5-4-2024

Processing Priorities, Researcher Use, and Programmatic Improvement

Megan M. Mummey

University of Kentucky, megan.mummey@uky.edu

Processing Priorities, Researcher Use, and Programmatic Improvement

MAC 2024

Megan Mummey, Director of Manuscript Collection, University of Kentucky Libraries

From the Louis Edward Nollau F Series Photographic Print Collection

Today I'm going to talk about a data analysis project I've been working on since 2018. In short, I've been using the data generated by our research requesting system to make decisions and set priorities for the archival processing of manuscript collections. And let me tell you I find this process to be one of the single most enjoyable things that I do. A. I love spreadsheets and structured data! B. ask any processing archivist and they will tell you that there is nothing like the thrill (the absolute high) of seeing someone use a collection you've poured your blood, sweat, and tears into (not literally).

How are processing priorities set?

- Subjectively
- "Problem in front of you"
- Processing on demand
- As acquired (MPLP)
- Donor needs
- Use based

One of my "problems": 36 ft collection found in storage - unprocessed with no accession record - acquired in 1978

So I first want to talk about how processing priorities are set. There are many different methodologies for this and in a nutshell most of the time it is incredibly subjective. Some of it is "the problem in front of you", like this photograph of a problem I found in storage. You know...just a small 36 cubic foot collection with no documentation. We've had it since 1978. You can do "processing on demand" (so when a researcher wants access to an unprocessed collection you minimally process it before they arrive). You can do it in a "as collections are acquired approach." And still sometimes you have to drop everything because of donor relations. Not all of these methodologies are wrong or necessarily bad, some are better than others. Oftentimes the way we determine our priorities at UK takes a little bit from all of these things.

UKL SCRC Research Room

Slide content:

Context

- UK SCRC fully implemented AEON in 2016
- In 2018 (with 18 months of data) began analyzing check-out data
- Have done this process once a year on use data from 2016-2022
- Never had a true backlog busting project or made accession information available
- Many requests for collections in our backlog

But in 2016 when my institution implemented AEON – I had a question –what would happen if we looked at the data AEON collects and have it inform our processing decision making. What would the data tell us? This was the first time that my institution had any reliable system for tracking collection usage in our research room. Before it – everything was paper-based, and nothing was done with those paper slips. AEON collects a lot of data! Which makes it at times challenging to use said data.

So, in 2018 with about 18 months of well-formed, consistent data – I began analyzing which boxes and collections were checked out in our research room. In Spring 2018, I gave a presentation as a part of a panel at the National AEON symposium about my process, the results, and future directions. This presentation is essentially an update of that presentation with data from 2016-2022. (I have not had time to really dig into 2023 yet).

Other bits of context you might need to know is that we've never had a true backlog busting project or made our accession information publicly available. Now – we do get a lot of requests for unprocessed collections in our backlog though. Because of past management decisions there are a lot of citations floating around in the scholarship for collections in our backlog.

Process

- AEON Transaction Report - Frequency
- Exclude staff check-outs
- Add in collection titles
- Clean up data
- Check online status
- Code collections - subject, collecting area, formats
- Analyze data and generate visualizations
- Help set priorities

|	Accession	Title	Number of uses	Online	Topic	Topic 2
2	91M2	Ann and Harry Caudill papers	60	yes	Authors	Appalachia
3	72M2	Frederick Moore Vinson papers	51	yes	Judiciary	Public Policy
4	63M349	Wickliffe Preston family papers	50	yes	Kentucky -- History	Lexington
5	49w1	Samuel Wilson vertical file collection	48	no	Kentucky -- History	Lexington
6	1979ua003	Otis Singletary papers	42	no	UK President recor	University of Kentu
7	2001ua062	Margaret Lantis papers	34	yes	Faculty papers	Public Health
8	2011ms055	John Arthur Dearinger papers	32	no	Theater	
9	0000ua007	John W. Oswald papers	31	no	UK President recor	University of Kentu
10	1997MS271	Altrusa International of Lexington records	31	yes	Education	Non-Profit
11	2007ua023	Athletic Players Files	27	yes	UK Athletics	University of Kentu
12	76M2, 82M6, 84M	Earle Clements papers	26	yes	Public Policy	
13	1997MS213	Hillenmeyer family papers	25	yes	Kentucky -- History	Lexington
14	85m1	Frontier Nursing Service records	24	yes	Appalachia	Public Health
15	81M3	Stanley Forman Reed papers	23	yes	Judiciary	Public Policy
16	2009ms132.0081	Wade Hall: Jack and Norah Parker letters	23	no	WWI	Military history
17	63M202	Hunt-Morgan family papers	22	yes	Kentucky -- History	Lexington
18	2004ua046	UK photographic services negatives	21	no	University of Kentu	Photographs
19	1997MS282	John D. Whisman papers	21	yes	Appalachia	
20	2005ua015	Joseph Baber papers	20	no	Faculty papers	University of Kentu
21	2009MS172	Pettit, Duncan, Gibson family papers	20	yes	Kentucky -- History	Suffrage
22	2009ms132.0004	Wade Hall: George Canary letters	20	no	WWII	Military history
23	1997ms373	Episcopal Dioces of Lexington records	19	no	Lexington	Religion
24	ua	UK general reference files	18	no	University of Kentu	Reference files
25	2010MS041	James Clay family papers	18	yes	Clay family	Kentucky -- History
26	61M158	Linda Neville papers	18	yes	Appalachia	Public Health
27	80m1	John Sherman Cooper	18	yes	Public Policy	
28	90M1	Appalachian Regional Commission record	17	yes	Appalachia	
29	2004av001	John C. Wyatte Herald Leader photograph	17	no	Photographs	Lexington
30	82M9	John Jacob Niles papers	16	yes	Appalachia	Music
31	2014ms0041	Kentucky Hemp Growers Association and	15	yes	Hemp	Public Policy
32	46M53	John Winston Coleman, Jr. collection on s	14	yes	Slavery	African-Americans
33	xx	WWII posters	14	no	WWII	Military history
34	58M25	Cora Wilson Stewart papers	14	yes	Appalachia	Women
35	77M1	Happy Chandler papers	13	yes	Public Policy	Civil Rights
36	2009ms132.0084V	Wade Hall: Dessie Gum Sharp papers	12	no	WWII	WWI
37	2009ms043	E. I. "Buddy" Thompson papers	12	no	Women	University of Kentu
38	96M8	Frontier Nursing Service medical surveys	12	yes	Appalachia	Public Health
39	1997ms295	Laura Massie papers	12	yes	Civil Rights	African-Americans
40	0000UA107	Margaret Ingels papers	11	yes	Faculty papers	STEM
41	0000ua197	UK Training of the Fighting Mechanic reco	11	yes	WWI	Military history
42	2004ua015	UK Public Information Audio Collection	11	yes	University of Kentu	A/V

My data process is relatively simple, and only uses excel, but of course it involves cleanup and takes time. AEON spits out a frequency report that gives me a spreadsheet of collection identifiers and numbers of box checkouts. Staff check-outs are excluded from this report. Unfortunately, this report does not differentiate between books and collections, so those have to be weeded out. Also the identifiers are only as good as the data entry. Sometimes there is some noise, junk identifiers, identifiers that are essentially the same but entered differently (you know normal human error; capitals vs. lowercase; extra spaces). So there's data cleanup that I perform as well as adding collection titles to be more human readable than just the collection identifier. I also check whether these collections have an online guide including whether it has a full box listing vs. just a collection-level guide. Over the years, I've added some data points such as year the collection guide went online and what program they're a part of (these are totally artificial, but can be helpful – so manuscripts, university archives, and audiovisual archives). Once my data is well-formed, I start coding the collections by general topic. I've developed a controlled vocabulary for this process, that I usually add to every year. I try to keep the coding consistent from year to year, so I look at how I've coded the collections in past years. I use my knowledge of the collections to assign between 1 and 6 topics. These can be collecting areas (like Kentucky history, Appalachia, public policy), subjects like specific wars or subject areas like medicine, STEM, suffrage, etc. and sometimes formats microfilm, photographs etc. I compile the number of box checkouts and the number of collections used for each subject. Then I look at the data, analyze it, generate some visualizations, and draw some conclusions

to help set future priorities.

2016-2017 Research Room box Check-outs by Subject
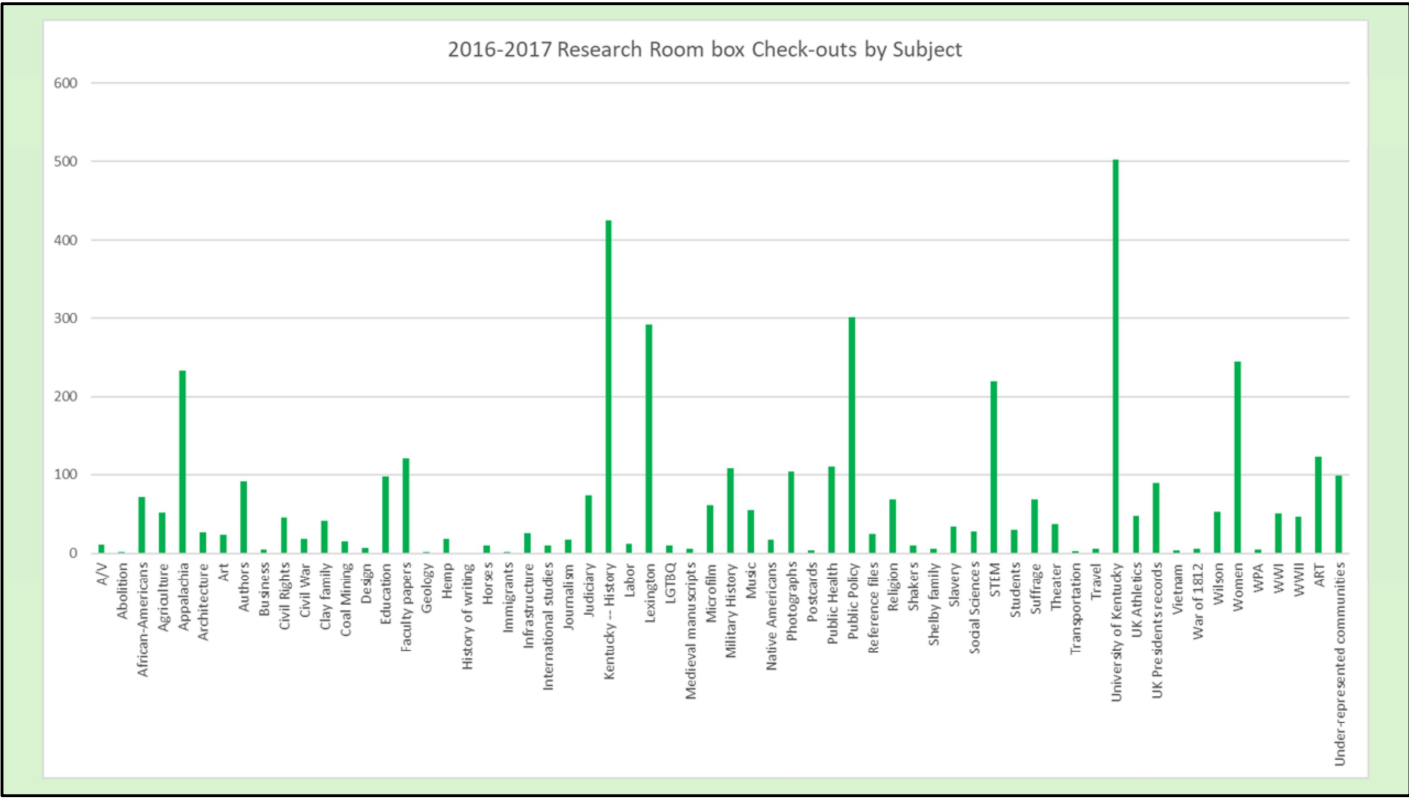
One of the visualizations I generate is a count of each coded term for both collections and box checkouts. This bar graph represents the subject-based topics by box checkouts. This is the first subject term graph I generated from the 2016/2017 data. These largest spikes here, not surprisingly, correspond with our largest collecting areas: Appalachia, University of Kentucky, Kentucky history, Politics and Public Policy.

2022 Research Room Box Check-outs by Subject
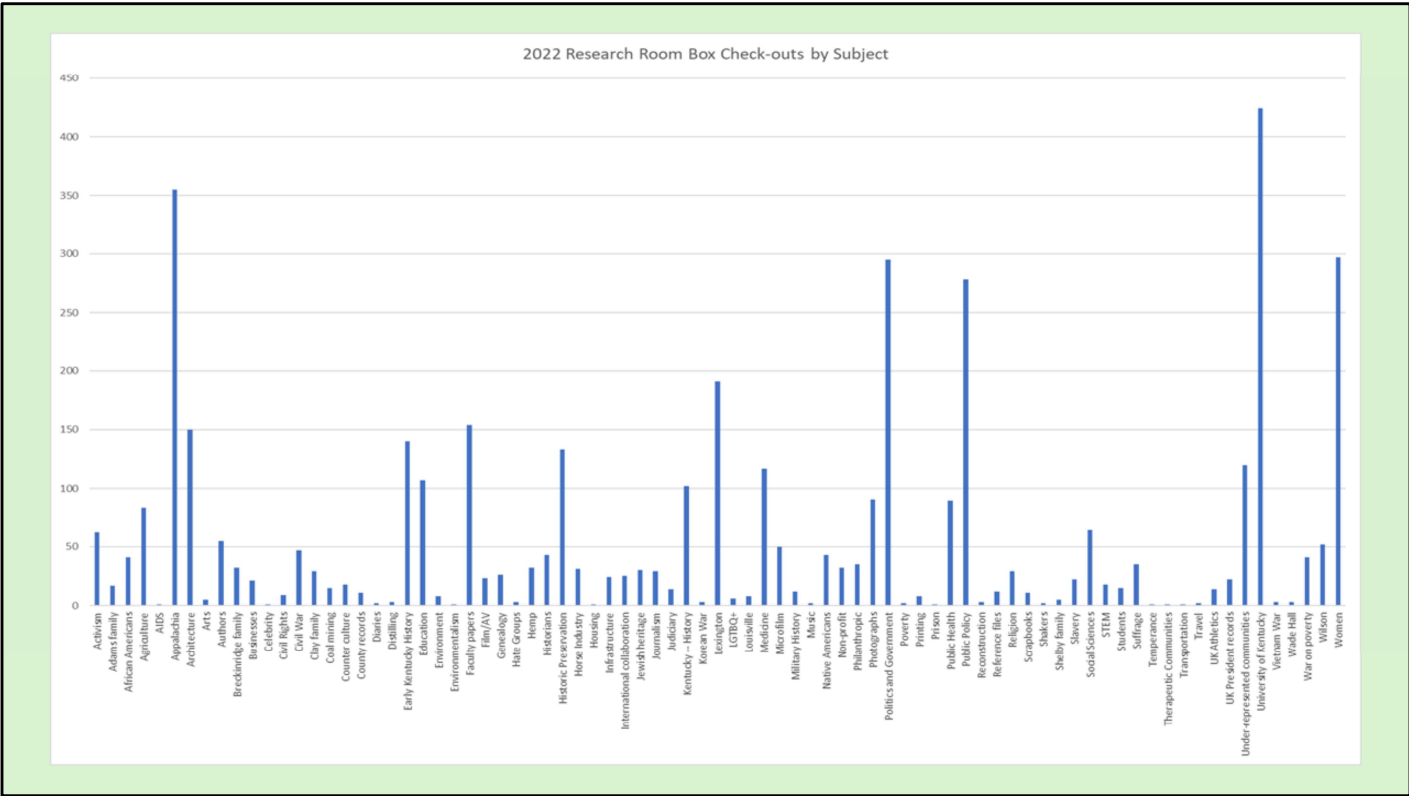
Here is the latest graph I have of data from 2022. As you can see – the large spikes still correspond and haven't changed much. We still do a lot with Kentucky history, Appalachia, politics and public policy, and UK. We're always going to prioritize our main collecting areas when we process. But what I'm really interested in with these graphs – is these smaller categories, which rise and fall in interesting ways from year to year. I find this is where the most meaningful analysis lives and where I can draw conclusions of what to process.
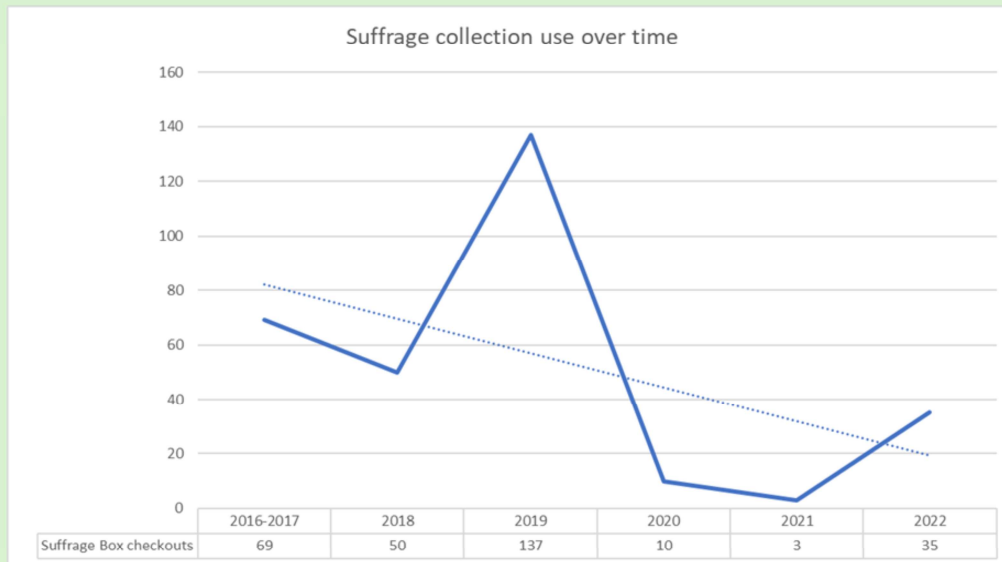
Programmatic Improvement

- Set processing priorities for new collections and the backlog
- Helps identify problem areas:
  - Legacy collections in need of better access
  - High-use collections without an online finding aid
- Helps identify collections for digitization
- Has implications for collection development
- Can be used for advocate for high use areas in need of support

I use this process to make some decisions to set the direction for the manuscripts processing program. The subject analysis helps me set priorities for backlog processing as well as for new material coming in. (Granted I use this analysis along with some of the considerations I discussed on the first slide – donor promises, responding to patron requests, etc.) This process has also helped identify legacy collections that are highly used and in need of better access as well as high-use collections that for whatever reason do not have an online collection guide. It can also help us decide whether to digitize a high-use collection for better access and better preservation (you know fewer people touching delicate things is always good). Looking at micro-trends in research at our own institution can also inform collection development decisions about whether or not to take a collection on a certain topic or how to aim our purchasing funds for the year. The data can also be used for internal advocacy, so if there is an area, program, etc. that has high use and low access I can use this data to back up requests for support.

Examples - Suffrage

Suffrage collection use over time

	2016-2017	2018	2019	2020	2021	2022
Suffrage Box checkouts	69	50	137	10	3	35

So now instead of doing a deep (and possible boring) dive into this body of data – I want to use my remaining time to talk about a few real-world examples of how we've used this data to make decisions. This first example is about identifying research trends to inform processing and digitization. This graph shows the rise and fall in use for suffrage related collections. Starting in 2016/2017 the in-person use of our collections about suffrage increased rapidly. This was of course due to an increase in scholarship about suffrage surrounding the various anniversaries of the 19th amendment in 1919-1920 (it was passed in 1919 and completely ratified by the states in 1920). The box checkouts shown in this graph peaked in 2019 and cratered out in 2021. I think if not for the pandemic the use in 2020 would still have been high. We worked hard in late 2018/early 2019 to process, improve description for collections dealing with women's suffrage. Additionally, we digitized some highly used collections that were starting to experience wear and tear from repeated use. These images came in handy when we were all sent home in 2020 during the pandemic.

Examples - Finding Aid Access Project

Collections checked out by online status 2016-2017

54% finding aid
15% collection-level
31% none
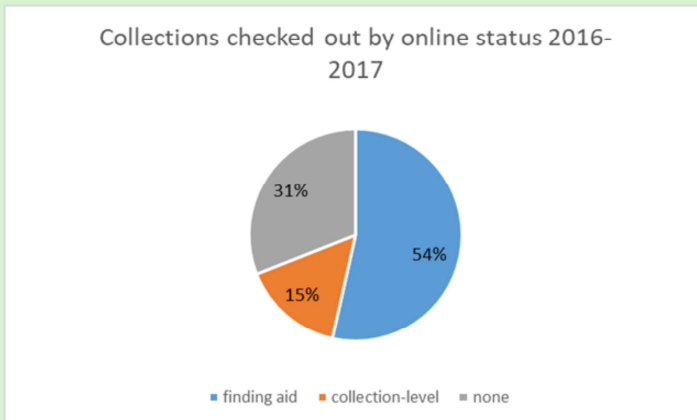
- From 2016-2018 - ⅓ of collections checked out were not online
- Represents a lot of work for research desk staff and subject specialist staff
- In 2019 ran a project to address this:
  - Identified 110 collections in this gap
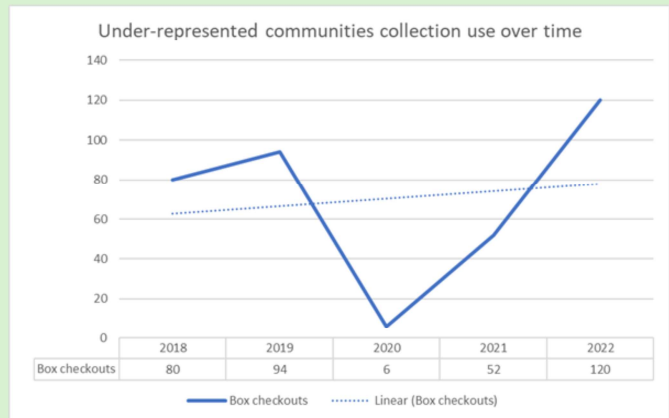 - 68 have gone online

So when I first did this process with the 2016-2017 data – I noted that a third of the collections checked out in our research room had no online access. Remember I mentioned how there are citations for backlog collections out there... We have to generate pdfs of collection guides in progress (if they exist) and send them to researchers, we have to assess these collections to see if they can even be served to patrons (are they stable? Are they in good enough boxes? Are they in any order at all?). Also, if they're not online – then patrons cannot help themselves by finding it. This represents a lot more work for our research room staff as well as our processing staff, and that work has to happen in the moment. In 2019 – I ran a project to identify some "abandoned" collection guides in our Manuscripts holdings, categorize them by difficulty to complete, resolve their problems, and put them online. I identified 110 collections in this gap. 68 have now gone online. (53 were categorized as easy; 9 medium; and 6 hard). We're still working on some of the ones identified as hard and medium, but we're definitely still making progress on them!
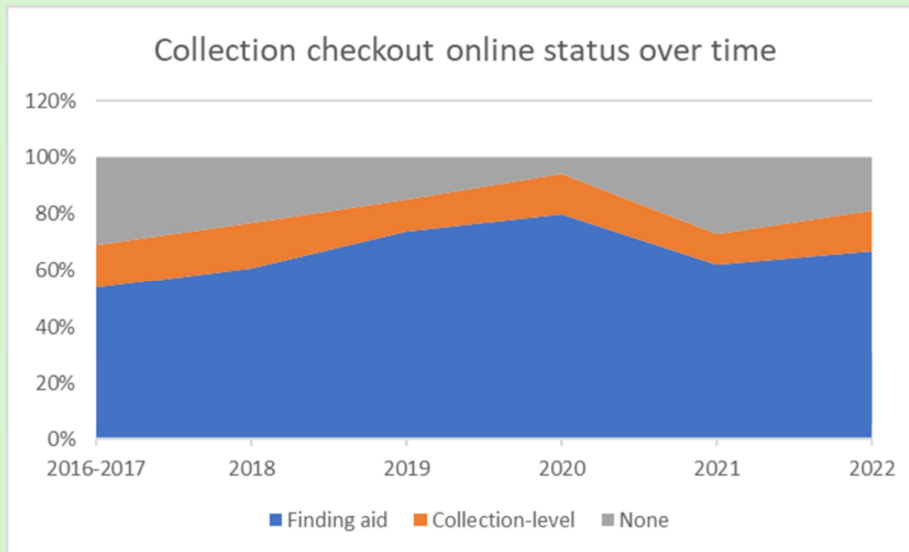
By 2020 you can see that these use numbers drastically improved. Use of collections in our research room that have no online access declined 15% from 2016-2019! Which is awesome. It's been climbing back up again – But that's just another opportunity to push some collections online!

Examples - Advocacy

Faculty papers collection use over time

	2018	2019	2020	2021	2022
Box checkouts	36	119	10	55	154

Under-represented communities collection use over time

	2018	2019	2020	2021	2022
Box checkouts	80	94	6	52	120

You can also use this data to advocate for certain types of collections! This graph shows the impact of faculty papers (something my institution is so-so on collecting). But they're actually used pretty heavily as long as the content is of high quality. I also have a graph here showing how often collections documenting under-represented communities are used. I'm happy to report that use is growing! This is a legitimate research area of interest. And we should continue to prioritize and process them because people want to use them! And...I had to put in these trend lines because of the pandemic...my numbers crater in 2020 and 2021! The pandemic ruined so many things, but it also ruined my dataset!

Examples - Bad Data Points

Checkouts by year EAD online 2022
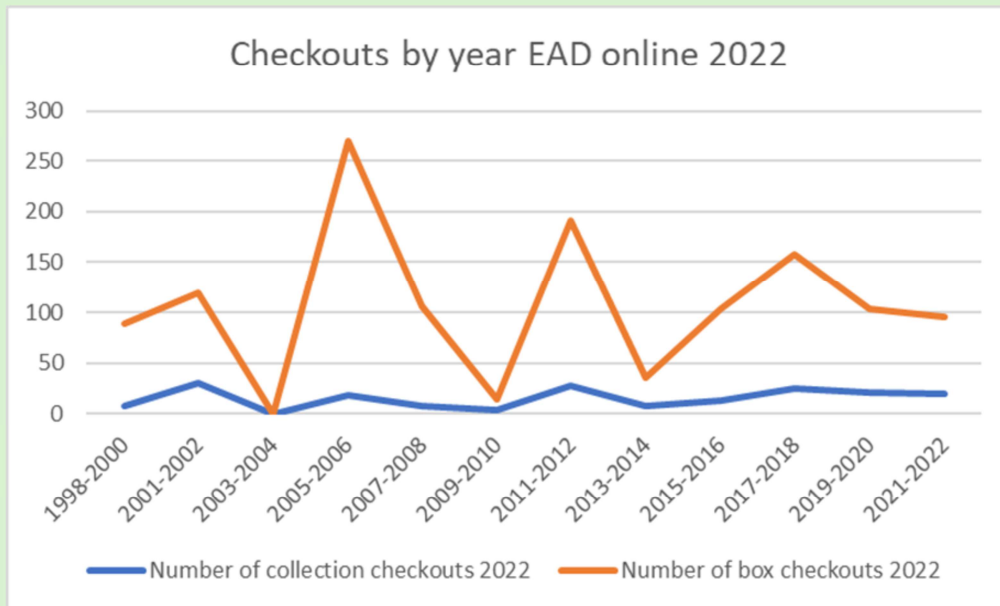
Now I want to take a minute to discuss failures. Not every data point I have investigated over the years....has panned out. I decided to look at the checkouts and numbers of collections used by when their finding aid went online. My results were not great. These spikes in 2005-2006 actually show when some of our biggest, most well-known, and highly used collections went online – like the frontier nursing service...this spike is when we had a project to put legacy finding aids online in 2011/2012. This downturn in 2009/2010 is when our new digital library was being built and we couldn't put anything online. I was hoping this data would show the longer something has been online the greater the use...but right now it just points out years we put a lot of finding aids or extra large finding aids online!

Examples - Bad Methods

Term	Count
kentucky	557
ky	192
states	170
united	163
lexington	159
history	156
county	97
john	94
war	88
family	85
clay	76
william	76
century	62
james	62
women	61
university	58
american	57
records	52
henry	46
1865	45
appalachian	43
1861	39
robert	39
correspondence	38
1945	37
african	37
world	37
20th	36
company	36
government	36
politics	36
civil	35
region	35
court	34

Another investigation that went nowhere...extracting and analyzing the subject headings from finding aids for collections used in our Research Room. This was a suggestion from someone who saw my 2018 presentation. I created an xslt to batch extract our library of congress subject terms from individual EAD. Then I ran the corpus through Voyant tools (which is a web based text analysis tool) and was extremely disappointed.

Examples - Bad Methods

Worst word cloud ever

Term	Count
kentucky	557
ky	192
states	170
united	163
lexington	159
history	156
county	97
john	94
war	88
family	85
clay	76
william	76
century	62
james	62
women	61
university	58
american	57
records	52
henry	46
1865	45
appalachian	43
1861	39
robert	39
correspondence	38
1945	37
african	37
world	37
20th	36
company	36
government	36
politics	36
civil	35
region	35
court	34

It created basically the worst wordcloud ever! 46 Henrys...62 James...170 states

Examples - Bad Methods

Term	Count	Length
shelby isaac 1750 1826 shelby james 1784 1848 sh	3	13
breckinridge robert j robert jefferson 1800 1871	2	12
alexander robert spreul crawford aitcheson 1819	2	10
church and state united states history 20th cent	2	8
marshall george s marshall silas norchutt joseph	2	8
aspects appalachian region coal mines and minir	3	7
united states history civil war 1861 1865	9	7
african american fraternal organizations kentuck	2	6
agricultural and mechanical college of kentucky	2	6
aspects kentucky coal mines and mining	2	6
coleman j winston john winston 1898	2	6
first african baptist church lexington ky	2	6
lincoln grant high school covington ky	3	6
african american baptists kentucky lexington	2	5
african american musicians kentucky lexington	2	5
baptist church fayette county ky	2	5
birney james gillespie 1792 185	2	5
black james dixon 1849 1938	2	5
bullock waller overton 1875 1953	2	5
church records and registers kentucky	2	5
fouse william henry 1868 1944	2	5
history civil war 1861 1865	6	5
jewell robert berry 1896 1986	2	5
kentucky lexington women college students	2	5
kinkead george blackburn 1849 1940	2	5
may andrew jackson 1875 1959	2	5
mcdowell henry c henry clay	2	5
morrow edwin porch 1877 1935	2	5
politics and government 1861 1865	2	5
siebert wilbur henry 1866 1961	3	5
u.s university of kentucky chapter	2	5
underwood warner lewis 1808 1872	2	5
united states history 20th century	4	5
united states politics and government	2	5

It's just as bad when it looked at the phrases rather than the individual words. The whole process was time intensive. And to draw any conclusions from it – I would need more programming skills and text mining knowledge than I currently have.

Conclusions

- Data is messy; you have to clean it up!
- Data cleanup can take a long time
- Look for data your institution may already be collecting
- Sometimes your data points mean nothing
- Use data can help point you in the right direction
- Results
  - More finding aids online!
  - More resources
  - Can give direction when faced with huge backlogs

A Suffragette Primer, 1911, from the Laura Clay papers

In conclusion I have some takeaways and reflections on my results. Data is messy – you've got to clean it up! Data cleanup can take a long time. Look for data your institution may already be collecting. Sometimes your data points mean nothing. But collection use data can help point you in the right direction! Because of this process we've put more finding aids online! Ones that are used! I can advocate for more resources in a high use area because I have numbers. Administrators love numbers. There's never a shortage of things to process or improve access to – but looking at the data can give you direction when faced with a huge backlog! And I'll repeat it again I find it so affirming to see that people are using our collections – it always makes me feel optimistic and positive.