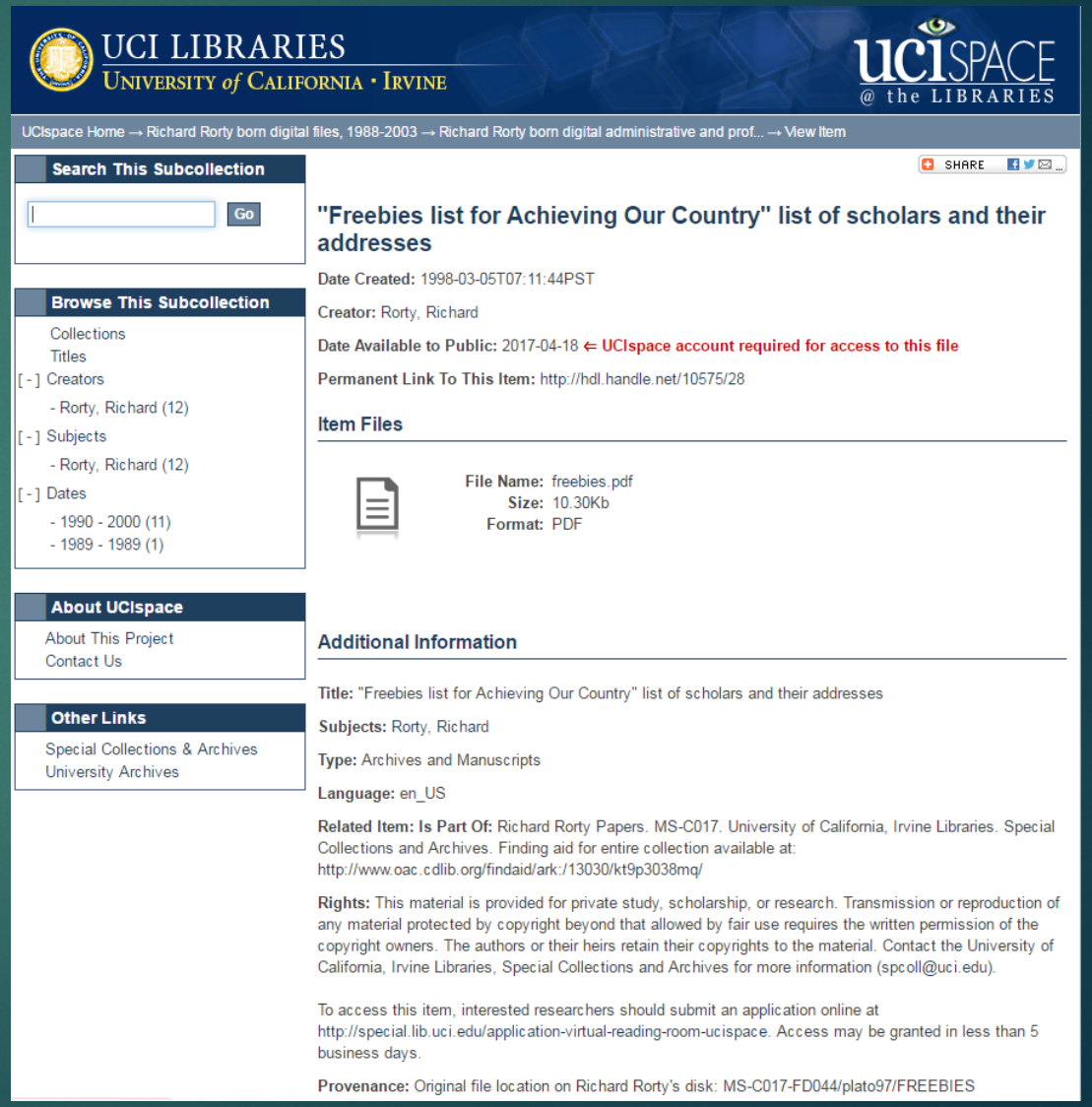
Created access copies of the "to keep" materials

60-80% of collection had access copies

2016: Me [redacted]

Richard Rorty Files in VRR

- 1027 files arranged in 8 subseries
- Item level processing
- Each file has metadata/description



UCI LIBRARIES
UNIVERSITY of CALIFORNIA · IRVINE

uciSPACE @ the LIBRARIES

UCIspace Home → Richard Rorty born digital files, 1988-2003 → Richard Rorty born digital administrative and prof... → View Item

Search This Subcollection

Go

Browse This Subcollection

Collections
Titles
[-] Creators
- Rorty, Richard (12)
[-] Subjects
- Rorty, Richard (12)
[-] Dates
- 1990 - 2000 (11)
- 1989 - 1989 (1)

About UCIspace

About This Project
Contact Us

Other Links

Special Collections & Archives
University Archives

SHARE

"Freebies list for Achieving Our Country" list of scholars and their addresses

Date Created: 1998-03-05T07:11:44PST

Creator: Rorty, Richard

Date Available to Public: 2017-04-18 ⇐ UCIspace account required for access to this file

Permanent Link To This Item: http://hdl.handle.net/10575/28

Item Files

File Name: freebies.pdf
Size: 10.30Kb
Format: PDF

Additional Information

Title: "Freebies list for Achieving Our Country" list of scholars and their addresses

Subjects: Rorty, Richard

Type: Archives and Manuscripts

Language: en_US

Related Item: Is Part Of: Richard Rorty Papers. MS-C017. University of California, Irvine Libraries. Special Collections and Archives. Finding aid for entire collection available at: http://www.oac.cdlib.org/findaid/ark:/13030/kt9p3038mq/

Rights: This material is provided for private study, scholarship, or research. Transmission or reproduction of any material protected by copyright beyond that allowed by fair use requires the written permission of the copyright owners. The authors or their heirs retain their copyrights to the material. Contact the University of California, Irvine Libraries, Special Collections and Archives for more information (spcoll@uci.edu).

To access this item, interested researchers should submit an application online at http://special.lib.uci.edu/application-virtual-reading-room-ucispace. Access may be granted in less than 5 business days.

Provenance: Original file location on Richard Rorty's disk: MS-C017-FD044/plato97/FREEBIES

Mark Poster Files in VRR

- 1 GB of material
- Divided into subseries
- Within subseries, contains zip files with the files
- Description includes CSV files containing the file names within zip files

SHARE

Mark Poster "Notes" files, 1992-2004

Date Created: 1992

Creator: Poster, Mark

Description: This subdirectory includes books and journal articles by other authors, selected texts and citations of research interest to Mark Poster, as well as a curriculum vitae and email correspondence. This sub-directory has been appraised and packaged for access as a .zip file by the UCI Libraries. Researchers may search the contents of the .zip file after downloading and unzipping it. The .zip file is accompanied by a .csv file that lists the contents of the .zip file. Only the .csv file is searchable within UCISpace. This .csv file may be opened using a spreadsheet program such as Microsoft Excel.

Date Available to Public: 2017-04-18 ⇐ **UCIspace account required for access to this file**

Permanent Link To This Item: http://hdl.handle.net/10575/5647

Item Files

File Name: notes.zip
Size: 32.27Mb
Format: Unknown

File Name: posterinventory_notes.csv
Size: 112.3Kb
Format: CSV file

Additional Information

Title: Mark Poster "Notes" files, 1992-2004

Subjects: Internet -- Social aspects | Information technology -- Social aspects | Mass media -- Social aspects | Postmodernism -- Social aspects | Critical theory

Type: Archives and Manuscripts

Duration: 409 files, 35.3 MB

Related Item: Is Part Of: Mark Poster Papers. MS-C018. University of California, Irvine Libraries. Special Collections and Archives. Finding aid for entire collection available at: http://www.oac.cdlib.org/findaid/ark:/13030/kt809ng0c3

An Idea for a Plan

ALL the files

List of all file names and corresponding subfolders

Into the Virtual Reading Room

attached to finding aid

Files organized in subfolders, by digital object number (i.e. original disk media)

MSC013_DIG008_Working	3/13/2014 6:50 PM	File folder
MSC013_DIG023_Original	3/17/2017 7:13 AM	File folder
MSC013_DIG023_Working	3/14/2014 10:55 AM	File folder
MSC013_DIG025_Original	3/17/2017 7:13 AM	File folder
MSC013_DIG025_Working	7/9/2014 5:41 PM	File folder
MSC013_DIG026_Original	3/17/2017 7:13 AM	File folder
MSC013_DIG026_Working	3/14/2014 12:43 PM	File folder
MSC013_DIG029_Original	3/17/2017 7:13 AM	File folder
MSC013_DIG029_Working	3/17/2017 7:13 AM	File folder
MSC013_DIG030_Original	3/17/2017 7:13 AM	File folder
MSC013_DIG030_Working	3/17/2017 7:13 AM	File folder
MSC013_DIG031_Original	3/17/2017 7:13 AM	File folder
MSC013_DIG031_Working	3/17/2017 7:13 AM	File folder

Processing Plan Template

Estimated processing hours (refer to chart in Section 4.5 of the processing manual). Comment on condition (i.e. barriers to access).

! Not applicable to born digital !

How will you organize the collection? Is there any existing meaningful order? What series will you use?

Born digital files are organized by digital object number, which corresponds to the piece of digital media that they came in on. This organization will be maintained. There will only be one series for born digital materials. If the collection receives a lot of use, it may be beneficial to arrange the files differently. Time, and patrons, will tell.

Processing Level	Component Level	Title	Description	Arrangement	Preservation	Appraisal
Low	Series	Born Digital materials				
			Consolidate only the “preservation” folders			
			Change folder names (remove the word “preservation”)			
			Determine which folders have nothing in them (create a print of the access copies folder, examine that)			
			Delete empty folders			
			Create inventory of file names corresponding to folder name/digital object number			
			Decide if you want to zip one large folder, or one for floppies, one for each hard drive—thinking one for the			

Uh oh…

- Only 500 MB converted to access copies.
  - Realized after segregating the folders that held the access copies
- Contained correspondence, which has a donor-imposed 25 year restriction

Current Finding Aid for J. Hillis Miller Born Digital Materials

C013-018-A,
Box 45, Item MS-
C013-017-A

Born digital files Series 9. **circa 1980s-2000s**

Physical Description: 11.8 Gigabytes

Scope and Content Summary

The **J . Hillis Miller papers** have an unprocessed digital component not yet available to researchers. The electronic content primarily consists of drafts, correspondence, photographs, and notes. Many documents are duplicated in the analog collection accessible in the Special Collections and Archives Reading Room.

Lessons Learned

- Minimally processing born digital materials ≠ minimal effort
- What does processing even mean in a born digital landscape?
- Documentation may not be complete, needs to be clearer
- Help is needed!

Born Digital Processing Framework

- 9 archivists
- Came from the Born Digital Archiving eXchange
  - Unconference at Stanford

Survey the collection
Create processing plan
Rehouse physical media if necessary
Decision - do you keep physical media or not?
Assign identifier to physical media
Photograph/document physical media
Consult collection materials (ie deed of gift, digital material survey, etc)
PII risk assessment
Create file directory list (file-level metadata)
Perform file format analysis
Identify deleted/temporary/system files
Image media (but is this more of an acquisition task?)
Scan for PII

Remove or otherwise segregate PII that is found
Identify and describe restrictions based on PII found
Identify duplicate content
Delete (or otherwise identify) duplicate content
Determine volume of materials (in M/G/T/P bytes)
Virus scan
Describe content at appropriate level
Add description to a finding aid (what kind of description)?
Determine arrangement
Determine level of description
Arrange materials intellectually

Understand correlation between any analog/physical material
Arrange files according to intellectual decisions
Extract descriptive metadata
Weed/separate material that doesn't fit collecting scope
Extract technical metadata
Record technical metadata
Record administrative metadata
Make preservation decisions - how will files be made available?
Determine which files need to be migrated
Migrate materials in need of migration
Create a directory list

For each activity (e.g. create file directory list) decide the following:

1. Where in the lifecycle it falls, e.g. description, preservation, wrap up work
2. If it should be included in min. processing requirements
3. If source of the content affects the activity
4. If format of content affects the activity
5. How important the task is to the workflow

Thank you!

and feel free to contact me:
lugleanj@uci.edu

UCSF Digital Collections

Migrating to Nuxeo/Calisphere

Kelsi Evans, Project Archivist
UCSF Archives
kelsi.evans@ucsf.edu

UCSF Digital Collections

Welcome, kelsi Log Out

- Dashboard
- Items
- Collections
- Item Types
- Tags
- Exhibits
- Dropbox
- Simple Pages
- Simple Vocab

Edit Item #3234: "University of Califor...

Search +

Dublin Core | Item Type Metadata | Files | Tags

Dublin Core

The Dublin Core metadata element set is common to all Omeka records, including items, files, and collections. For more information see, http://dublincore.org/documents/dces/.

Title — A name given to the resource

Add Input

University of California College of Dentistry "Earthquake Class" of 1906

Use HTML ☐

Subject — The topic of the resource

Add Input

Use HTML ☐

Description — An account of the resource

Add Input

UC Affiliated Colleges, College of Dentistry / Dental Department class of 1906. The class graduated after the San Francisco earthquake of 1906.

Use HTML ☐

Creator — An entity primarily responsible for making the resource

Add Input

Save Changes

View Public Page

Delete

Public: ☑ Featured: ☑

Collection

Select Below

UCSF Japanese Woodblock Print Collection

Title:
Meiji no kusuri no kōkoku
Translated Title:
drug advertisement given to Meiji-za theatre
Creator/Contributor:
Toyohara, Kunichika, 1835-1900, Artist
Abstract:
a drug advertisement
Date:
1897
Subject:
Advertising Japan History
Drugs Japan
Japan
Toyohara, Kunichika, 1835-1900
Type:
triptych
woodblock print
Ukiyo-e
advertisement
Physical Description:
35.8 x 72.7 cm
Language:
Japanese
Identifier:
ucsf_p084
Origin:
Japan
Collection:
UCSF Japanese Woodblock Print Collection
Contributing Institution:
UC San Francisco, Special Collections

printable version
image only | image with details

nuxeo

HOME | WORKSPACE | SEARCH | Kelsi.Evans@ucsf.edu

Document, User, Group

IMPORT

Asset Library (89)
- Admin (0)
- UCB (0)
- UCD (0)
- UCI (0)
- UCLA (0)
- UCM (56)
- UCOP (0)
- UCR (0)
- UCSB (0)
- UCSC (0)
- UCSD (0)
- UCSF (0)
  - "A History of UCSF" website images (0)
  - AR 2014-32 Development/Alumni Relations (0)
  - AR 2015-4 School of Dentistry (0)
  - AR 2017-16 Base Hospital 30 (0)
  - AR 90-60 UCSF 125th Anniversary (0)
  - Archives Classification (0)
  - Berne, Eric Collection (0)
  - Charlie (0)
  - Glantz (0)
  - Health Sciences Artifact Collection (0)
  - Homeopathy Collection (0)
  - Japanese Woodblock Prints (0)
  - MSS 2000-31 AIDS Ephemera Collection (0)
  - MSS 2001-04 Sally Hughes AIDS Research (0)
  - MSS 2001-05 GASP of Colorado (0)
  - MSS 2001-33 Shimp (0)
  - MSS 2002-08 Radiologic Imaging Laboratory (0)
  - MSS 2003-12 Eric Berne papers (0)
  - MSS 2005-08 Eric Berne papers (0)
  - MSS 2007-21 J. Michael Bishop (0)
  - MSS 2007-33 Hahnemann Hospital (0)
  - MSS 2009-01 Eddie Leong Way (0)
  - MSS 2011-23 Robert L. Day (0)

> Asset Library > UCSF

UCSF

Used: 1,104.36 GBytes

Content | Summary | Edit | Files | History | Manage | Quota

New | Filter | ...

Items/page 20

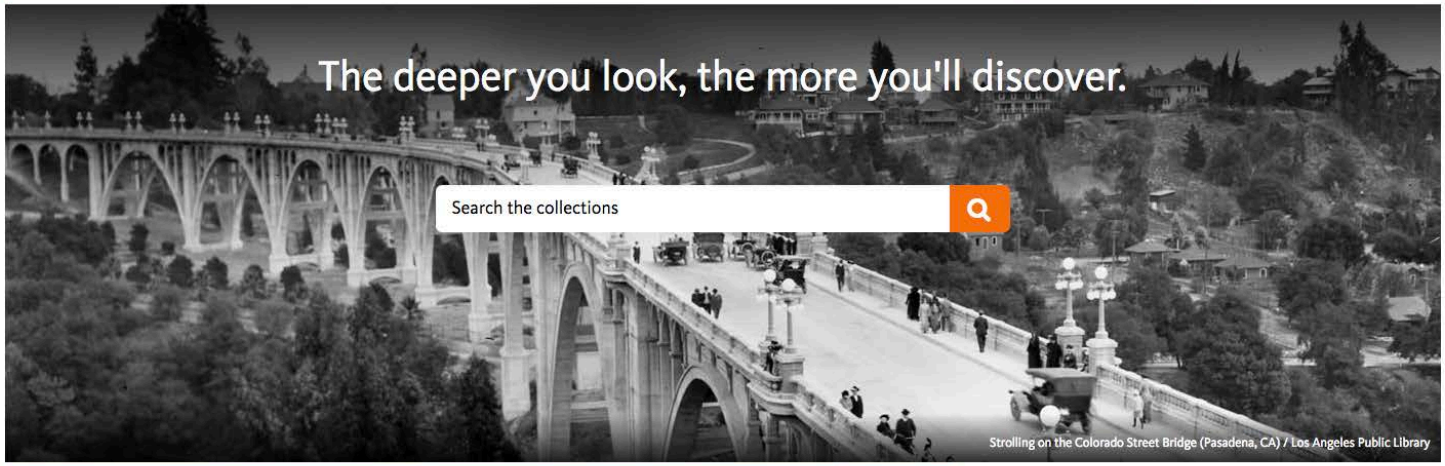
1/3

Title	Modified	Last contributor	Version	State
"A History of UCSF" website images	12/15/2016	Kelsi.Evans@ucsf.edu		Project
AR 2014-32 Development/Alumni Relations	1/5/2017	Kelsi.Evans@ucsf.edu		Project
AR 2015-4 School of Dentistry	3/17/2016	Barbara.Hui@ucop.edu		Project
AR 2017-16 Base Hospital 30	4/19/2017	Kelsi.Evans@ucsf.edu		Project
AR 90-60 UCSF 125th Anniversary	1/26/2016	Kelsi.Evans@ucsf.edu		Project

About | Contact | Help | Terms of Use

CALISPHERE
UNIVERSITY OF CALIFORNIA

Contributing Institutions | Collections | Exhibitions



Calisphere is a gateway to digital collections from California's great libraries, archives, and museums. Discover over 750,000 images, texts, and recordings—and counting.

UCSF DIGITAL COLLEC

Search +

Browse Items | Browse

FEATURED

Dentistry recruitm poster with group

UCSF School of Dentistry Admissions recruitment p offering professional care dentistry and dental hygie Photo by Bob Vogel.

FEATURED

Mary B. Olney pape

Selected documents and imag 1938, Dr. Olney founded the f

FEATURED

Cholera: The Reinh

The Reinhard S. Speck Chole was a Professor of Microbiolo

Browse Items · Browse Coll

Proudly powered by Omeka.

UCSF DIGITAL COLLECTIONS

Search +

Browse Items | Browse Collec

BROWSE IT

Browse All | Browse by Tag | S

< 2 of 118 >

Bertram Katzung planting flowers o Chauncey Leake Day

Francis A. Sooy w UCSF marching band on Chaunce Leake Day

Browse Items | Browse Collections | Browse Exhibits

BERTRAM KATZUNG PLANTING FLOWERS ON CHAUNCEY LEAKE DAY

Dublin Core

Title
Bertram Katzung planting flowers on Chauncey Leake Day

Description
Bertram Katzung planting flowers on UCSF Chauncey Leake Day, June 2, 1981, in front of the Medical Sciences Building

Source
Photograph collection, Chauncey Leake Day

Publisher
Regents of the University of California

Date
1981-06-02

Contributor
UCSF Archives and Special Collections

Rights
Regents of the University of California

Format
Photographic print

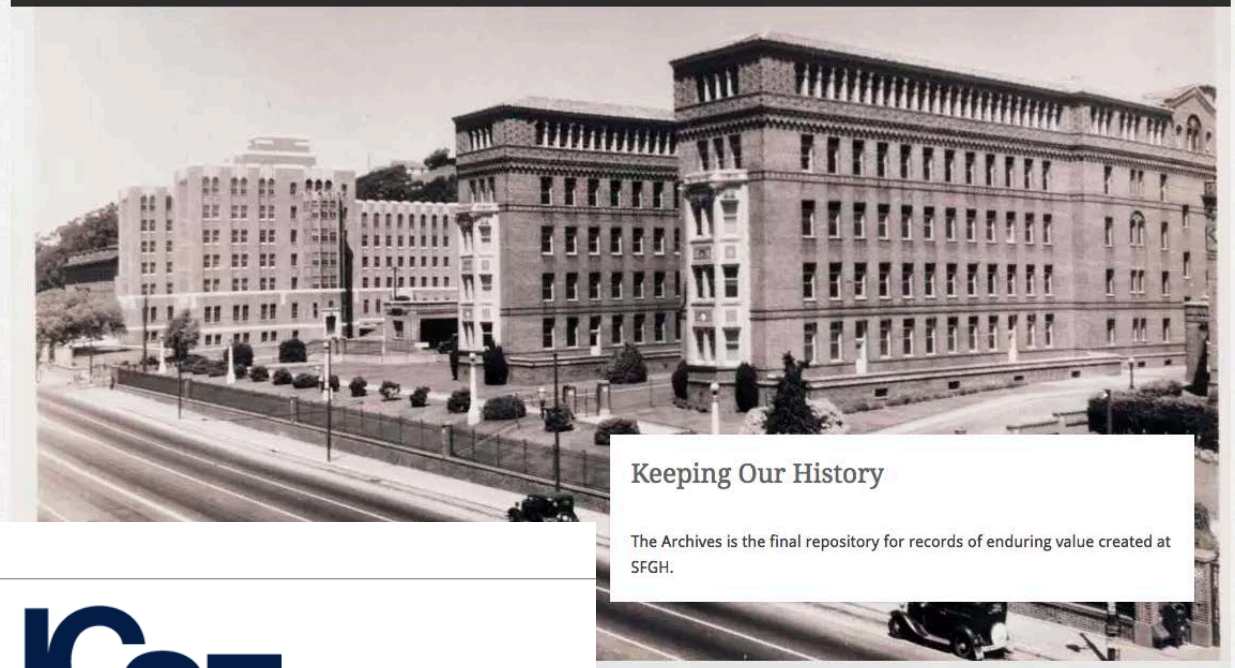
Type
Image

Identifier
photocoll_chaunceyleakeday1981_katzung

Files

San Francisco General Hospital Historical Archives

HOME | ZSFG HISTORY | ABOUT THE ARCHIVES | COLLECTION DEVELOPMENT POLICY | GET INVOLVED | VIEW OUR COLLECTIONS



Keeping Our History

The Archives is the final repository for records of enduring value created at SFGH.

UC San Francisco

View featured image

UCSF

Location: San Francisco, CA
Phone: (415) 476-8112
Email: http://www.library.ucsf.edu/collections/archives/contact
Website: http://www.library.ucsf.edu/

- Collections at UC San Francisco
- Contributors at UC San Francisco
- Search UC San Francisco

Contributors at UC San Francisco

H.M. Fishbon Memorial Library
Library, Industry Documents Library
Library, Special Collections
Library, Tobacco Control Archives
Library, University Archives

Making the Move to Nuxeo

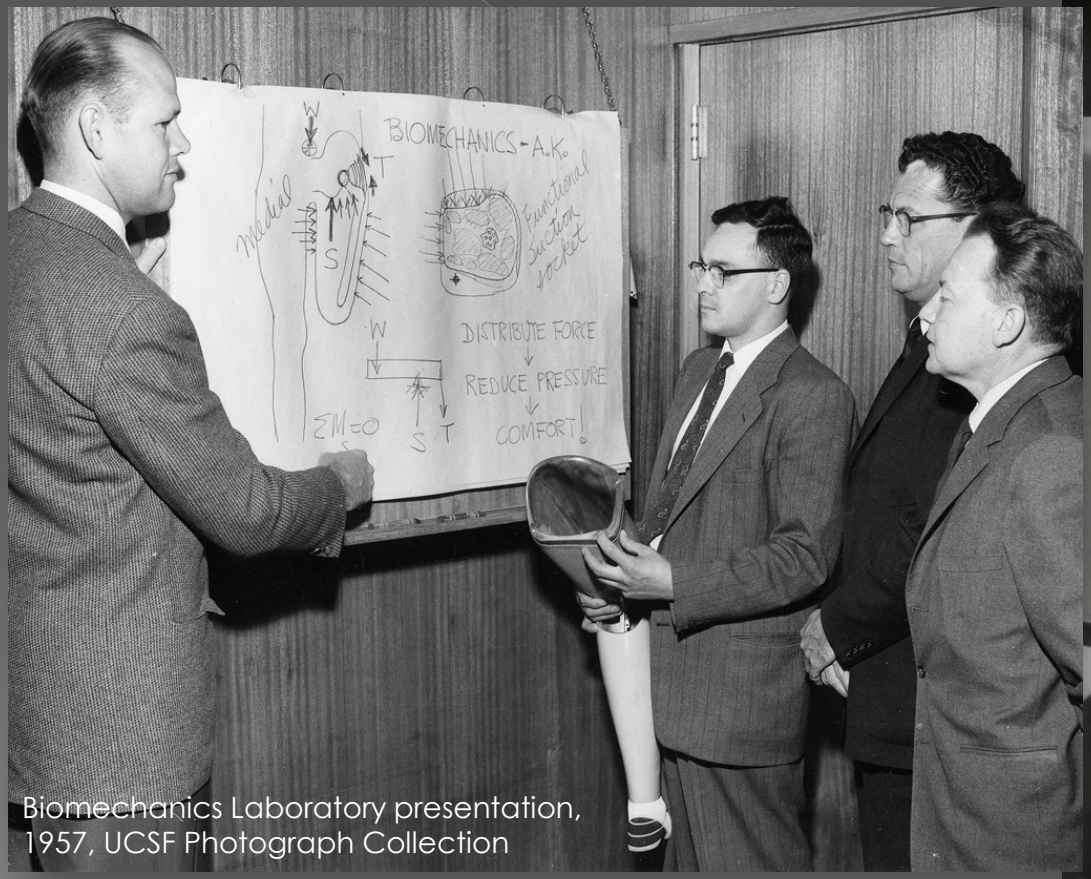
- Unifies collections
- Increases searchability
- Better file management
- CDL support

UCSF School of Medicine picnic, 1959, UCSF Photograph Collection

Aerobics class at Millberry Union, 1982, UCSF Photograph Collection

Implementing the System

- File migration



Biomechanics Laboratory presentation, 1957, UCSF Photograph Collection

UCSF Library computer, 1984, UCSF Photograph Collection

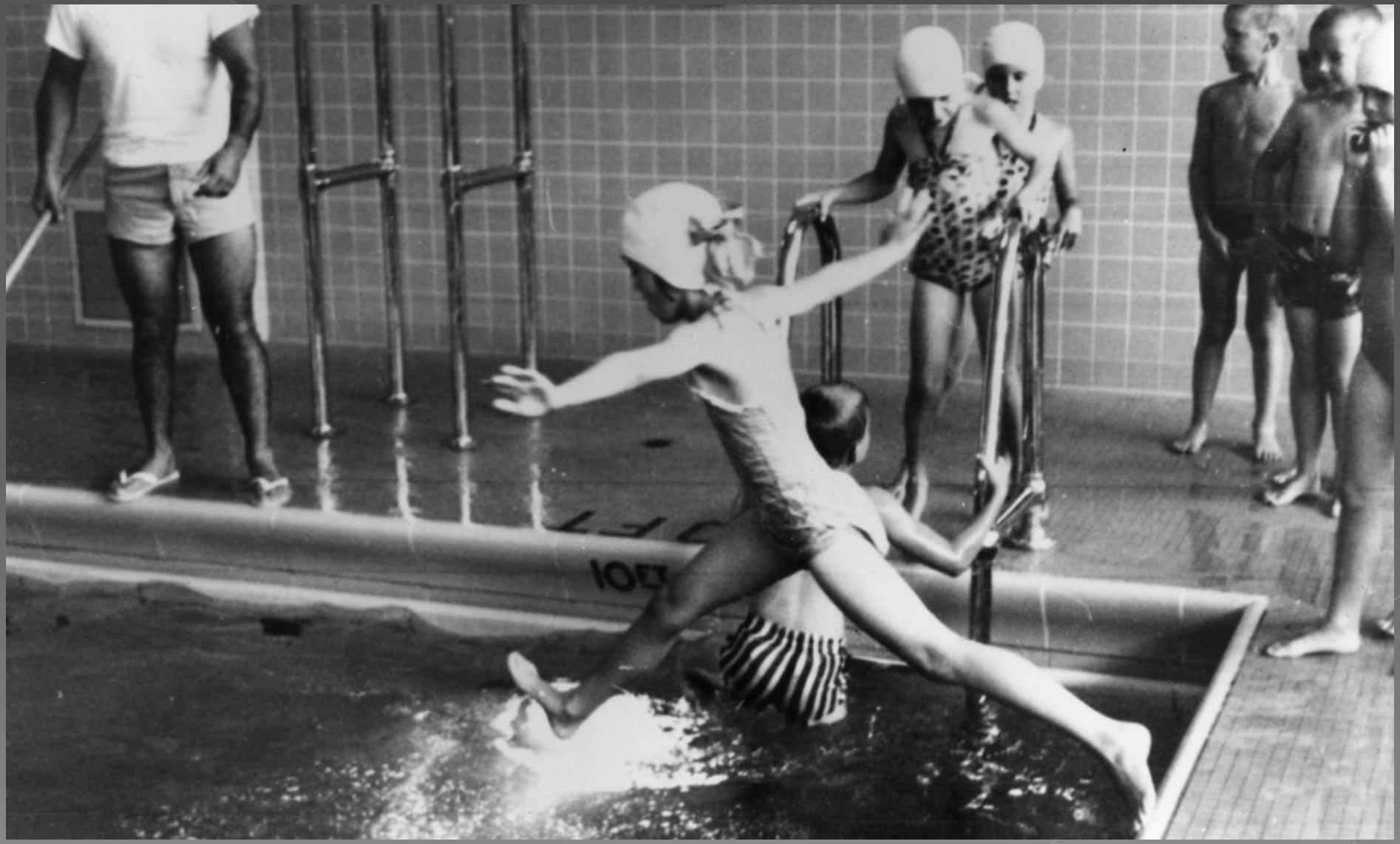
Implementing the System

- Metadata cleanup

Library staff member Phyllis Gross in UCSF Library Current Journals area, 1958, UCSF Photograph Collection

Library staff member Charles Stuckey at card catalog, 1969, UCSF Photograph Collection

Thank you!



Kelsi Evans – SCA AGM talk - 2017
Rough notes for talk

At UCSF we **migrated our digital collections** from a few different locally-managed sites, including an Omeka instance, to Nuxeo/Calisphere.

In this talk, I'll go over the **considerations that went into the decision** to migrate our material and discuss some of the major issues we had to work through in the initial phases of the migration. My **colleague David Krah will go into more depth** into some of our current projects and how our tools and processes have continued to evolve.

UCSF Archives set up an **Omeka instance in 2008** to serve digitized material to the public. If you are unfamiliar with Omeka, it is "free, flexible, and open source web-publishing platform for the display of library, museum, archives, and scholarly collections and exhibitions." The **focus of the platform is really display and web-publishing**, less on digital asset management, and we'll come back to that in a minute.

[show backend, talk about Dublin Core fields, our instance was set up to handle small jpgs and pdfs]

In addition to the Omeka instance, we also had some sites that I'm going to call **project sites**. These were basically created to serve one collection and were connected with a funded project. One of these was the **Japanese Woodblock Print Collection.**

[show site, includes over 400 images with descriptive metadata of our medical-related Japanese prints. It has a lot of information but lives as this kind of siloed site on the internet, so it can be difficult to direct researchers to and it lives out of context of the larger UCSF archives collections]

Around 2014 we began evaluating the new system being offered by CDL, Nuxeo backend with public display on the new Calisphere.

We were frankly eager to try a new system because the **limitations of our Omeka instance** and other sites were becoming clearer and clearer, especially as we started undertaking more large-scale digitization projects and our grants more regularly included some sort of digitization component.

To highlight a few of the limitations – **Omeka is open source**, which means there is a robust user community which can be great…but to be part of that, you really need staff with some programming and development expertise or at least the time and energy to devote to developing those skills. We just didn't really have this, so our

instance was just the bare bones, which has limited search functionality and not the best user interface.

[show page with limited images, and no space for complex objects, or really anything more complicated than an image or pdf]

Additionally, if you remember I mentioned that it is a platform **focuses on display and web publishing, not asset management, and that's really how it had been used by our institution.** And because of this our backend had become a real hodge-podge of collections and exhibits and stand alone objects that were thrown up to make stuff available but not really with a mindful intention of managing robust, large digital collections with complex objects.

Finally, a major limitation was **Omeka's inability to serve as a unifying platform** for collections and I mean this at a couple different levels. **One, it couldn't easily and clearly bring together the different contributing institutions** that live under the umbrella of UCSF, including San Francisco General Hospital (which maintains its own website), and the Mount Zion campus (which had digital objects on oac). **Two, because of technical limitations**, we couldn't easily migrate the material that lived on project sites into Omeka, so what we were left with were several stand alone sites along with the Omeka site that we were trying to get users to navigate through, and that was becoming really confusing.

Nuxeo and Calisphere offered solutions to a lot of these issues, including the ability to unite collections under one **UCSF umbrella** and take that material that lived on siloed sites and put it in conversation with other collections in a much more **search friendly** interface, **manage complex objects and different file types** (especially the high res preservation copies) and **CDL was going to be there to offer support**, so we could actually push the boundaries of the system in a way our staff limitations had not allowed us to do with Omeka.

So with all this in mind in 2015 we decided to migrate as much of our material as was possible into Nuxeo with public display on Calisphere

One of the first steps was migrating material from the stand alone sites, which was relatively straightforward. We had the tifs that we wanted to manage in Nuxeo which would be automatically served as low res jpgs on Calisphere, so we sent those to CDL on harddrives, which CDL then loaded into Nuxeo under the appropriate project folder, and CDL and our team did some metadata field matching and then did a mass migration of that data from one site to the other.

We did **some of that same process with the Omeka material but then we started to run into some issues**. The first was that when our team and CDL bulk pulled files from Omeka, **all they got were low res access copies** being generated by Omeka. These were not the high res tifs that we want to manage in the long run. So we had to go back to our campus server and track down these tifs using the file

name and then go through the process of matching these tifs with the affiliated metadata on Omeka, using the file name to match items. This was eventually **effective but not really efficient a**nd we definitely learned some lessons and we're trying to implement those now with an eye toward future migrations.

Another large issue for us was the fact that a lot of the material on **Omeka had very limited metadata, s**ometimes just a title and an unhelpful file name, so things like "building" with a file name of "building". I imagine this was done to serve material to researchers, so again that focus on publishing and less on asset management in Omeka. So years later, we didn't know really basic information like which collection the image came from.
So for this, **we created an intern project** of tracking down some of that material and having them update metadata. So this was effective but time and staff intensive. To help manage that time and staff investment, we used this **project to reevualate some of the material and decide if it was worth migrating for public display**. We knew we had the tifs for preservation on our server but we made decisions about some of those "buildings" images and they just didn't make the cut for migration.
This whole process really helped us establish better metadata standards, file naming conventions, and digitization best practices guides, again with an eye toward future migrations; really having an understanding that this is an iterative process that we need to be prepared for for long term stewardship of digital items.

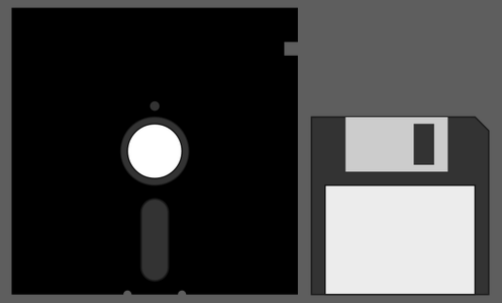
At this point we've successfully migrated all of our omeka material and the majority of the project sites onto Nuxeo and started building new collections.
From a little over 2000 items on omeka and a few hundred on project sites to almost 30,000 on Calisphere with definite plans for growth in the future.

Crossing the Knowledge Gap: Effective Documentation's Role in Creating Digital Preservation Workflows

Victoria (Tori) Maches
MLIS student, UCLA

The problem

- Need tech documentation to start program, document processes to maintain it
- Tech documentation assumes background archivists may lack
- Gaps in documentation affect developing programs
- Clear documentation needed to get started and maintain program

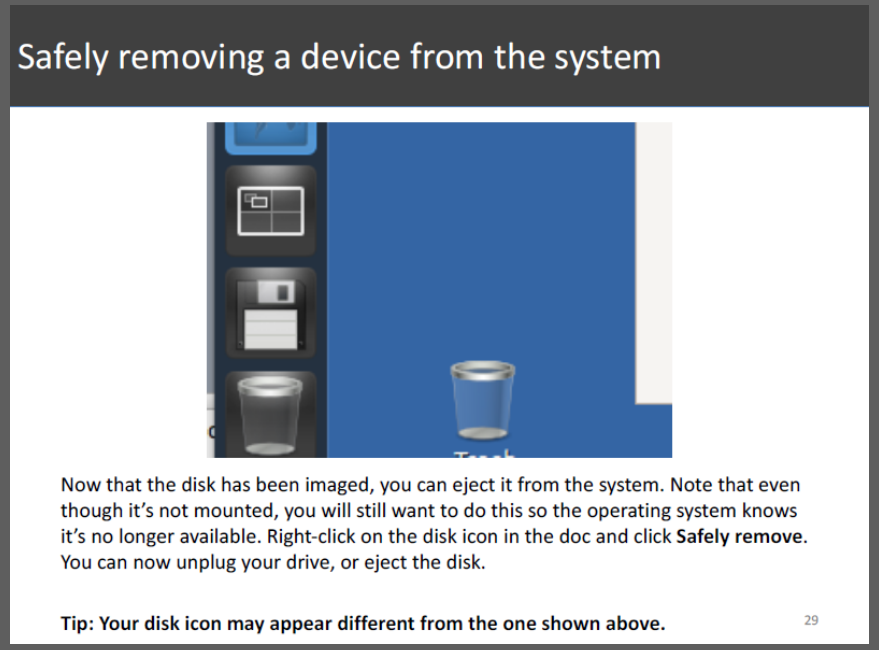


Steps for new practitioners

- The focus: address knowledge gaps, develop skills
  - Ask questions
  - Tutorials and alternate documentation
  - Look outside archives-specific contexts
  - Document everything
  - Pay attention to what you don't know

Steps for documentation creators

- The focus: What would you have wanted to know?
  - Step-by-step instructions
  - Explain how/why it works
  - Screenshots/photos
  - Take advantage of born-digital medium
  - Assume inexperienced audience
- Keep future practitioners in mind



Conclusion

- Need clear documentation to create workflows, maintain program
- Start now and future documentation will fill these gaps
- Combine short- and long-term approaches

LOCKSS LOTS OF COPIES KEEP STUFF SAFE

Born Digital: Care, Feeding, & Intake Processes at LOCKSS

Mary-Ellen Petrich - @mellen22
Digital Preservation Specialist, LOCKSS
Stanford University Libraries

Society of California Archivists
April 2017

me

- engineering -> **library science**
- hired to **catalog** the preservation collection and **test software**
- developed **processes**, and **scripts** that direct the preservation process at **LOCKSS**

LOCKSS?

- lots of **copies** and **communities** keep stuff safe
- a LOCKSS network is a **peer-to-peer** network
- websites are **not predictable**
- LOCKSS addresses issues of **data relationships** and **metadata**

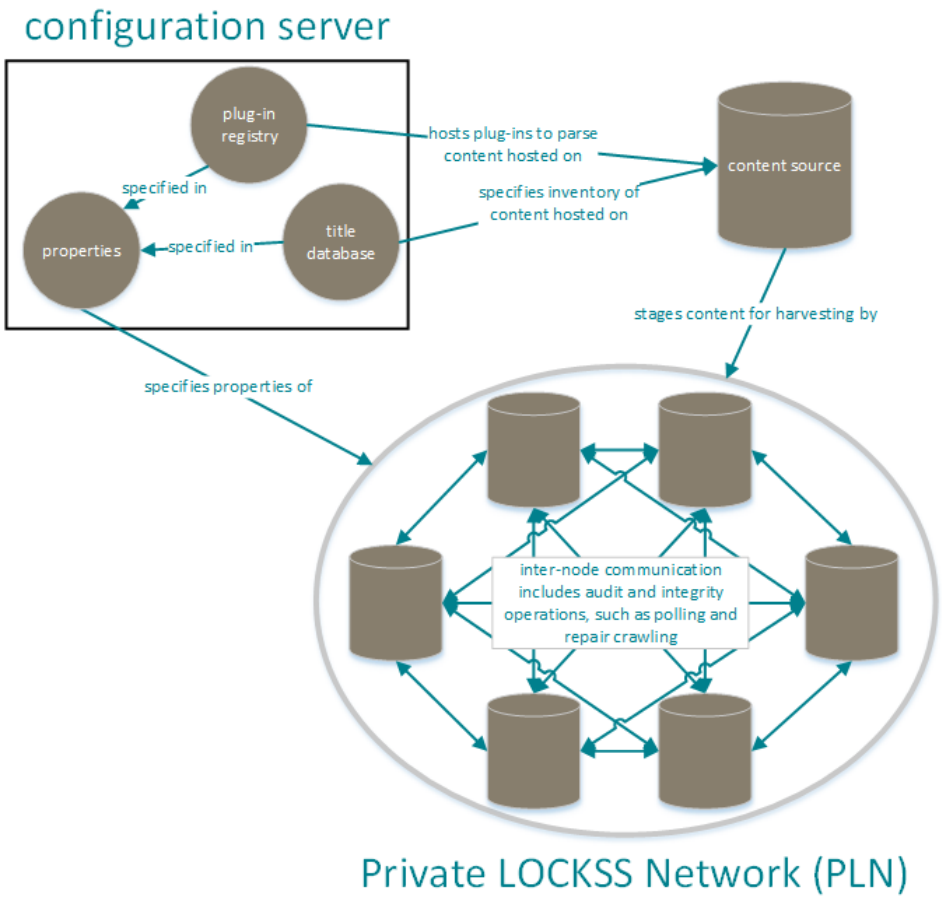
inception

- Founded in 1999
- By a serials librarian and a computer scientist

- print journals → Web
- **conserve library's role** as preserver
  - **collect** from publishers' websites
  - **preserve** w/ cheap, distributed, library-managed hardware
  - **disseminate** when unavailable from publisher

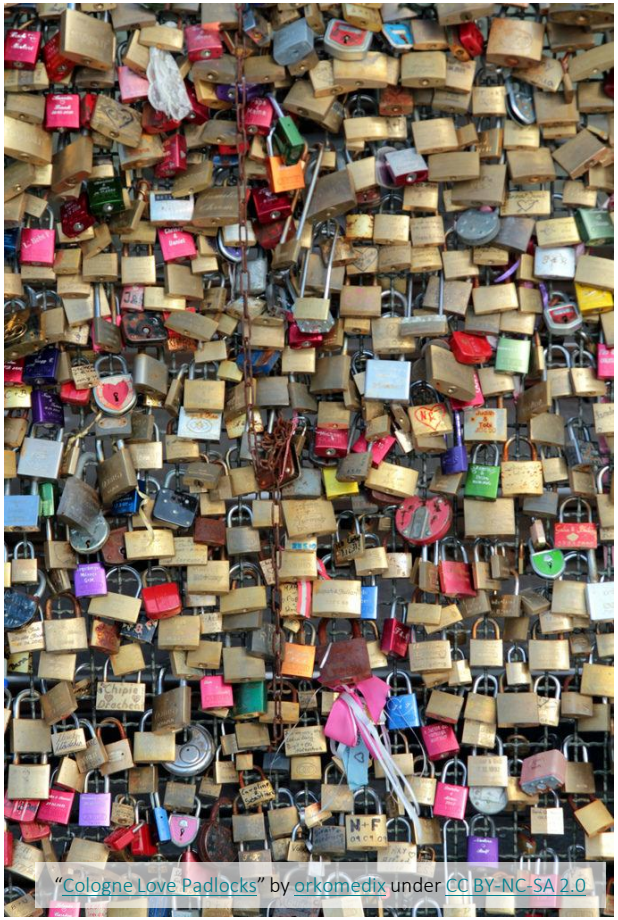
what is a LOCKSS network?

- Peer-to-peer network of web servers
- Journals and other archival information on the Web
- A set of independent, low-cost, persistent Web caches that cooperate to detect and repair damage to their content by voting in "opinion polls."



lots of LOCKSS

- LOCKSS (principle)
- LOCKSS (program)
- LOCKSS (software)
- Global LOCKSS Network (GLN)
- Private LOCKSS Networks (PLNs)
- CLOCKSS



Private LOCKSS Networks (PLNs)

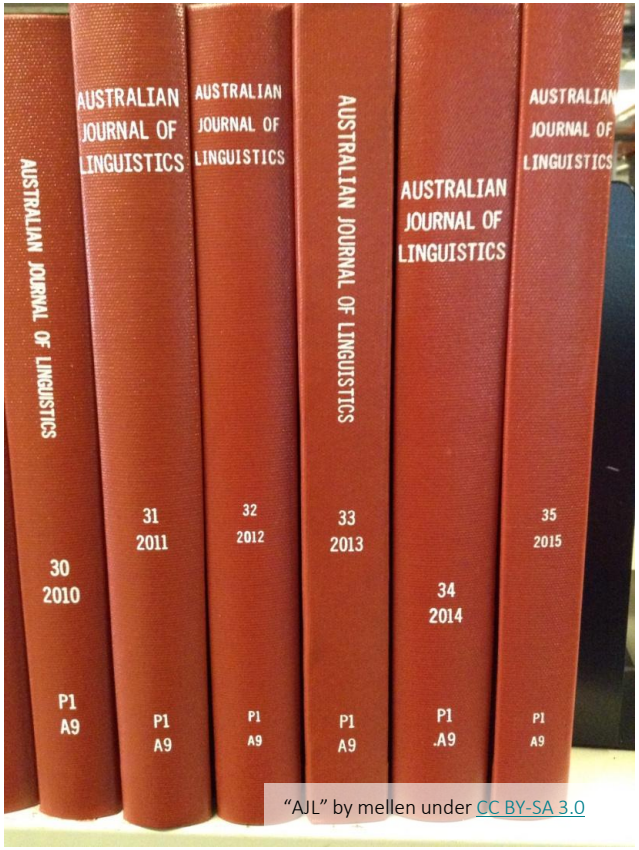
- what are they?
  - community of interest
  - jointly designate content
  - run distributed nodes
  - establish governance
  - preservation via diverse technologies, institutions, networks

Controlled LOCKSS (CLOCKSS)

- what is it?
  - library/publisher partnership
  - preserve the scholarly record
  - 12 globally-distributed nodes
  - **dark** until no longer accessible
  - triggered content world-accessible

Global LOCKSS Network (GLN)

- ~150 **Libraries**, >600 **Publishers**
- released:
  - **~9,000 journals**
  - **~110,000 Archival Units (AU)**
  - **~15-20 terabytes**
- dark web / subscription materials
- what is it?
  - **conserve library's role** as preserver
  - **collect** from publishers' websites
  - **preserve** w/ cheap, distributed, library-managed hardware
  - **disseminate** when unavailable from publisher

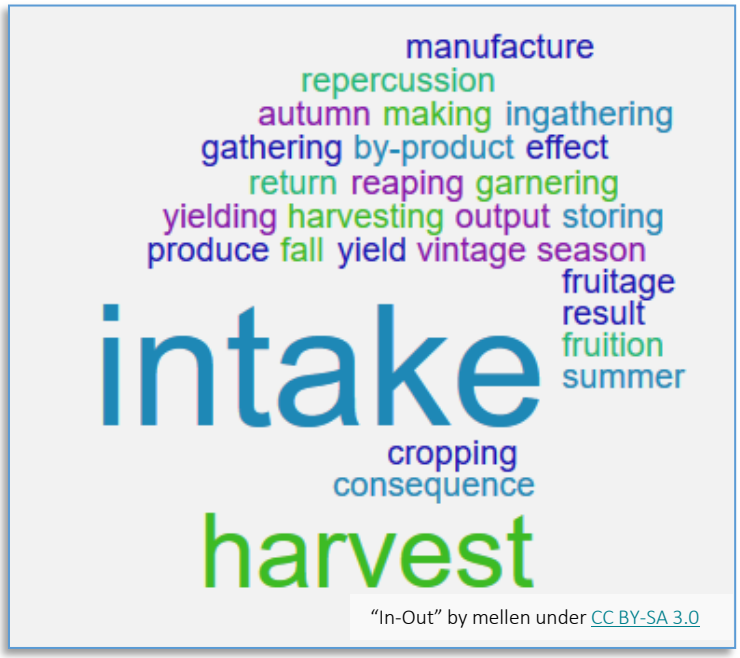


collection methods

- WARC
  - Hand-crafted
  - Quick & Dirty
  - Small single journals
- File Transfer
  - FTP or snail mail
  - Publisher Driven
- Harvest
  - Acting like a browser
  - LOCKSS Driven
  - Preserves file relationships
  - Parses out metadata

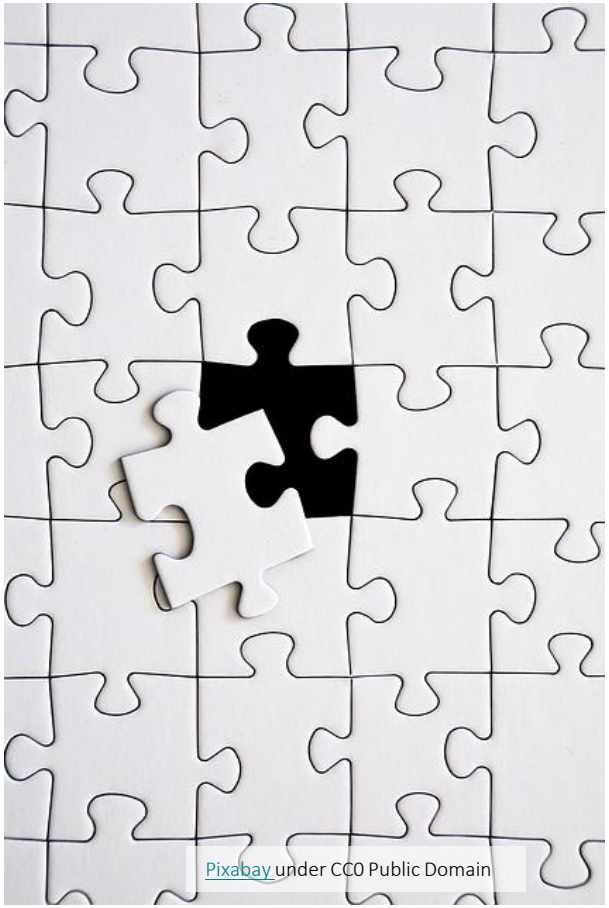
publisher setup for harvest

- Archival Unit (AU)
  - Volume of a Journal
  - Volume or Chapter of a Book
  - A closed collection of documents
  - Up to ~500 GB
- Subscription
  - IP Address access
- LOCKSS Permission Statement
  - Site, Journal, or Volume level
  - *LOCKSS system has permission to collect, preserve, and serve this Archival Unit*
- Manifest page
  - List of journal issues
  - Bottom of the tree



publisher plugin to the LOCKSS daemon

- Collection
  - Start URL
  - Link extraction
  - Crawl Rules - Exclude & Include
  - Crawl filters
- Validation
  - Mime type
  - Html error codes
  - Login page identification
  - Substance checking
- Metadata Collection
- Polling filters



title database (tdb file)

- catalog records ++
- basic metadata
  - publisher, title, publication year, issn/isbn
  - in case metadata is missing
- parameters for each AU
  - url & volume or year or others
  - defines the AUid
    - passes parameter values to the publisher plugin
    - unique key
- status
  - human readable preservation stage
  - LOCKSS daemon: recognize, crawl, don't crawl

digital workflow

- ***doNotProcess*** ignore this AU
- ***doesNotExist*** AU does not exist
- ***expected*** not known if AU exists on the publisher's web site
- ***exists*** known that AU exists on the publisher's web site
- ***manifest*** permission page and manifest verified
- ***wanted*** higher priority for testing
- ***testing*** someone is testing this AU
- ***notReady*** testing has failed
- ***ready*** testing is completed and the AU is ready for release

<-->

- ***released*** released for collection
- ***down*** no longer collected, unavailable through the publisher
- ***superseded*** this volume is no longer collected, but is available with another platform

```
1 {
2
3   publisher <
4     name = Taylor & Francis ;
5     info[tester] = 6
6   >
7
8   plugin = org.lockss.plugin.taylorandfrancis.TaylorAndFrancisPlugin
9   param[base_url] = http://www.tandfonline.com/
10
11  {
12
13    title <
14      name = Archives and Manuscripts ;
15      issn = 0157-6895 ;
16      eissn = 2164-6058
17    >
18
19    param[journal_id] = raam20
20
21    implicit < status ; year ; name ; param[volume_name] >
22
23    au < manifest ; 2012 ; Archives and Manuscripts Volume 40 ; 40 >
24    au < down ; 2013 ; Archives and Manuscripts Volume 41 ; 41 >
25    au < down ; 2014 ; Archives and Manuscripts Volume 42 ; 42 >
26    au < down ; 2015 ; Archives and Manuscripts Volume 43 ; 43 >
27    au < manifest ; 2016 ; Archives and Manuscripts Volume 44 ; 44 >
28
29  }
30
31 }
```

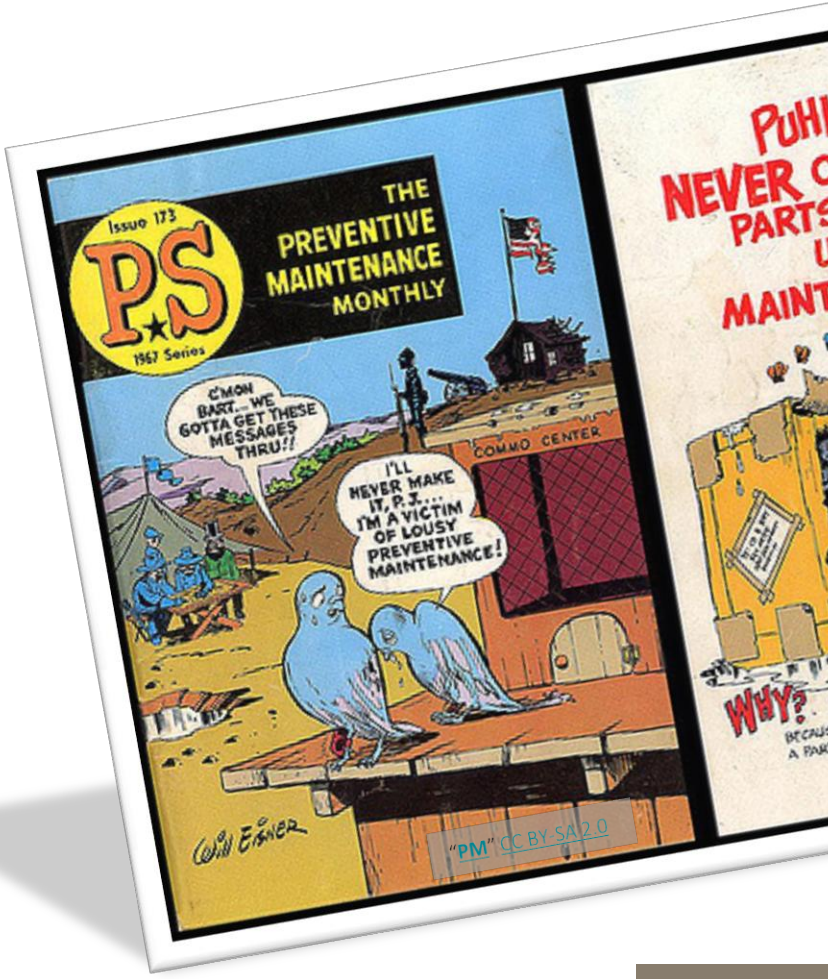
new material

Defined by scripts that parse the title database: shell, perl, python, awk

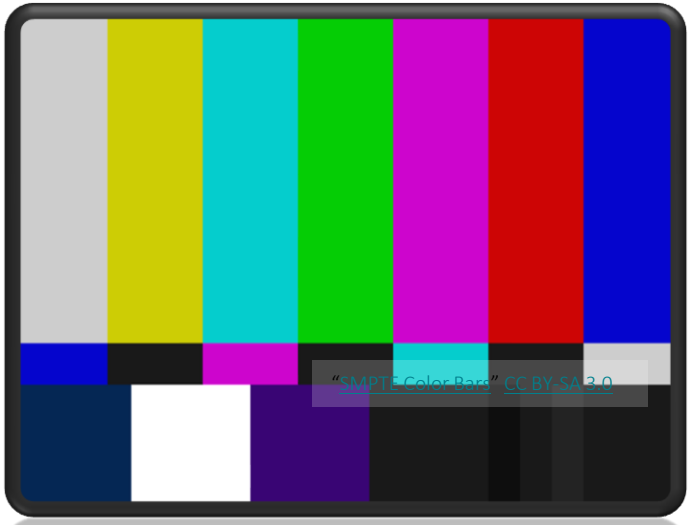
- add new publishers & journals
- new volumes to add, predictable & unpredictable
- find new manifest pages (1x/wk)
- content releases to GLN (~1x/mo)

preventative maintenance

- old volumes have moved, developed problems
- merge metadata for multiple networks
- compare the catalog to the network
- QA. typos, duplicate ISSNs, duplicate volumes, malformed parameters



testing



- Test content against software
  - two servers
  - 12 hours apart
- Errors
  - No subscription
  - Permission statement missing or malformed
  - No volume exists
  - Malformed lists of issues, articles, or links
  - URL redirects (journal has moved)
  - No articles
  - HTML crawl errors (can't access, taking too long, missing, moved)
  - Transient changes, rotating ads, dynamic content, dynamic file generation, watermarking

LOCKSS?

- lots of **copies** and **communities** keep stuff safe
- a LOCKSS network is a **peer-to-peer** network
- websites are **not predictable**
- LOCKSS addresses issues of **data relationships** and **metadata**

have we collected it?

- How do we know?