Tracking and Analyzing Digital Records Processing Time at the University of Minnesota Archives and Special Collections

February 2025

Lara Friedman-Shedlov, Digital Records Archivist
Carol Kussmann, Digital Preservation Analyst
University of Minnesota Libraries

Abstract

The Archives and Special Collections Department of the University of Minnesota Libraries began tracking the time spent to ingest and process born digital archival materials in 2014. Following further refinements to the data collection process in 2020, we were able to accrue over four years worth of data documenting the time required for about a dozen different tasks in the workflow. Cross-referencing this time tracking data with other information about the accessions, such as storage size at various points in the workflow, has been instructive. This paper discusses observations we were (and were not) able to make based on this data, and the ways we've been able to use it to improve our workflow and acquire additional resources. In addition, the challenges of gathering and analyzing this type of data will be discussed and suggestions made for future work.

Background and overview of the project

In 2014, the University of Minnesota Libraries established a task force to explore and develop capacity of the Libraries to preserve and provide access to born digital archival material. As the organization began ingesting and processing this material, one of the first decisions made was to track time spent by personnel on various tasks in the workflow. The initial motivation for implementing this practice was to better understand the necessary human resource requirements. Like much collection management and stewardship work, the labor of digital records processing is less visible or recognized[1]. As noted in our 2015 report of this early work[2], industry standards for paper collections typically estimate processing time by cubic foot, but there were no similar standards for estimating processing time for electronic records[3]. A decade later, there are still no established metrics. Since it therefore goes largely unmeasured, this work is easily overshadowed by collection development and reference activity, often leaving it under-resourced.

Initially our time tracking efforts were relatively broad and informal. Data was captured at a macro level in a spreadsheet based on broad categories of tasks. The earliest data (2014-2015) documented how long it took to ingest each accession from the original media to a dedicated workstation[4]. We continued to track our time for the next 18 months (July 2015-December 2016) additionally capturing time spent on processing tasks as well as ingest tasks[5]. Gathering this data provided us with a general understanding of what types of accessions take the most time and allowed us to very roughly estimate how much time it would likely take to process our backlog. This data proved instrumental in making a compelling case for establishing a dedicated position in the department to focus on born digital archival records, as well as an ongoing Electronic Records Management Group (ERMG) to ensure the work begun by the task force could be systematized and operationalized.

After the formal establishment of the ERMG, its co-chairs considered whether to continue tracking time for born digital records work. Although uncertain about how we would use the

---

[1] As discussed by, among others, S. Wilson in “Implications of Archival Labor: If we want respect for our labor, we need to value it more” (Medium, 2016, https://medium.com/on-archivy/implications-of-archival-labor-b606d8d02014) and R. Clarke, K. Stanton, A. Grimm, and B. Zhang in "Invisible Labor, Invisible Value: Unpacking Traditional Assessment of Academic Library Value" (College & Research Libraries, Volume 83 Number 6 (2022) https://doi.org/10.5860/crl.83.6.926)

[2] Kussmann, Carol; Nelsen, R. Arvid; University of Minnesota. Electronic Records Task Force. (2015). Electronic Records Task Force Final Report. Retrieved from the University Digital Conservancy, https://hdl.handle.net/11299/174097.

[3] See the discussion in Shein, et. al, "Balancing the Art and Science of Archival Processing Metrics" Journal of Western Archives: Vol. 11: Iss. 1, Article 1. DOI: https://doi.org/10.26077/3c95-8ef9

[4] Table 1 on page 25 of the 2015 Electronic Records Task Force Final report (https://hdl.handle.net/11299/174097) summarizes the number of collections, amount of data, hours to process and the time spent ingesting content, providing an estimate on how long it might take to ingest the backlog at the time.

[5] Table 1 on page 15 of the Electronic Records Task Force Phase 2 Final Report (https://hdl.handle.net/11299/189543) summarizes the number of people who ingested or ingested and processed collections, the number of GB, and the time it took for that work. An average of GB/hour or hours/collection are provided.

information, we decided to continue to gather it, expanding the categories of information we collected by identifying more specific tasks to track and simplifying the collection process by using Google Forms to track them. The form improved consistency in how the data was categorized and pulled it all into one spreadsheet that would be easier to use later for analysis.

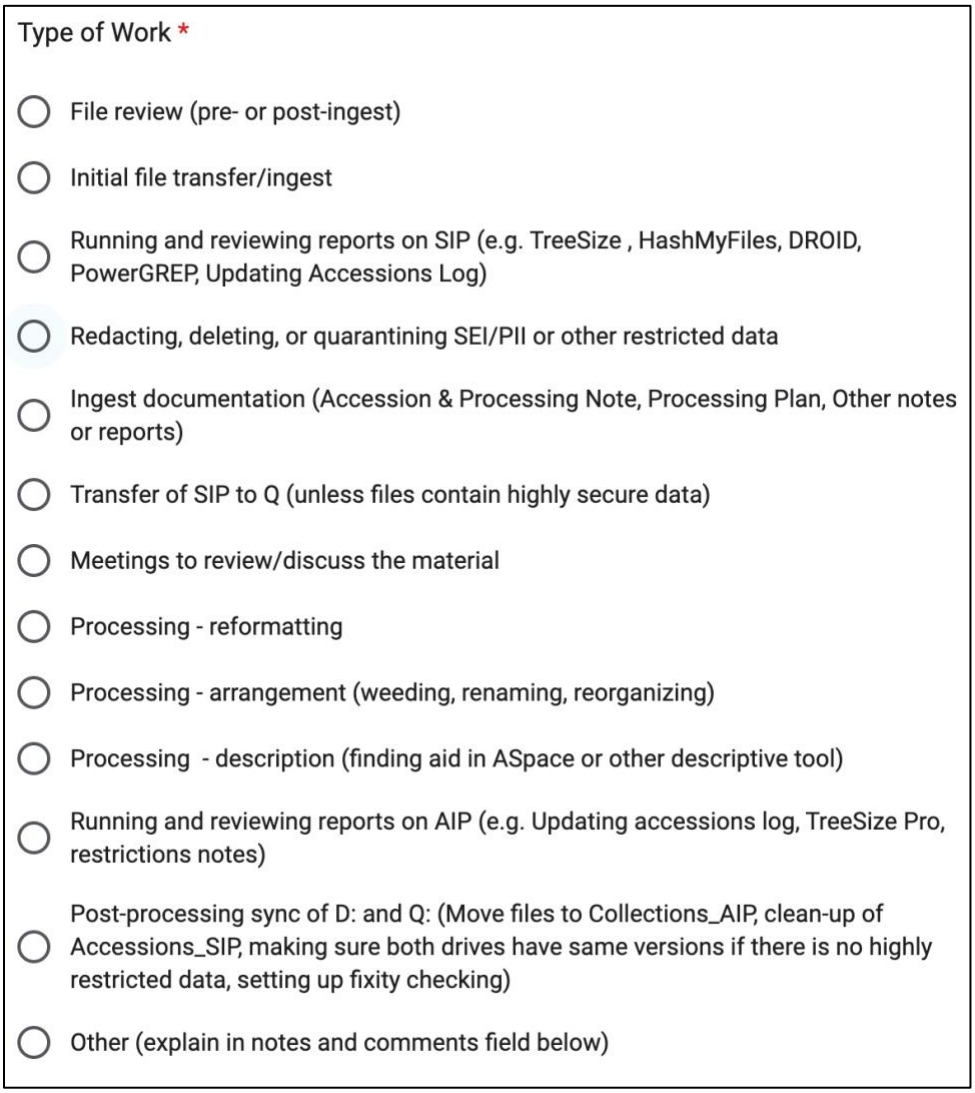

Figure 1: Screenshot of the 'types of work' on the form for which time is being tracked

In addition to collecting information about individual tasks or activities, the form also included fields to capture an identifier for the project being worked on (an accession or collection ID), as well as the number of minutes spent on the selected task. Staff were asked to fill out the form any time they worked on born digital records, submitting it as many times as necessary to document each task they may have undertaken and/or each chunk of time spent on the same task over the course of one or more days.[6]

[6] In practice this form was not filled out consistently by staff who infrequently did this work. Therefore, the vast majority of the data collected and analyzed represents work completed by the Digital Records Archivist.

Data analysis

Once we had several years of data tracked using this new collection tool, we wondered what patterns we might see if we analyzed this data. Some of the questions we thought we might be able to answer were:

- Is there a factor that could be used to estimate the time it takes to ingest and process a collection?
- Is there a correlation between the time it takes to process a collection and the size of the collection?
- Does the time it takes to complete certain tasks or projects change over time, and if so, is there evidence that we become more efficient? Did hiring a full time Digital Records Archivist to do this work actually result in greater productivity?

To help us with the analysis and visualization of our data, we utilized Tableau, a business intelligence and analytics software platform[7].

Data preparation and review

A total of 976 entries were collected using Google Forms during the period from January 2020 through March 2024. Prior to ingesting the data into Tableau, the data was reviewed and minor corrections were made to formatting as needed to ensure the consistency needed for linking related records. A decision was made to focus our analysis on the data entered by the Digital Records Archivist, as it was not only the vast majority of the data (819 of the 976 entries), it also represented a more consistent subset. These entries represented a total of 118 projects, 55 of which were completed (ingested and completely processed).

After data was ingested into Tableau, we were able to sort and filter the data as needed for various types of analysis. For example, we were able to distinguish between projects that were completed (fully processed) vs. projects that were just ingested or at some other intermediate stage. This allowed us to more accurately determine the amount of time that was tracked to bring projects to various points in the workflow. Tableau assisted greatly in our ability to identify and adjust the analysis to focus on specific factors needed to understand the data. We were also able to exclude outliers as needed to hone in on the most common data to determine if there were any trends.

Data limitations

Initial efforts to apply data analysis with Tableau to answer our questions quickly revealed gaps and other issues with the data collection process. We found that we were not necessarily gathering the correct data to answer these specific questions. Additionally we observed a lack of consistency in how tasks were recorded and at what level of accuracy they were recorded.

[7] Tableau (www.tableau.com) is a product of Salesforce. We would like to acknowledge our colleague Mark Engelmann, Business Intelligence Analyst for the University of Minnesota Libraries, who provided major assistance in using this software.

One of the issues revealed was that time tracking categories on our form were not as clear or unambiguous as we thought. For example, does a follow up email to a curator with an update on a project get tracked as a 'documentation' task or a 'meetings' task? We included an "other" category as a catch-all for tasks that weren't otherwise listed, but some tasks occasionally tracked in the "other" category could have been tracked in another existing category. For example, the task of preparing media for ingest (photographing and numbering original media) could be considered part of the "initial file transfer/ingest" task or the "file review" task, but was sometimes recorded in the "other" category.

Another issue with our data collection was that the amount of time for a task was sometimes recorded well after it was completed or was estimated, leading to perhaps inaccurate data. In addition, we weren't always performing just one discrete task during a time period but moving back and forth between multiple tasks, so the specific time recorded for individual tasks was not always exact. While the majority of this work was completed and recorded by the Digital Records Archivist, other Archives and Special Collections staff are frequently involved in parts of the workflow, including processing and description tasks. These colleagues were also requested to track their time. This contributed further to inconsistencies as non-dedicated staff have struggled to regularly and accurately log the work they completed.

There was also a lack of consistency of how tasks that run in the background (i.e., processes that are started by staff and analyzed when finished, but that do not require hands-on staff time while running, such as fixity checking or generating large reports) are documented. We didn't distinguish or have a way to indicate whether we were tracking only the 'active' amount of time a person needed to complete the task or the total 'machine' time that a process took.

We also observed the limitations of our practices for tracking the size of projects. We were initially influenced by the approach typical for analog collections, where we measure the size in terms of storage space, e.g. cubic feet or number of boxes, not the number of pages in a box. We paralleled that practice by consistently tracking the accession/collection by the number of bytes. Tracking the amount of space used is important to understand storage needs over time, and is often a more meaningful metric than the number of files for understanding the size of digital material. Some of our largest collections have relatively few files and take up a lot of space (e.g. video files), while other collections considerably smaller in storage size may have hundreds or thousands of files. In retrospect it might have been interesting or helpful to also track and analyze the change in number of files as an indicator of the amount of processing work done, especially since, unlike boxes of paper files, this data is automatically visible and easy to capture from digital file systems.

Because we sometimes merge multiple accessions together for processing or merge accruals into existing collections, calculating an accurate change in size can be difficult or can create outliers of data. If two accessions are ingested separately but later merged for processing, it is no longer possible to determine the amount of time each individual accession took from start to finish, or to distinguish the processed sizes of what were originally separate accessions. Similarly, when an accrual is initially ingested, it is tracked as a new accession. However, eventually the accrual is combined with an existing collection for final processing, at which point

it is no longer practical to track the size of the new accession separately from the size of the entire collection.

These shortcomings were not surprising considering we didn't have these specific research questions in mind when we originally started to collect data. Despite these caveats, analysis of the data we did collect was an exercise that yielded many worthwhile insights.

Findings and discussion

Some of the interesting things we were able to learn from the data included better understanding the overall size of projects being brought into the Libraries and confirming that the time spent processing materials is valuable, as overall it decreases the amount of required storage space. We were also able to confirm assumed ideas around the time it takes to process collections and better understand our workflows. The following sections describe these findings in more detail. Graphs below were created from Tableau and are provided with the intent to illustrate general trends or visualizations, not details for each data point[8].

Distribution of project size

Using Tableau, it was easier to see the distribution of project sizes. Once we plotted the total time spent compared to the size of each project, it was easy to see that the majority of the projects were tightly clustered in the lower left corner of Figure 2 below. Of the 111 projects tracked in the data, 41% (46) were under 1 GB and 80% (88) were under 100 GB. Only four projects were over 1TB.

[8]It was not always possible to configure and extract the graphs from Tableau in a way that showed the data we wanted to include and still be completely legible in the context of this report. In some cases we opted to include graphs for the sake of illustrating an overall trend, even when some of the text may be excessively small. We have provided explanations in the captions to assist in understanding relevant details in the figures.

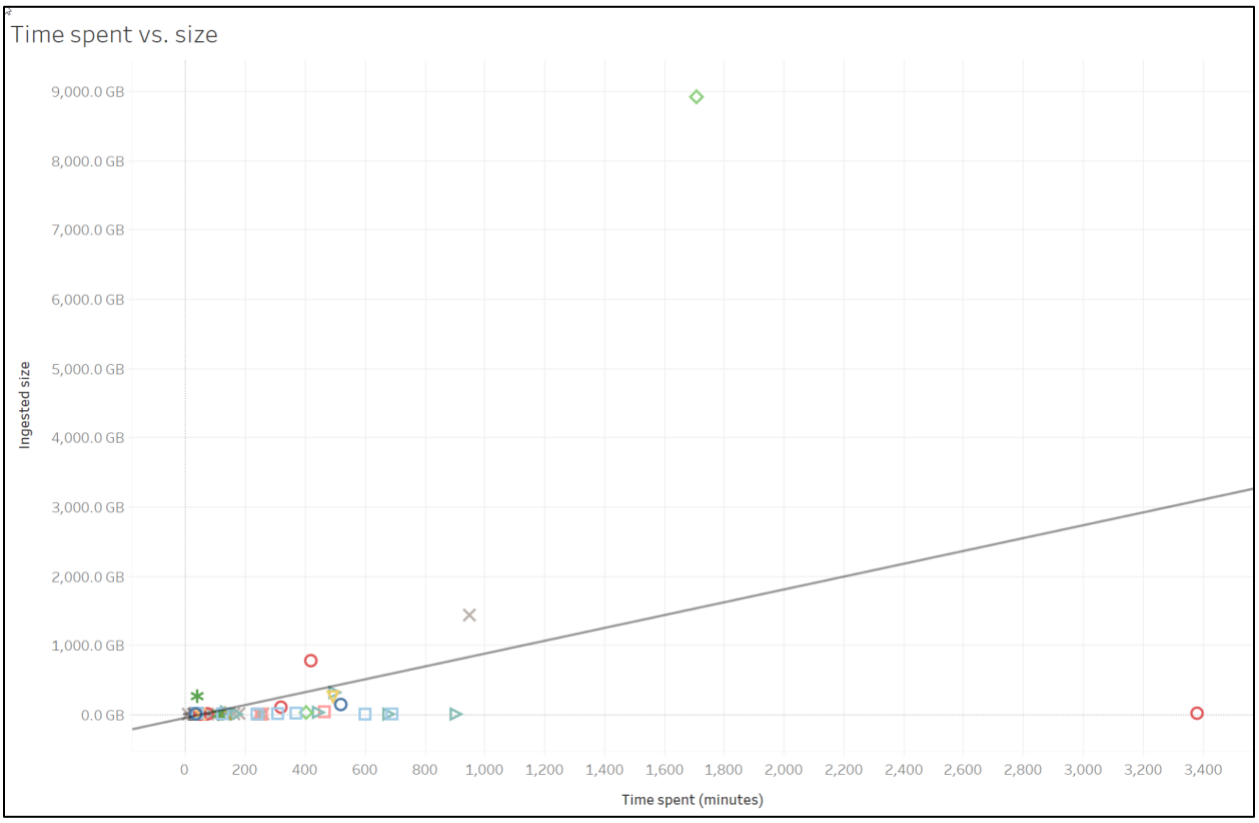

Figure 2: Shows that when all completed projects were graphed using the time they took to complete in relation to the size, they were heavily clustered in the area of the graph representing a size under 500 GB. The y axis of the graph is size at ingest from 0 to 9,000 GB in 1000 GB increments. The x axis shows time spent from 0 to 3,400 minutes, in increments of 200 minutes. The various symbols plotted on the graph represent individual projects.

Storage space and impact of processing

We saw an overall reduction in storage size of almost 80% after processing. This was a surprise and showed that reviewing materials and processing them, not just preserving everything that we get, is valuable to reducing the amount of storage space we are using. Storage space, while often said to be cheap, isn't free and isn't unlimited.

Completed Projects	Total Tasks Logged	Total time Spent in Minutes	Total Ingested Size (TB)	Total Processed Size (TB)	Change in Size from Ingest to Processed (TB)	Percent change in size
55	427	16,637	12.12 TB	2.44 TB	-9.69 TB	-79.9%

Table 1: Shows the number of projects and the size in TB before and after processing and the overall size reduction

While in general we saw a reduction in storage space being used after processing, we did find that in some cases the size actually increased during processing (Figure 3). Further investigation revealed this was usually due to unzipping container files such as zips or tar files.

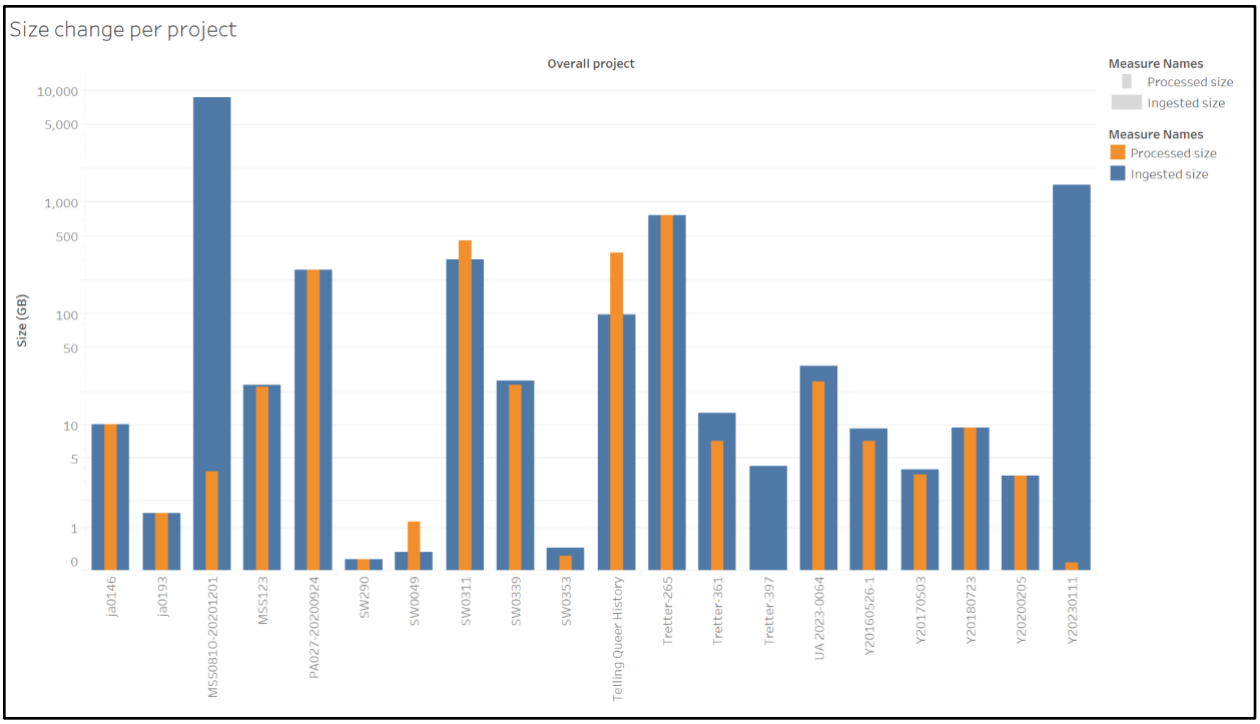

Figure 3: Graph showing a sampling of projects that demonstrate how projects changed in size from ingest through processing. The thick blue bar shows the ingested size. The thinner orange line shows the size after processing. These examples show that the ingest and processing size may remain the same; other times the processed size is less; and sometimes the processed size is larger than the ingested size (most often due to unzipping container files as part of processing). The y axis shows size from 0 to 10,000 GB with increments growing in orders of magnitude. The x axis shows the identifiers for various specific projects in this sample.

Excluding projects that grew in size due to file decompression, the data shows a general correlation between percent change in size (reduced) and time spent (Figure 4); in other words, spending more time processing generally results in savings in storage space. However the data visualization reveals many outliers, potentially providing case studies illustrating the pitfalls of predicting outcomes based on factors such as file size.

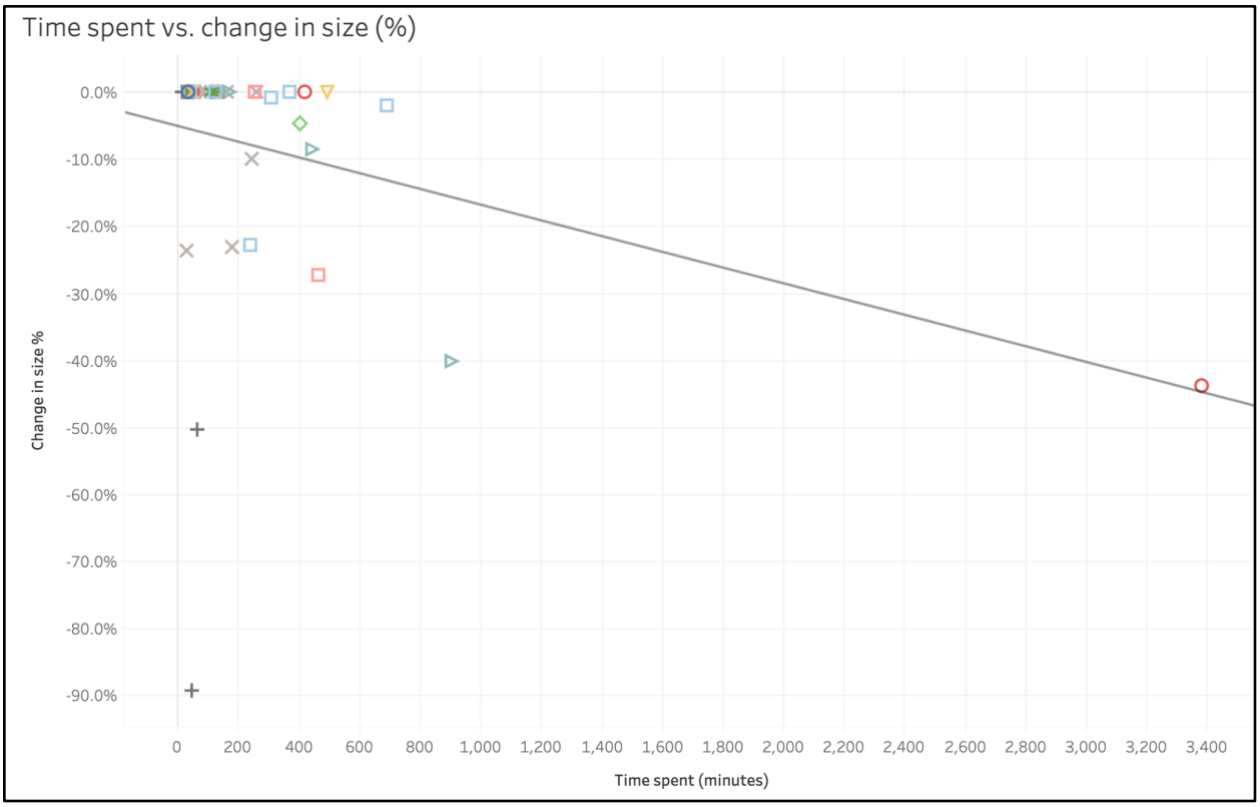

Figure 4: Graph with a trend line showing that the change in size of a project decreases as more time is spent working on it. The y axis shows percent change in size, ranging from 0 at the top to -90% at the bottom, in increments of 10. The x axis shows time spent from 0 to 3,400 minutes, in increments of 200 minutes. The various symbols plotted on the graph represent individual projects.

Relationship between project size and processing time

The data shows that, very generally speaking, the larger a project is in storage size the longer it takes to process. However, there are too many exceptions to suggest a standard, predictive metric for processing time. Some large collections take a small amount of time to review, while some small collections may take a lot of time. Figure 5 below highlights these outliers, where large collections take a short amount of time and small collections take a long amount of time to ingest and process.

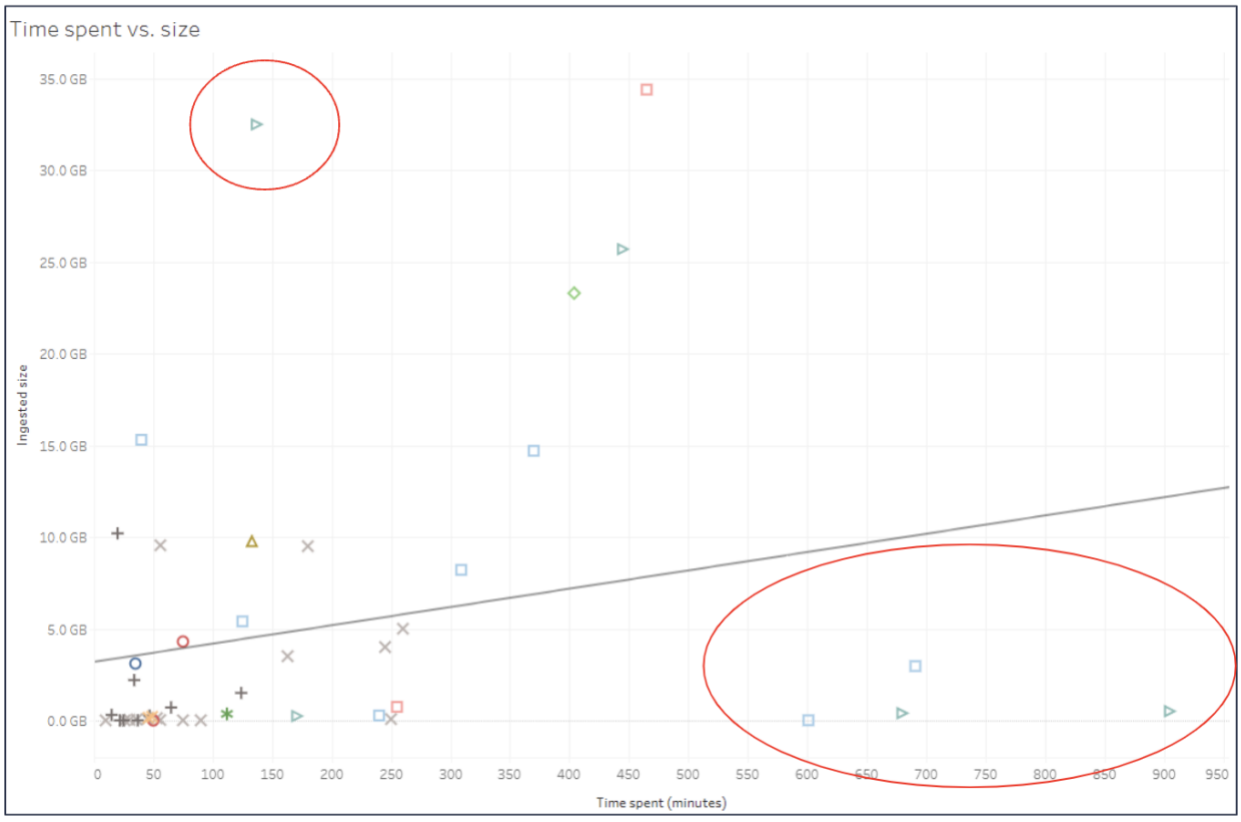

Figure 5: Graph representing the ingested size vs the time spent in minutes for 78 individual accessions/projects whose ingested size was under 35GB. The various symbols plotted on the graphs represent individual projects. The marks circled in the upper left and lower right show outliers in the data in this subset - indicating either a large collection taking a short amount of time or small collections taking a large amount of time. The y axis shows ingested size in increments of 5 GB. The y axis shows time spent from 0 to 950 minutes, in increments of 50 minutes.

The visualization above corroborated what we suspected and what others have already concluded: storage size is not a particularly useful factor to predict the time needed to process digital collections[9].

When we took a closer look at the outliers, we saw that, not too surprisingly, the type and number of media items (e.g. disks, drives, etc.) are clearly a significant factor influencing the amount of time required, particularly for initial transfer and ingest. However we did not typically track media type and number in a central place, and when we did, it wasn't captured in a consistent manner that lent itself to large-scale analysis. Using data gathered manually from other sources, we were able to identify five projects where we had data for media type and number of media items, all of which were within a relatively similar storage size range. Despite their similar size, the time required to ingest the materials varied greatly depending on the number and type of media present. One hundred floppy disks took 510 minutes and 75 floppy

[9] See, for example, Chela Scott Weber's discussion in "Time Estimation for Processing Born-Digital Collections" (2020) in Hanging Together, the OCLC Research Blog, https://hangingtogether.org/time-estimation-for-processing-born-digital-collections/

and optical disks took 517 minutes, whereas one flash drive of about the same size took just seven minutes (Table 2).

Project	Media Type/Count	Size (GB)	Time (min) (transfer/ingest)
N300	100 floppy disks	47	510
MSS095	75 floppy/optical disks	40	517
UA2023-0064	10 DVDs	34	390
Tretter-361	1 network drive	13	13
SWHA349	1 flash drive	32	7

Table 2: This table documents the size (GB) and transfer time (minutes) for five projects with different media types.

Efficiency

We hypothesized that an individual's efficiency would increase over time as they gained proficiency. We were initially surprised to see that the data did not strongly reflect that happening (Figure 6).

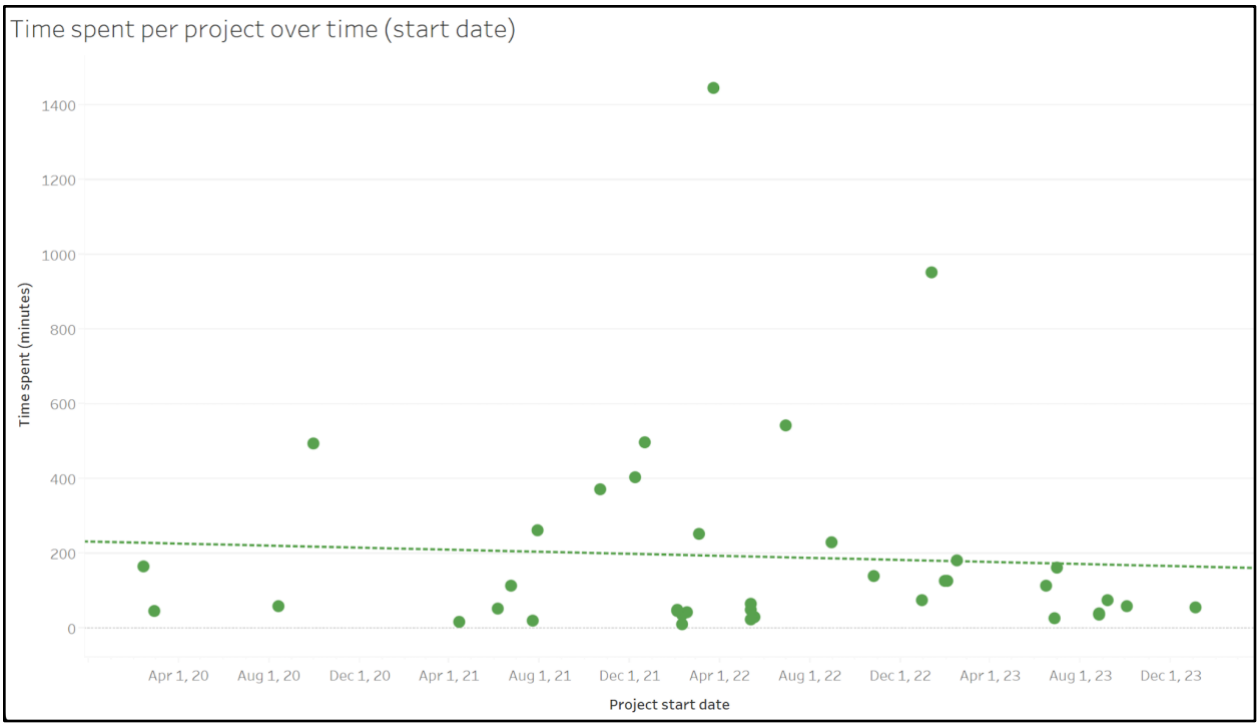

Figure 6: Trend line showing that, for projects started after January 2020 and completed before December 2023, the time it took to complete all tasks in the workflow went down only slightly over time. The y axis shows time spent from 0 to 1,400 minutes in increments of 200 minutes. The x axis shows start dates, with increments every 4 months.

Trying to understand why the data did not reflect more growth in efficiency led to a few insights into the data and our practices. One factor may be prioritization of work: easier, more straightforward projects were tackled first, and more challenging ones were addressed in recent years. So the fact that more recent projects took longer is at least in part related to their complexity, rather than the proficiency of the archivist. We also noted that while the time required per project did not decrease, the overall number of projects completed each year rose steadily over the course of the four years. This trend suggests that efficiency did indeed increase, perhaps due to an growing ability of the archivist to multitask.

- 2020 - 6 projects
- 2021- 14 projects
- 2022 - 40 projects
- 2023 - 53 projects

Passive (machine) vs. active time

Another key realization, as stated earlier, was that our time tracking did not consistently distinguish between ‘active’ time spent by the archivist and passive ‘machine’ time required for processes to complete. Certain tasks, especially generating reports on or moving very large directories of files, might run for hours. If we are looking at the data to help us determine personnel requirements, it might be more useful to focus mainly on the amount of active time

required to complete tasks. If on the other hand, we are interested in the total time required to get an accession through the workflow and ready for access, then also counting the machine time might be relevant. Creating a way to collect separate statistics on time that required a person's full attention as opposed to those that run for lengthy periods, largely unattended, is therefore desirable.

Conclusion and next steps

While the available data could not fully answer the questions we posed, it was nevertheless enlightening. In addition to the insights described above, the analysis facilitated review of our internal workflow and project status. The visualizations clearly revealed which collecting units within our department were making progress on processing digital materials and which were not. This information is helpful for us in considering how to prioritize our efforts going forward, including training and other types of support for unit staff.

We were also able to use time tracking data to justify the purchase of specialized equipment to make ingesting optical disks, one of the most time-consuming media to handle, more automated and efficient. As we continue to track ingest processes, we can see that the time to ingest disks using the new equipment is indeed less than the time it was taking to do this work manually (Table 3).

Transfer method	Average transfer time (in minutes) per disk
Manually (disk by disk)	8.5
Autoloader (batch processing)	3.8

Table 3: Data showing average time to ingest optical disks one-by-one using an optical disk drive, vs. using an autoloader that allows transferring files from batches of up to 100 disks at a time.

Based on the insights from this analysis, we are continuing to collect data with a number of refinements and additions to the process. These updates to the process include the following:

- Clarification of task categories on the form
- Systematic tracking of the number and types of media ingested
- Tracking of and differentiation between ‘machine’ and ‘active/person’ time for all tasks

While we still have no clear way to predict how long individual accessions will take to ingest and process, collecting this data has been valuable. This data not only documents the time this often invisible work takes, but shows that this work is worthwhile as it reduces the overall storage space needed and results in more organized records to share with our users.