MARAC Fall 2024

Session 12

Suddenly Digital: Acquiring Pragmatic Digital Archives Skills on the Job

Thanks to our sponsor

Suddenly Digital: Acquiring Pragmatic Digital Archives Skills On the Job

Emily Higgs Kopin

Head of Digital Collections Strategy, Swarthmore College

Agenda

5 minutes: Housekeeping and intro

40 minutes: Presentation (with handout)

30 minutes: Crowdsourced Q&A

The recorded presentation, slides, and handout will all be available online after this session.

Slides: bit.ly/EHK-MARAC-2024

Handout: bit.ly/EHK-Handout

“Learn to code” is bad advice.

“Learn to sew.”

What tasks do I need to accomplish?

Example “Digital Archivist” responsibilities

- **Acquisition of born-digital materials**. Acquires born-digital content as identified by the Curator as within collecting scope. In consultation with the Head of Digital Collections Strategy, advises the Curator on appraisal strategy, privacy concerns, and security, as they apply to the context of born-digital records. Administers file transfer following best practices and utilizing appropriate tools.
- **Ingest and processing of born-digital materials.** Follows best practices and utilizes appropriate tools, such as BitCurator, to process born-digital archival materials.
- **Management and preservation of born-digital and digitized content.** Ingests content from digitization projects and newly acquired born-digital records, as well as legacy content from other systems, into digital asset management system. Ensures that content is reliably preserved. Monitors file format obsolescence and other risks.
- **Access to born-digital and digitized content.** Works to ensure that digital content in the collections is as accessible as possible to appropriate audiences, following the strictures of privacy concerns, copyright regulations, and donor restrictions. Works with Head of Digital Collections Strategy towards improving user experience across discovery systems and access tools. Works toward including content in multi-institutional aggregators.
- **Metadata management.** Ensures accuracy of descriptive, structural, and administrative metadata. Maintains an awareness of metadata standards for digital collections (e.g. DACS,, DCMI, MODS), including best practices for socially-conscious and reparative (re)description, and ensures conforming metadata standards across collections. Demonstrates a facility with authorities/thesauri (e.g. LCSH/LCNAF, AAT), including linked data services. Audits and performs clean-ups of metadata as needed.

What if I don't have tasks yet? I don't even know what I'm supposed to be doing.

Digital Preservation Coalition: https://www.dpconline.org/

- Novice to Know-How (Free overview course)
- Issue-specific Technology Watch Publications
- Digital Preservation Handbook

Theory and Craft of Digital Preservation by Trevor Owens

https://saaers.wordpress.com/

Will a coding class help? Yes!

Taking an “Intro to coding” class...

- …can give you a basic understanding of computing logic.
- …can introduce you to syntax of a particular coding language.
  - Syntax varies across programming languages.
- …can help you identify tasks that can be accomplished computationally.
  - Counting and math
  - Matching patterns
  - Sorting

I need to run digital archiving software.

Method/Tool	**Learning Path(s)**
Command Line (CLI) - instructions given to your computer through typing out commands. Many digital archives tools are run this way, and common functions for processing and manipulating files in bulk are easily accomplished using the command line.	• Ashley Blewer • Library Carpentry • Learn Enough Command Line to be Dangerous (Michael Hartl) - Freemium • Bite Size Command Line by Julia Evans • Jason Raitz
BitCurator - a pre-assembled toolbox filled with digital forensics tools that help digital archivists with core archival functions. It can be installed on a Linux computer or run in a virtual machine (check out VirtualBox if you're new to virtual machines). Many of the tools included in BitCurator can be installed and run individually on your computer, and the BitCurator documentation often includes helpful information about how to use these tools (even outside of the BitCurator environment) in addition to the official documentation for each program	• BitCurator Docs ○ or View html pages on Github • Quick Start Guide • Overview of Tools The BitCurator Consortium maintains resources, guides, workflows, and a listserv that are great for general digital archives queries in addition to BitCurator-specific issues.

```
tesseract myimage.png output
```

Most programs will have CLI instructions in the documentation somewhere online.

For a program you need help with, check the “man.” (or “--help”)

```
man grep
```

NAME

grep, **egrep**, **fgrep**, **rgrep**, **bzgrep**, **bzegrep**, **bzfgrep**, **zgrep**, **zegrep**, **zfgrep** – file pattern searcher

SYNOPSIS

grep [**-abcdDEFGHhIiJLlMmnOopqRSsUVvwXxZz**] [**-A** num] [**-B** num] [**-C** num] [**-e** pattern] [**-f** file] [**--binary-files=**value] [**--color**[**=**when]] [**--colour**[**=**when]] [**--context=**num] [**--label**] [**--line-buffered**] [**--null**] [pattern] [file ...]

DESCRIPTION

The **grep** utility searches any given input files, selecting lines that match one or more patterns. By default, a pattern matches an input line if the regular expression (RE) in the pattern matches the input line without its trailing newline. An empty expression matches every line. Each input line that matches at least one of the patterns is written to the standard output.

grep is used for simple patterns and basic regular expressions (BREs); **egrep** can handle extended regular expressions (EREs). See re_format(7) for more information on regular expressions. **fgrep** is quicker than both **grep** and **egrep**, but can only handle fixed patterns (i.e., it does not

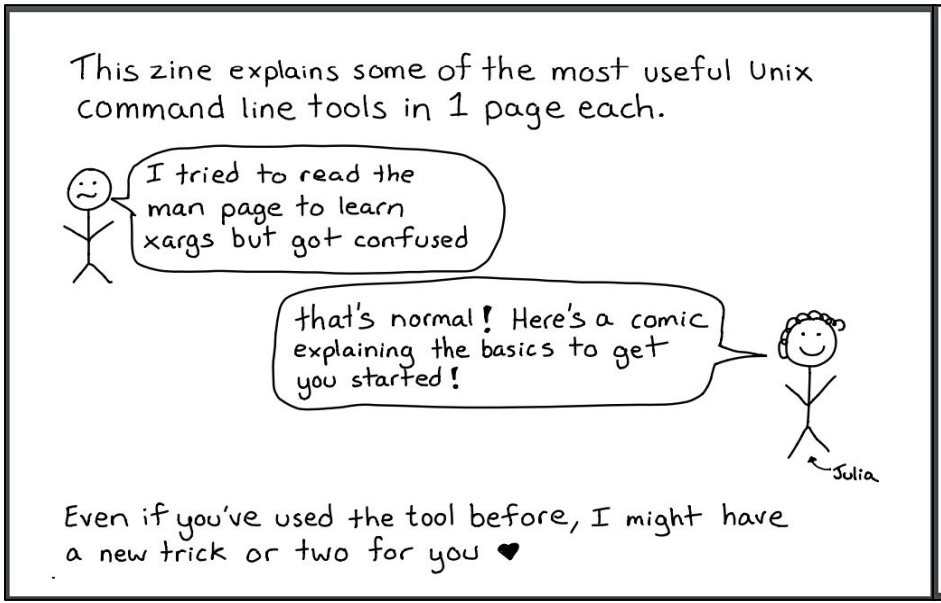
tesseract --help

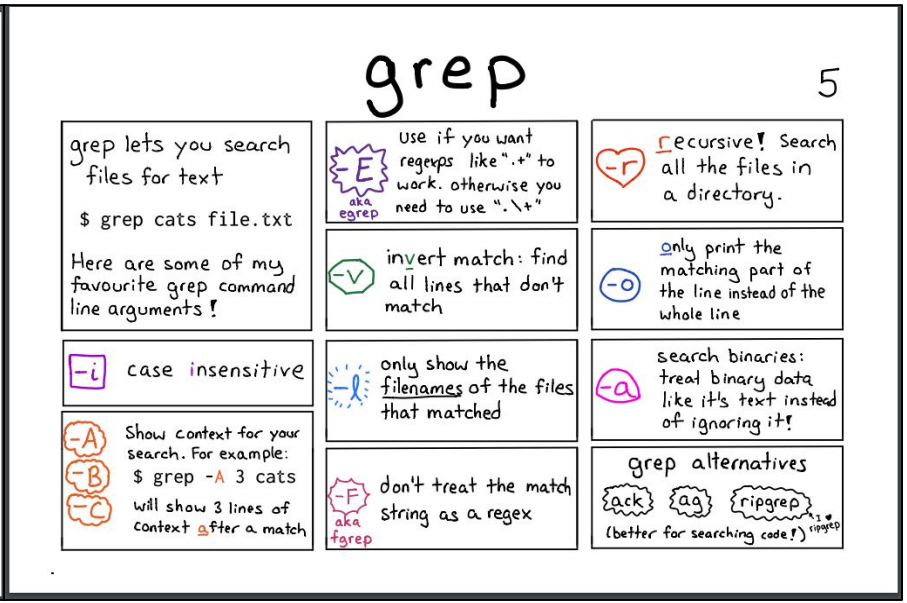
```
Usage:
  tesseract --help | --help-extra | --version
  tesseract --list-langs
  tesseract imagename outputbase [options...] [configfile...]

OCR options:
  -l LANG[+LANG]        Specify language(s) used for OCR.
NOTE: These options must occur before any configfile.

Single options:
  --help                Show this help message.
  --help-extra          Show extra help for advanced users.
  --version             Show version information.
  --list-langs          List available languages for tesseract engine.
```

Julia Evans's "Bite Size Command Line" - https://wizardzines.com/





A brief interlude: what is BitCurator?

BitCurator is a pre-assembled toolbox filled with digital forensics tools that help digital archivists with core archival functions.

“Bitcurator” also sometimes refers to the community that creates, maintains, and uses digital forensics tools through BitCurator and related digital archives software.

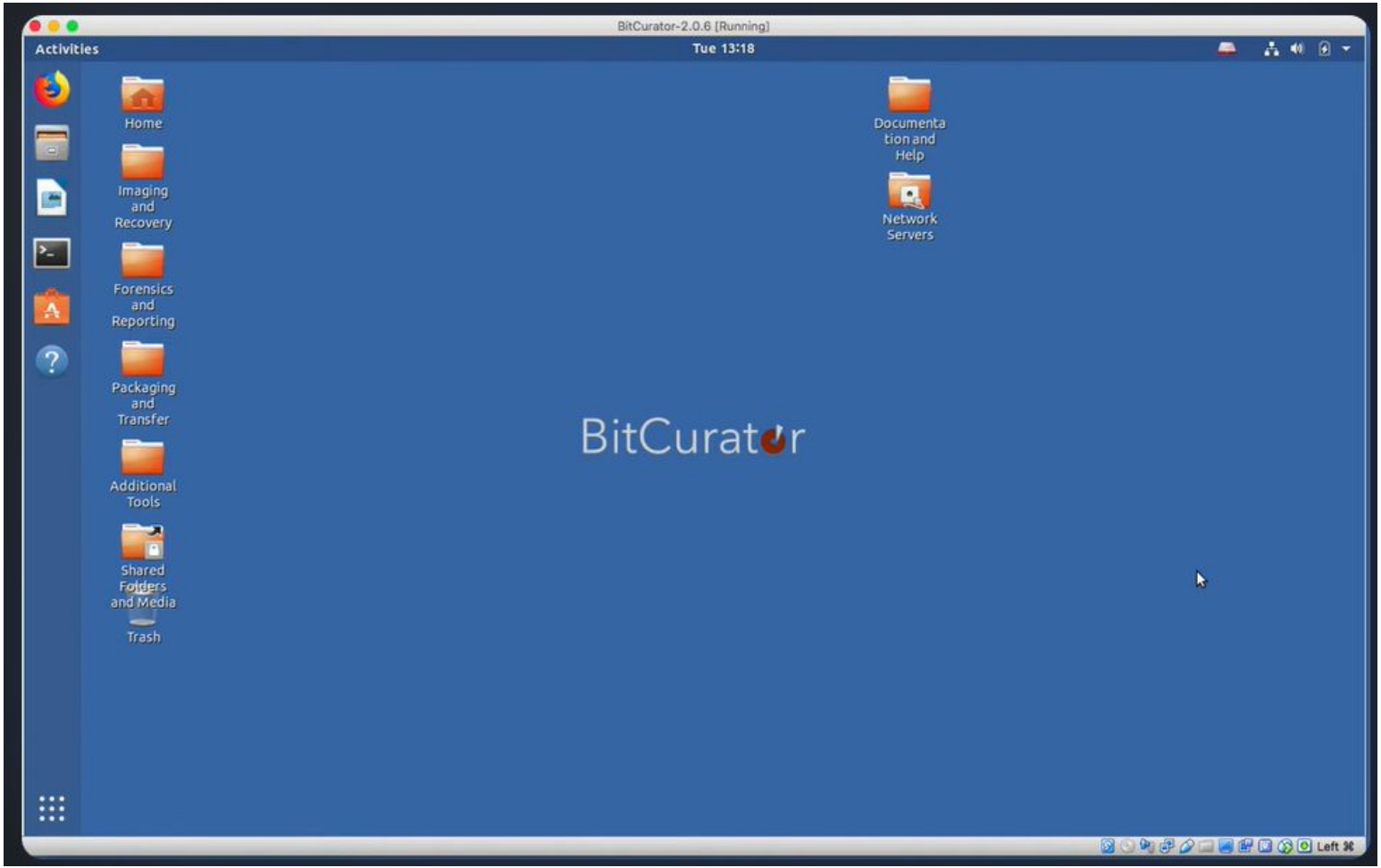


Image courtesy of BitCurator documentation

- 1 Imaging and Recovery
  - 1.1 Safely Mount Devices
  - 1.2 Create Disk Images
- 2 Forensics and Reporting
  - 2.1 Extract metadata from disk images and files
  - 2.2 Identify potentially private and sensitive information
  - 2.3 Generate Forensic Reports
  - 2.4 Deduplicate Files
  - 2.5 File Export
  - 2.6 Forensics and Reporting Nautilus Scripts
- 3 Package and Export
  - 3.1 Package files and metadata with BagIt Python
  - 3.2 Transfer files and metadata with grsync

Image courtesy of BitCurator documentation

I need to transfer files off of storage media and/or between storage locations.

Method/Tool	Learning Path(s)
Disk imaging - creating an exact copy of a storage device (like a hard drive, CD, or floppy disk) with the entire structure of the disk, including system files and sometimes hidden/deleted files. Popular tools for disk imaging include FTK Imager, Guymager, GNU ddrescue, and others.	• Digital Archival traNsfer, iNgest, and packagiNg Group • Ashley Blewer • BitCurator Quick Start Guide (Guymager)
A/V Formats and Tools (there are many)	Ashley Blewer: https://ashleyblewer.com/ • Pragmatic Audiovisual Preservation • A/V Artifact Atlas • Open Workflows for A/V Archiving
Bag-It Python Library - a python utility used to structure files and metadata according to the commonly-used BagIt specification. This particular way of ordering files is compatible with a variety of programs (including checksum validators) and helpfully stores metadata and reports alongside the data files.	• BagIt at the Library of Congress • GitHub • Ethan Gates • NYPL Documentation

search

- AV Artifact Atlas
- About AVAA
- Browse by Tags
- Browse by Categories
- Browse by Formats
- Gallery
- Unknown Artifacts
- Contributors Guide
- Glossary
- FAQ

You can browse the site using the search box in the sidebar, the tag cloud below, or this list of artifacts.

2-inch Quad | Abrasion | Analog | Audio | Azimuth | Baking | Balanced Cables | Banding | Betamax | Black Levels | Cable | Cassette | Chrominance | Cleaning | Clicks | Clipping | Color | Color lock | Color smearing | Comet Tails | Common Artifacts | Curvature error | DV | DV25 | Deck Failure | Deformation | Device Error | Digital | Digital Clicks | Digitization | Dirt | Disc Playback | Discrete Cosine Transform | Dot Crawl | Dropout | Dying Batteries | Edge curl | Editing cut | Electrical | Error | Expansion | Extended Play | Film | Flange pack | Flicker | Frequency Loss | Ghosting | Gouging | Head Clog | Hue | Hum | Interference | LP | Levels | Levels Too Hot | Long Play | Luminance | Luminance Blooming | MPEG | Magnetic Tape | Magnetostrictive Action | Media Damage | Media Failure | Mono | Noise | Open Reel Tape | Operator Error | Overs | Oxide Out | PAL | Phase | Pitch | Playback Adjustment | Popped strand | Post-echo | Power Supply | Pre-echo | Processing Artifact | Production Error | Pulse | Quilting | RF

training.ashleyblewer.com

Computers (Understanding)

一、Computers
二、Fixity
三、Networks
四、Storage
五、Databases

Computers (Talking to)

一、Command Line Interface
二、(Bash) Scripting
三、Git 100
四、Git 101
五、Metadata

Digital Preservation

一、Digital Preservation
二、OAIS
三、Fixity
四、Storage
五、Digital Forensics

FFmpeg

一、FFmpeg
二、FFplay
三、FFprobe
四、FFV1
五、FFmpeg + Art
六、FFmpeg + Preservation

Media Analysis

一、Video
二、MediaInfo
三、MediaConch
四、QCTools
五、FFprobe
六、(Other) Analysis Tools

```
myfirstbag/
|-- data
|   \-- 27613-h
|       \-- images
|           \-- q172.png
|           \-- q172.txt
|-- manifest-md5.txt
|     49afbd86a1ca9f34b677a3f09655eae9 data/27613-h/images/q172.png
|     408ad21d50cef31da4df6d9ed81b01a7 data/27613-h/images/q172.txt
\-- bagit.txt
      BagIt-Version: 0.97
      Tag-File-Character-Encoding: UTF-8
```

I need to identify/isolate patterns in data.
I need to normalize, clean, and edit data.

Method/Tool	Learning Path(s)
Regular Expressions - a language for pattern-matching embedded in (or compatible with) *many* other computational tools.	• Library Carpentry • RegexOne • The Programming Historian • Regular-Expressions.info • Core RegEx Cheatsheet • Python's RegEx flavor • James Truitt
OpenRefine - an open-source application for data cleaning and transformation. It is commonly used for spreadsheets, but can handle other data formats too.	• Library Carpentry • University of Illinois Libguide • Scotty Carlson • Code4Lib 2018 Cheat Sheet

Python - a general-use programming language often recommended for its legibility.
 	
Pandas - a Python library specifically designed for dealing with messy data.
	• Michelle Janowiecki:
	  ○ Python Lessons for Librarians
	  ○ Pandas for Metadata
	• Library Carpentry
	• ITHAKA Constellate

rexegone.com

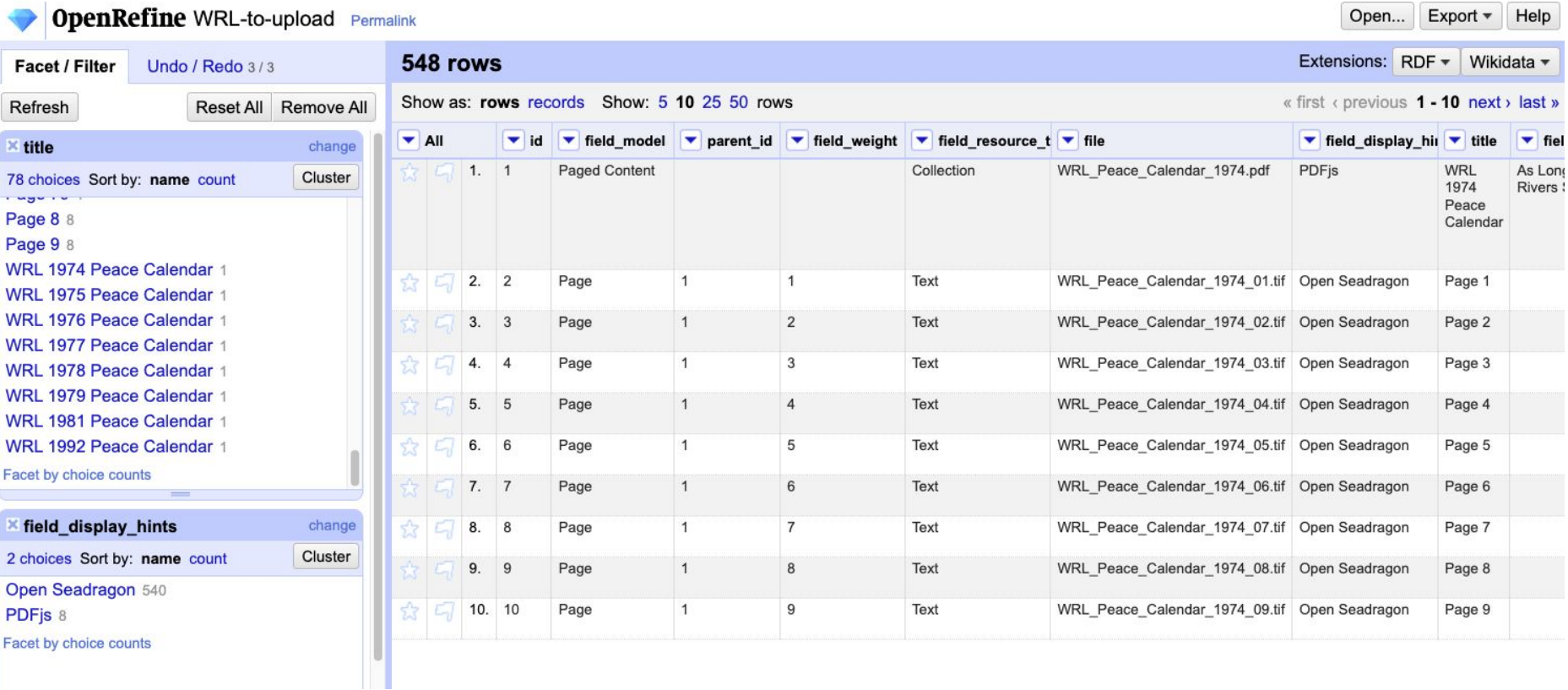
Exercise 4: Excluding Characters

Task	Text	
Match	hog	✓
Match	dog	✓
Skip	bog	

[^b]og

Continue ›

Solution The simplest solution to match any line that ends in 'og' but is not 'bog' would be the expression '[^b]og'. Alternatively, you could use what we learned from the previous lesson and use '[hd]og' to match 'hog' and 'dog' but not 'bog'. Note that it is slightly more restrictive expression because it limits the strings it can match.


OpenRefine WRL-to-upload
Permalink
Open...
Export
Help
Facet / Filter
Undo / Redo 3 / 3
548 rows
Extensions:
RDF
Wikidata
Refresh
Reset All
Remove All
Show as: rows records
Show: 5 10 25 50 rows
« first ‹ previous 1 - 10 next › last »
title
change
78 choices Sort by: name count
Cluster
Page 8 8
Page 9 8
WRL 1974 Peace Calendar 1
WRL 1975 Peace Calendar 1
WRL 1976 Peace Calendar 1
WRL 1977 Peace Calendar 1
WRL 1978 Peace Calendar 1
WRL 1979 Peace Calendar 1
WRL 1981 Peace Calendar 1
WRL 1992 Peace Calendar 1
Facet by choice counts
field_display_hints
change
2 choices Sort by: name count
Cluster
Open Seadragon 540
PDFjs 8
Facet by choice counts
All
id
field_model
parent_id
field_weight
field_resource_t
file
field_display_hi
title
fiel
1. 1 Paged Content Collection WRL_Peace_Calendar_1974.pdf PDFjs WRL 1974 Peace Calendar As Long Rivers
2. 2 Page 1 1 Text WRL_Peace_Calendar_1974_01.tif Open Seadragon Page 1
3. 3 Page 1 2 Text WRL_Peace_Calendar_1974_02.tif Open Seadragon Page 2
4. 4 Page 1 3 Text WRL_Peace_Calendar_1974_03.tif Open Seadragon Page 3
5. 5 Page 1 4 Text WRL_Peace_Calendar_1974_04.tif Open Seadragon Page 4
6. 6 Page 1 5 Text WRL_Peace_Calendar_1974_05.tif Open Seadragon Page 5
7. 7 Page 1 6 Text WRL_Peace_Calendar_1974_06.tif Open Seadragon Page 6
8. 8 Page 1 7 Text WRL_Peace_Calendar_1974_07.tif Open Seadragon Page 7
9. 9 Page 1 8 Text WRL_Peace_Calendar_1974_08.tif Open Seadragon Page 8
10. 10 Page 1 9 Text WRL_Peace_Calendar_1974_09.tif Open Seadragon Page 9

Home

Last edited by **Michelle Janowiecki** 4 years ago

Welcome to the Python lessons for librarians wiki!

Why Python? Why librarians?

Python is a great and popular coding language in libraryland! I am a metadata librarian, and I frequently use Python to:

- Clean up metadata stored in spreadsheets
- Get metadata out of repositories and applications using APIs
- Put metadata into repositories and applications using APIs
- Extract metadata from MARC records
- Get URIs for controlled vocabulary terms
- Convert metadata from one format to another (CSV --> JSON, MARC --> CSV, RDF--> CSV)
- Merge metadata based on identifiers

Why these lessons?

In my Python journey, I was extremely lucky to have a mentor at my library who encouraged me to learn Python, helped me with projects, and answered my never-ending questions. I know though that this kind of support and encouragement is unfortunately rare in our busy (and often dysfunctional) libraries. I think of these lessons as a way to try to provide my fellow library workers with what I've learned and offer a space for growth and support.

jupyter jupyter-interactive-visualization Last Checkpoint: 3 minutes ago (autosaved) Quit

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Markdown

Interactive Python Visualization Libraries

In [ ]:
```
import numpy as np
import plotly.offline as py
import plotly.figure_factory as ff
from bokeh.models import HoverTool, WheelZoomTool
from bokeh.plotting import figure
from bokeh.io import show, output_notebook
output_notebook()
```

Plotly

In [ ]:
```
py.init_notebook_mode()
t = np.linspace(-1, 1.2, 2000)
x = (t**3) + (0.3 * np.random.randn(2000))
y = (t**6) + (0.3 * np.random.randn(2000))

colorscale = ['#7A4579', '#D56073', 'rgb(236,158,105)', (1, 1, 0.2), (0.98,0.98,0.98)]

fig = ff.create_2d_density(
    x, y, colorscale=colorscale,
    hist_color='rgb(255, 237, 222)', point_size=3
)

py.iplot(fig, filename='histogram_subplots')
```

Bokeh

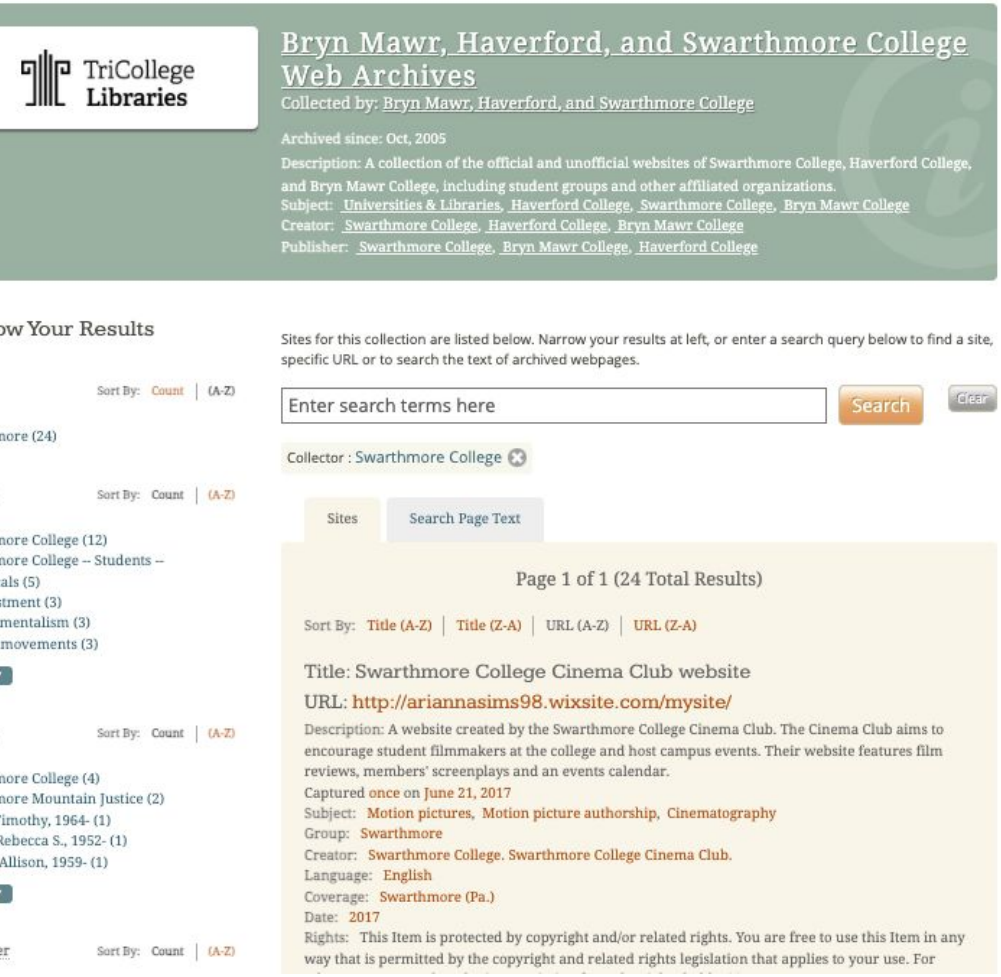
In [ ]:
```
n = 500
x = 2 + 2*np.random.standard_normal(n)
y = 2 + 2*np.random.standard_normal(n)

p = figure(title="Hexbin for 500 points", match_aspect=True,
           tools="wheel_zoom,pan,reset", background_fill_color='#440154')
p.grid.visible = False

r, bins = p.hexbin(x, y, size=0.5, hover_color="pink", hover_alpha=0.8)
```

I need to collect and preserve web-based content.

Method/Tool	Learning Path(s)
Archive-It - a web archiving service for cultural heritage institutions that crawls web pages and stores copies in the Internet Archive, accessible through the Wayback Machine.	• Guide for New Archive-It Users • Archive-It Video Curriculum • Intro to Archive-It Community Webs Webinar
Conifer - a web-based web archiving service (formerly webrecorder.io) that creates and stores an interactive copy of a web page through recording your interactions. Free and paid accounts are available.	• Quick Start User Guide • Norfolk Record Center Tutorial



Articles in this section

- Archive-It Video Curriculum
- What's the difference between the General Archive and Archive-It?
- What is web archiving?
- WARC Naming Conventions
- Archive-It workflow
- Web Archiving Lifecycle White Paper
- Archive-It Crawling Technology
- Want to know more about Archive-It?
- Glossary of Archive-It and Web Archiving Terms

Archive-It Video Curriculu

Sylvie Rollason-Cass
Updated 10 months ago

On this page:

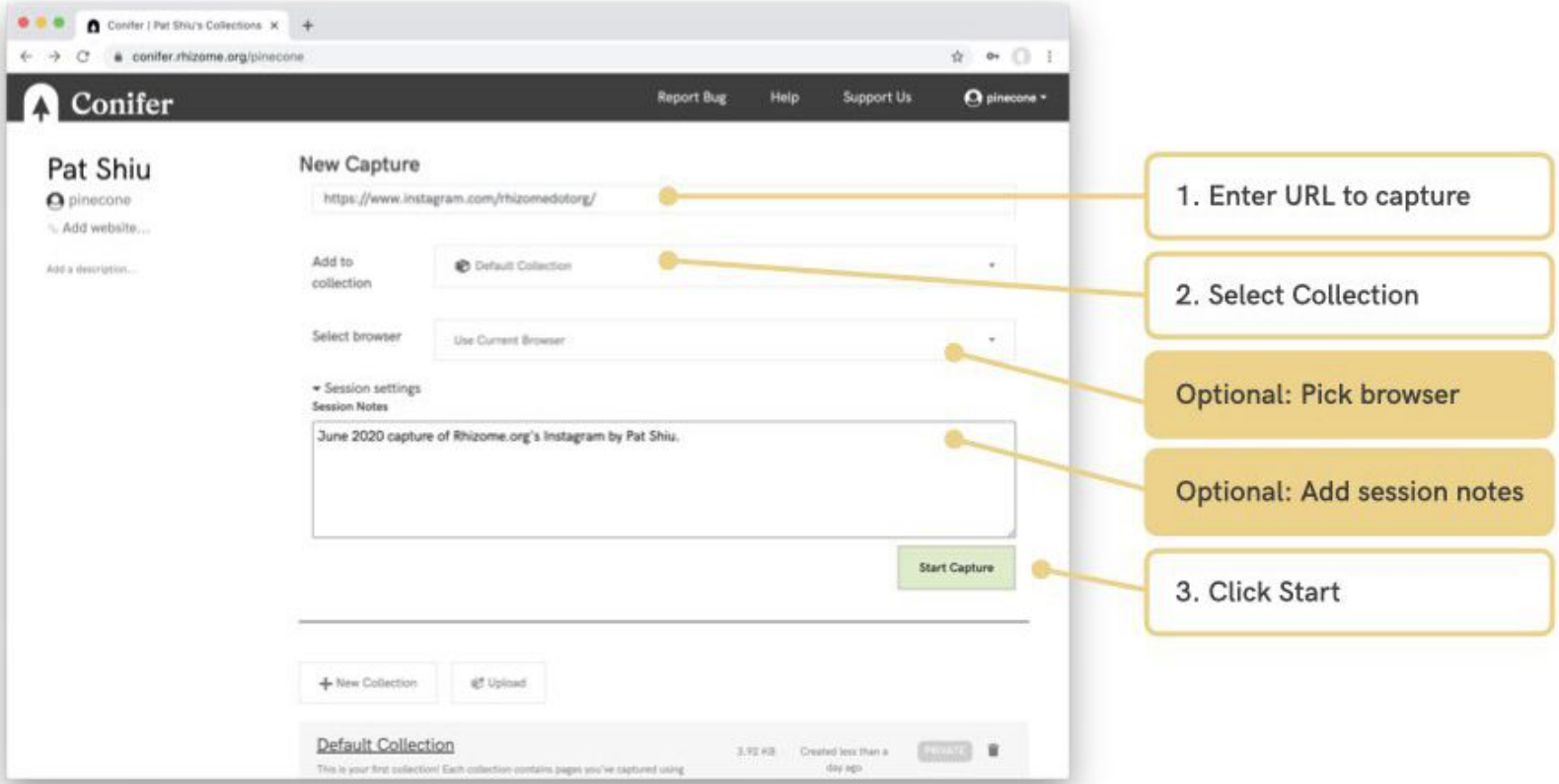


Image courtesy of Conifer documentation

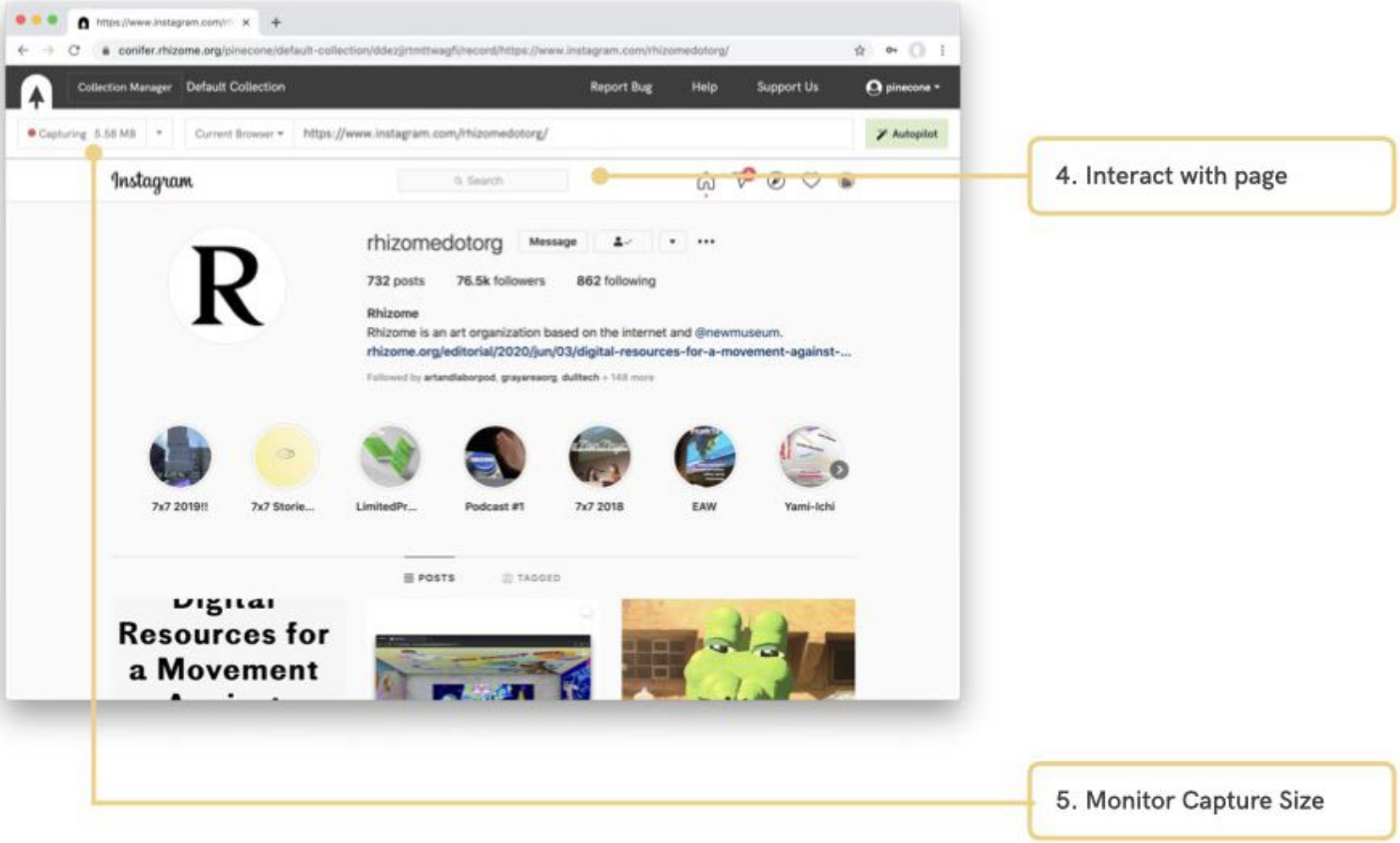
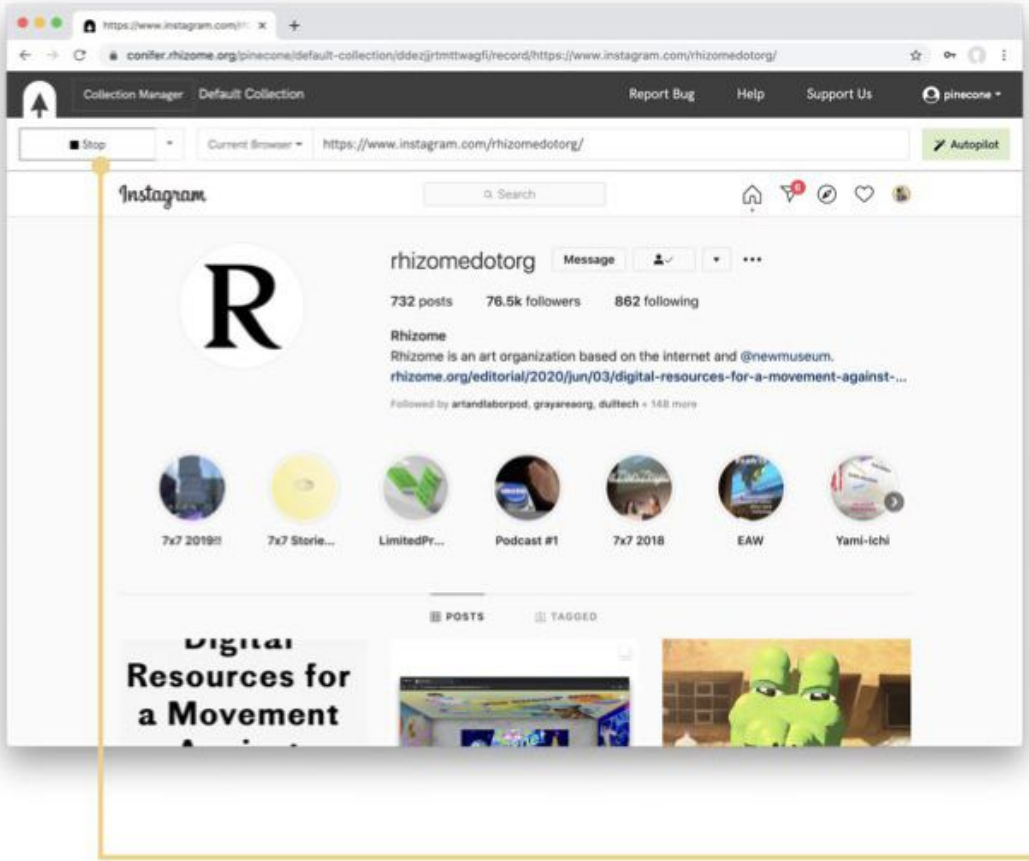


Image courtesy of Conifer documentation



6. Click Stop to finish capture

Image courtesy of Conifer documentation

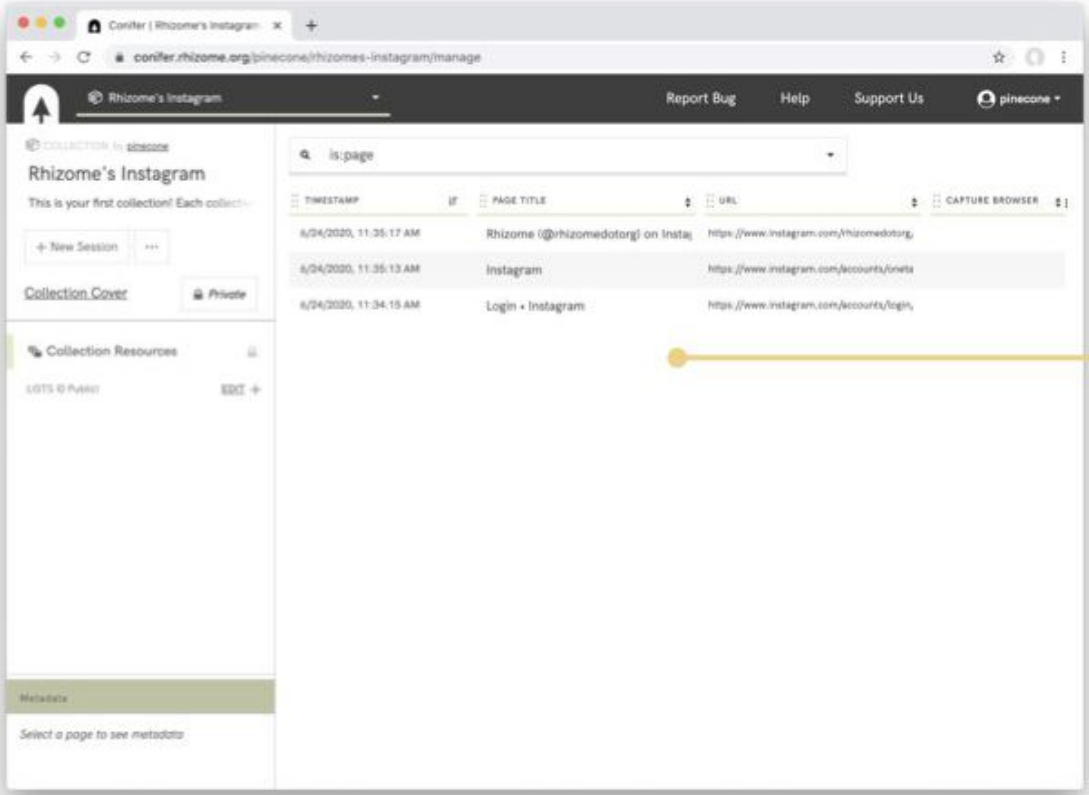


Image courtesy of Conifer documentation

I need to collect and preserve email.

Method/Tool	Learning Path(s)
ePADD - an open-source application to appraise, process, search, and view archived emails.	● User Guide v11 ● ePADD at iPres 2022
Mailbag - a specification and packaging tool that saves emails in preservable packages, including pdfs for access.	● Mailbagit on Github ● Documentation ● Greg Wiedeman Presentation

ePADD | Import | Browse | Search | Export | Save

Appraisal | Dashboard

Email Archive of Jeb Bush

9,838
Correspondents

8,120
Entities

8
Labels

1,003
Image Attachments

2,126
All attachments

4
Folders

Lexicon Search

Reports

More

2.1 Example of a mailbag

```
<base directory>/
     |
     +-- bagit.txt
     |
     +-- bag-info.txt
     |
     +-- mailbag.csv
     |
     +-- manifest-<algorithm>.txt
     |
     +-- tagmanifest-<algorithm>.txt
     |
     +-- data/
          |
          +-- mbox/
          |     +-- [payload files]
          +-- pdf/
          |     +-- [payload files]
          +-- warc/
          |     +-- [payload files]
          +-- attachments/
                +-- [Mailbag-Message-ID]/
                |     -- attachments.csv
                |     -- [payload files]
                +-- [Mailbag-Message-ID]/
                      -- attachments.csv
                      -- [payload files]
```

I need to figure out how to deal with this weird format.

Method/Tool	Learning Path(s)
JHOVE - an open-source tool for file format identification, validation, and characterization. A well-loved tool for identifying uncommon or unknown file formats, and for embedding in archival processing workflows for digital objects.	• JHOVE Documentation • A Beginner's Guide
PRONOM - a registry of file formats, software products and other technical components maintained by the UK National Archives **DROID** - a tool that identifies file formats using PRONOM. It extracts other useful metadata as well.	• PRONOM Homepage • PRONOM in Practice Webinar • DROID User Guide
Library of Congress - maintains sustainability information about digital content and formats.	• Sustainability of Digital Formats • Recommended Formats Statement
r/DataHoarder - a subreddit dedicated to "data hoarders," i.e. hobbyist digital archivists. This is a great place to ask questions about niche data formats, obscure storage media, etc. (follow advice at your own risk!)	• DataHoarder Wiki

The technical registry
PRONOM

Welcome : About Add an entry
Search Help Information resources

Help : report on simple search

Search Results

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles | Migration Pathways

You searched for: "filemaker"

Save as... XML | CSV Print

page 1

FileMaker Pro Database (2)

FileMaker Pro is a cross-platform relational database application from FileMaker Inc. Early FileMaker files may lack a file format extension, particularly those created in an Apple Macintosh environment.

FileMaker Pro Database (1)

FileMaker Pro is a cross-platform relational database application from FileMaker Inc. Early FileMaker files may lack a file format extension, particularly those created in an Apple Macintosh environment.

FileMaker Pro Database (3)

FileMaker Pro is a cross-platform relational database application from FileMaker Inc.

FileMaker Pro Database (5)

FileMaker Pro is a cross-platform relational database application from FileMaker Inc.

FileMaker Pro Database (7+)

FileMaker Pro is a cross-platform relational database application from FileMaker Inc.

FileMaker Pro Database (12)

FileMaker Pro is a cross-platform relational database application.

Home
Popular
Explore
All

CUSTOM FEEDS
Create a custom feed

RECENT
r/DataHoarder
r/datahoarders
r/CrohnsDisease
r/neopets
r/OccipitalNeuralgia

COMMUNITIES
Create a community
r/AmateurRoom...

r/DataHoarder · 2y ago

Looking for an iOS app to make PDFs of a book. If this is not the right place for this post, please delete it.

0 votes · 18 comments

r/DataHoarder · 4y ago

MEGATHREAD: Archiving the Capitol Hill Riots

34K votes · 2.6K comments

r/DataHoarder · 2y ago

Any iOS file manager with SMB connectivity and a photo library sorted by file size?

3 votes · 12 comments

r/DataHoarder · 4y ago

Rescue Mission for Sci-Hub and Open Science: We are the library.

8.4K votes · 985 comments

r/DataHoarder · 1y ago

HELP!! How to sync pictures across iOS and Android

2 votes · 4 comments

r/DataHoarder · 5y ago

It's A Digital Disease!

This is a sub that aims at bringing data hoarders together to share their passion with like minded people.

786K Members

224 Online

Can I just get a list of tools please?

Share | Copy Link

Infra Finder: Discover open infrastructures that meet your needs

Search

92 Solutions

Sort by: Last Updated

Filters

Clear All

Update Results

Solution Category

Open Attributes

Technical Attributes

medRxiv

Cold Spring Harbor Laboratory

United States of America

medRxiv (pronounced "med-archive") is a free online archive and distribution server for complete but unpublished manuscripts (preprints) in the medical, clinical, and related health sciences.

Repository ser

DSpace

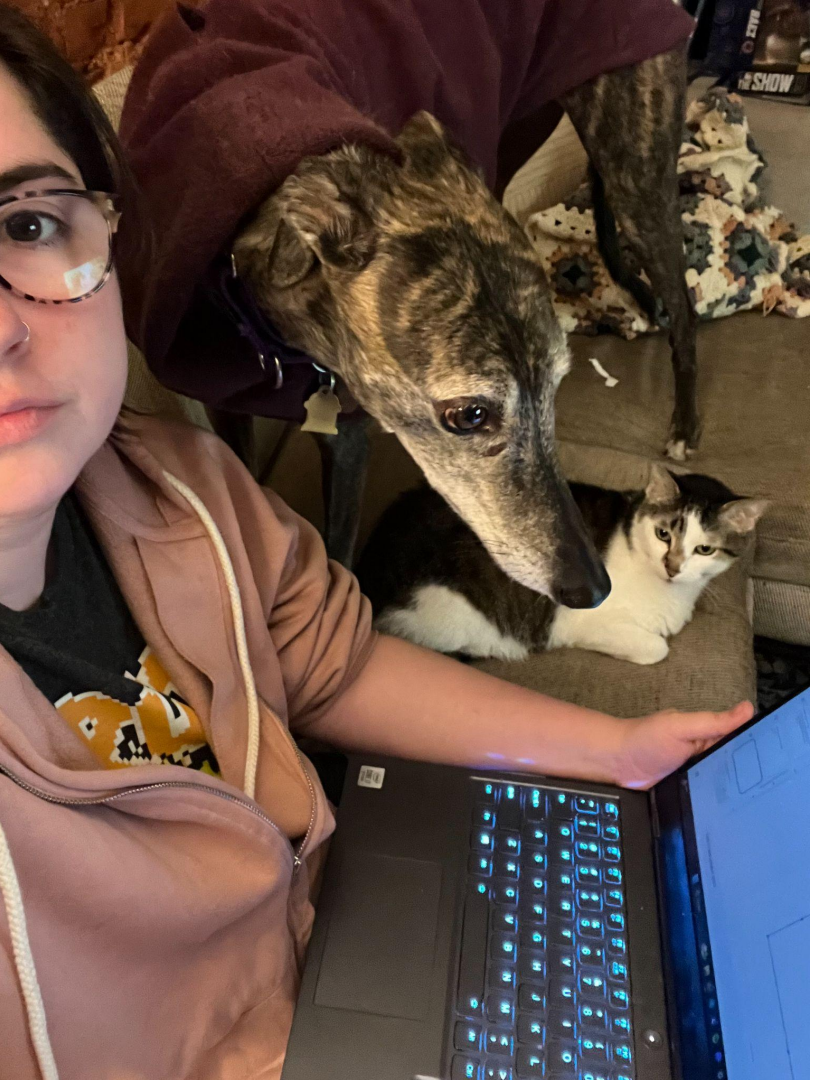
Lyrasis

United States of America

DSpace is a web application, allowing researchers and scholars to publish documents and data. While DSpace shares some feature overlap with content management systems and document management systems, the...

Repository software

Knowing how to learn something new *is* the defining skill of a digital archivist.



Special Thanks

James Truitt, Digital Archivist, Swarthmore College

Everyone at the DLF professional development networking session this year

Ashley Blewer, whose work I adore

My email: ehiggs1@swarthmore.edu

Crowdsourced Q&A

bit.ly/EHK-CommunityDoc

- Anonymity allowed, following the code of conduct.
- No question is too basic or too niche.
- Feel free to add your own tips and resources.