Making News Usage Tangible: A Exploratory Analysis of Usage Patterns in the Texas Digital Newspaper Program

Mark E. Phillips
University Libraries, University of North Texas, Denton, Texas, United States
E-mail address: mark.phillips@unt.edu

Kristy K. Phillips
College of Information, University of North Texas, Denton, Texas, United States
E-mail address: kristyphillips@my.unt.edu

Ana Krahmer
University Libraries, University of North Texas Libraries, Denton, Texas, United States
E-mail address: ana.krahmer@unt.edu

w

Abstract:

University of North Texas Libraries have been collecting, digitizing, and making accessible newspapers from around the State of Texas since 2009 and using them to build the Texas Digital Newspaper Program (TDNP). As the largest collection on The Portal to Texas History, these 980,000 newspaper issues comprise nearly half of the Portal's 2 million publicly-available items. This paper presents the outcome of an exploratory analysis of usage data for the TDNP collection, spanning 2009 to present and representing over 48 million total use events.

Keywords: digital libraries, newspapers, usage, digitized newspapers, collection use data

1 INTRODUCTION

Since 2004, University of North Texas Libraries (UNT) has operated The Portal to Texas History (https://texashistory.unt.edu/), an online platform for providing access to resources collected and curated by institutions across the State of Texas. Contributing partners include other state and private universities, public libraries, historical societies, local archives, and even personal collections. These digital resources are preserved at UNT as part of its larger UNT Libraries Digital Collections, topping just over 4 million digital objects. On the Portal alone, 2 million

digital objects are openly accessible with metadata description, contributed by over 500 partners.

A major component of cultural heritage collections are newspapers. Texas, as with other states in the United States, has a rich history in newspaper publishing, spanning by some estimates over 2500 newspaper titles (Allen et al., 2022) since the early 1800s, when Texas was still part of Spain. This publishing history encompasses metropolitan daily newspapers, rural local newspapers, advertising and commercial papers, and school and neighborhood papers. Widespread newspaper publishing in Texas had its beginnings between 1813 and 1846, with eighty-six different titles being published in this span, with major increases in output beginning after statehood in December 1845, with steady title publication continuing until the present day (2022).

UNT began the Texas Digital Newspaper Program (TDNP) to support statewide organization, guidance, and standards for preserving and building access to this rich newspaper publishing history. Since then, TDNP has become a well-known avenue for collection, digitization, description, preservation, and access to the newspaper publishing output of Texas. The program hosts materials from three types of newspaper content streams: 1) newspapers digitized from physical pages, 2) newspapers digitized from microfilm, 3) and born-digital newspapers (Krahmer & Phillips, 2014). Since its beginnings in 2009, TDNP has preserved and provided access to 980,000 issues of newspapers, entailing just over 11 million newspaper pages, at the time of this writing. Hosted on The Portal to Texas History (https://texashistory.unt.edu/explore/collections/TDNP/), TDNP newspaper items appear alongside photographs, books, oral histories, and other media types, supportive of serendipitous discovery of connections between different types of materials and collections.

Starting in 2009, the Portal has collected item-level usage statistics for human interactions with the digital resources it hosts (https://texashistory.unt.edu/stats/), offering collection managers and administrators direct evidence of usage impact for these online collections. Additionally, the Portal aggregates item-level usage statistics so that collection-level information is available for individual partner institutions. This usage data is openly accessible and serves as a way for both UNT and partner institutions to report on the overall value of their digital resources, as many of the newspapers added to the Portal have been digitized through external grant funding, with funders interested in learning about the impact of collection access.

In addition, Portal usage data in relation to the Texas Digital Newspaper Program (https://texashistory.unt.edu/explore/collections/TDNP/stats/) provides an opportunity to answer some questions about newspaper usage in these large-scale digital collections. The research questions driving this analysis are:

> R1: To what extent do newspaper issues accumulate usage within the system over time compared to their overall representation within the collection?
>
> R2: How has the usage of newspaper issues changed over time within the collection, and what trends can we identify?

2 BACKGROUND

To help answer the question of how the newspapers for the Portal are being used, it is important to describe the methods that are currently used to generate usage data. Usage data in digital collections, and to a greater extent within online platforms in general, is a challenging area. Multiple variables directly affect the quality of reported data, and determining the impact, relevance, and value of usage data is critical to sustaining digital collections. Multiple community and consortial initiatives exist that offer standards of practice and tools that speak to the importance of usage data including COUNTER, RAMP, the OA Book Usage Data Trust, and the DSpace Content & Usage Analysis module.

These initiatives address in common the specific need of equitable usage evaluation for collections hosted over time, and their work sets standards of practice for repository managers worldwide. One of the earliest usage evaluation practice communities was the COUNTER Methods Project. Established in 2003, COUNTER (https://www.countermetrics.org/) was formed to address the problem presented by the many disparities between how online publishing platforms reported on usage. Similar to COUNTER but for repositories, the Repository Analytics and Metrics Portal (RAMP) (https://rampanalytics.wordpress.com/) was launched in 2017 and was intended to serve as a "prototype implementation of a new model for reporting institutional repository (IR) metrics" and "although . . . initially conceived as a resource for individual IR managers, the combination of cross-platform data aggregation and data persistence have resulted in a dataset that is unique in size and scope" (Wheeler & Arlistch, 2020). Founded in 2015, the OA Book Usage Data Trust (https://www.oabookusage.org/) was developed, "To champion strategies for the improved publication and management of open-access books by exchanging reliable usage data in a trusted, equitable, and community-governed way" (OA Book Usage Data Trust, 2024). A common repository platform, DSpace, offers functionality to calculate and log usage statistics as a core function of the software (Lyrasis, 2023). Atmire, the DSpace and DuraSpace service provider, extends this statistics function with their *Dspace Content & Usage Analysis* module, to serve "to measure and report on the usage, content, growth and therefore the value of your DSpace repository" (Atmire, 2024). These initiatives have in common three key goals:

1. To understand user engagement versus bot engagement with collections.
2. To evaluate collection-level interaction versus individual result interaction within a repository.
3. To support data persistence for long-term usage analysis (Wheeler & Arlistch, 2020).

Most helpful to our own research is the guidance these initiatives offer on standard methods for aggregating usage statistics, as well as for defining what constitutes a valid use within a framework.

3 METHODOLOGY

The methodology for this work builds on the implementation of usage statistics that already exist in The Portal to Texas History. This implementation is informed by the COUNTER Code of Practice (Counter Metrics, 2023) and its guidance for processing log files to calculate usage statistics. The goal of the CCP is to improve usage statistics and reporting from a wide range of organizations that provide access to online resources. These practices include how to identify bot

traffic; what HTTP status codes should be counted vs ignored; and how to handle double-click events from users. First, we follow the recommendation that organizations should define the steps that they take to identify and remove traffic from non-human agents that interact with a collection in automated ways. Referred to as "robots," "bots," or "spiders," these agents interact with digital collections to download content, generate an index, and/or collect metadata for a wide range of other services. These bots can be well behaved and advertise their identity as robots as part of their *User Agent* string when they request content. These well-behaved bots are easy to identify and remove from access logs that are later used to aggregate usage data.

One challenge in recent years has been the increase in the use of screen scrapers to programmatically access resources in Portal collections. When an automated script or robot is not clearly labeled as a bot in its *User Agent* string, manual identification is difficult for the organization – aka the content provider – gathering the usage information. This can happen because the creator of the bot does not want to advertise the fact that it is a bot, to prevent being blocked by content providers. Content providers may wish to limit bot traffic due to high-resource tax on the underlying infrastructure of a system, and they do this by blocking IP addresses that represent a common source of problematic bots. To get around blocks, badly-behaved bots often masquerade as standard *User Agents*, making them much more difficult to identify. Another class of usage that has become more prevalent in recent years is the use of virtual private networks (VPNs) to mask user identities based on IP addresses. One feature of some of these VPNs is the ability to derive each server request from a different IP amongst a pool of IP ranges. This activity creates challenges for many standard ways of aggregating usage by a single User Agent from an IP address with a resource.

Another recommendation that COUNTER offers is a distinction between interactions with the resource overall versus interactions that result directly in the consumption of content. This consumption, or viewing of, content is defined as downloading the material. This contrasts with broader interactions with the resource overall, in situations where just the descriptive metadata or information about citing the resource is viewed but not directly downloaded. When evaluating aggregated usage data, framing this distinction as *downloading* versus *interacting with* content can be problematic, especially in systems for presenting digitized newspaper content. These systems often provide access to their resources through a tiled image viewer that enables zoomed viewing, to support display of these physically large items on a wide range of screen sizes, including everything from mobile phones to large computer displays. Zoom-tiling is enabled without direct download of images, in contrast to PDF download of content, where one file represents an entire newspaper issue.

For this research, we created the dataset by processing log files using a series of steps informed by CCP practices:

1. First, we aggregated log files on a single machine for processing, pulling these from the multiple different application servers that provide access to the system, combining and sorting these log files by date to ensure all steps could be run on a consistent dataset.
2. Next, we identified and removed traffic from the "well-behaved bots," or bots that identified themselves, utilizing lists maintained by the COUNTER program for this step. After removing these bots, we performed further filtering of access logs to remove access

to resources that lacked information about where their request originated from in other content in the system. For example, when a user accesses a search result and navigates to an item in the system, their browser should provide a *Referer* header as part of that request. This *Referer* header is sent when a user navigates between pages of a newspaper issue. This header includes a URL from the previous page the user interacted with, which links to the current page or resource.

3. When the *Referer* header is missing, that request has a high likelihood of being a bot, as a common pattern for simple bots and scripts used for downloading content is to create lists of things to access at a later date but not track where they got the link information for future requests.

Once this data has been removed from the access logs, we have a much cleaner set to use for the following.

4. The next step is to aggregate the usage data for a given user, specific to those objects they interact with. Essentially, a user session is defined as a thirty-minute window of activity with a resource. This gathers all interactions with a given item, from downloading the landing page for the item, loading the image files, and browsing to different pages of the item, and groups them together based on a user's IP address.
5. After gathering these interactions, we discard some types of content because some calls for those types of content do not correlate to a user's direct use of the object. An example of these non-direct usage events include displays of thumbnails for an item on search result pages. This is an example of use events documented by the application which do not equal to an intentional use of that item by a person. Thus, we do not include these in our data aggregation.

After we execute the above actions on the dataset, what remains is a set of accesses from a user, based on their unique IP address, engaging with the individual objects within a thirty-minute window of activity. We can see this final set of accesses aggregated on a per-item level, and this represents what the Portal terms as an "item use" or "usage" (see Figure 1). These individual item uses are logged and then aggregated for the day to provide daily use information for items in the system.

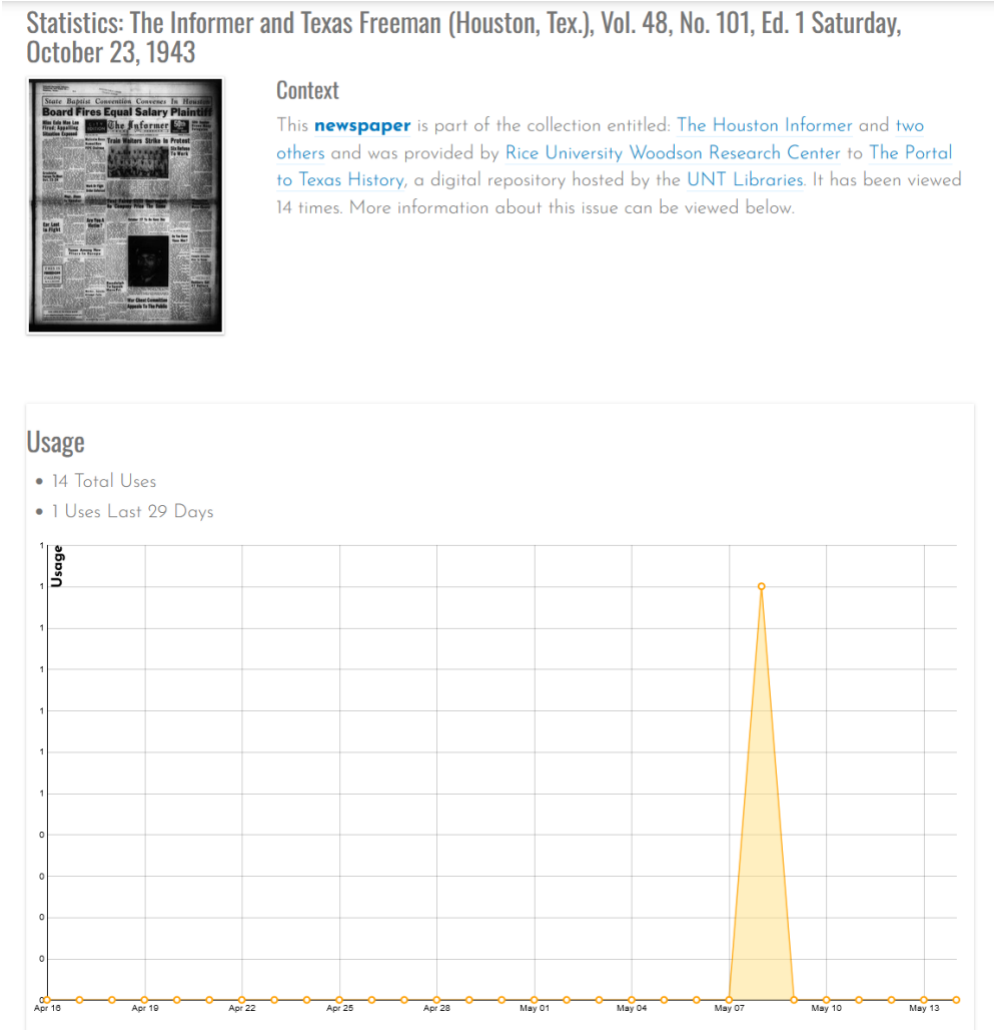

Figure 1: Showing an individual newspaper issue use record on The Portal to Texas History. (https://texashistory.unt.edu/ark:/67531/metapth1626771/stats/)

Generating the Dataset

We generated the dataset used in this paper by combining two sources of information from the underlying systems supporting the Portal to Texas History. First, the Archival Resource Key (ARK) identifiers for each of the newspaper issues in the Portal were identified in the metadata index. This index is a Solr search service that provides access to bibliographic information from the item's descriptive metadata. Specifically, we wanted a list of all newspaper issues and the date they were added to the system. The second source of information we pulled was the complete usage data for the system, aggregated by year. For example, the number of uses that a single newspaper issue in the system had in 2022. From these two combined sets, we were able to create one dataset that offers: 1) the ARK identifier, 2) the accession date for that item in the Portal, 3) the year from that accession date, and 3) the number of uses for each of the years from 2009 through 2023 (see Table 1).

meta_id	accession	year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
metapth41774	2009-02-02 14:49:57	2009	174	90	73	108	58	50	40	31	22	45	62	29	22	35	24
metapth41775	2009-02-02 18:56:00	2009	6	2	4	15	12	7	2	4	3	7	5	6	5	3	0
metapth41776	2009-02-02 18:54:41	2009	15	8	15	32	37	81	84	64	10	113	11	208	172	82	24
metapth41777	2009-02-02 18:53:43	2009	22	14	11	22	7	8	4	2	6	6	4	8	4	3	4
metapth41778	2009-02-02 18:53:40	2009	7	8	23	32	31	36	36	7	6	76	10	177	123	64	37
metapth41779	2009-02-02 18:51:49	2009	6	11	8	38	31	52	31	7	12	107	19	103	90	43	20
metapth41780	2009-02-02 18:50:34	2009	8	10	7	23	13	23	13	1	3	5	8	13	4	2	4
metapth41781	2009-02-02 18:50:16	2009	2	4	10	19	4	12	8	4	4	3	6	7	6	3	0
metapth41782	2009-02-02 18:48:58	2009	6	10	21	20	15	22	11	6	4	6	6	13	27	9	8
metapth41783	2009-02-02 18:47:06	2009	8	10	17	13	11	14	19	5	8	13	10	16	11	4	8

Table 1: Example rows from TDNP Usage Dataset presenting the number of uses per year for each item in the newspaper collection.

This dataset includes 946,107 rows that correspond to each of the newspaper issues loaded onto the Portal from January 1, 2009 through December 31, 2023. Notable "outlier" years were 2006, with 54 issues uploaded, and 2008, with 44 issues uploaded. Because we began aggregating usage statistics on June 22, 2009, 2009 data will include those 98 issues uploaded. As a result, we have removed 2006 and 2008 as outliers from this dataset and our subsequent analysis. Usage data for the dataset began on June 22, 2009, when we implemented usage statistics on the Portal. We have saved the dataset as both a comma-separated values (CSV) file and in the Parquet file format which were used for the following analysis.

The resulting dataset allows us to answer questions about how the TDNP collection has been used over time in the Portal. This first step in our analysis was to generate some overall statistics for each year so that we can better understand the growth and aggregate usage of items in the collection (see Table 2).

4 RESULTS

Year	Issues Added	% Total of Items	Use in Year	% of Use in Year	Use of Year	% of Use of Year
2009	7,262	0.77%	26,628	0.05%	1,210,154	2.48%
2010	44,789	4.77%	307,286	0.63%	6,604,281	13.56%
2011	32,652	3.48%	601,836	1.24%	5,984,721	12.29%
2012	30,933	3.29%	1,640,947	3.37%	4,504,513	9.25%
2013	49,533	5.28%	2,622,489	5.38%	3,973,591	8.16%
2014	93,637	9.97%	2,661,049	5.46%	7,337,588	15.06%
2015	56,959	6.07%	2,757,514	5.66%	4,837,713	9.93%
2016	90,398	9.63%	2,864,608	5.88%	5,354,337	10.99%
2017	62,389	6.65%	4,044,289	8.30%	2,878,900	5.91%
2018	117,840	12.55%	3,969,263	8.15%	2,826,017	5.80%
2019	73,985	7.88%	6,072,684	12.47%	1,317,687	2.70%
2020	59,218	6.31%	6,010,197	12.34%	742,120	1.52%
2021	75,334	8.02%	5,095,330	10.46%	527,868	1.08%
2022	67,247	7.16%	5,336,565	10.96%	422,327	0.87%
2023	83,933	8.94%	4,702,424	9.65%	191,282	0.39%
Total	946,107	100.00%	48,713,099	100.00%	48,713,099	100.00%

Table 2: Aggregate statistics by year for the items added, uses within a year, and uses of a year's items as part of the total uses.

An overview of the usage data for the TDNP collection in The Portal to Texas History is shown in Table 2. It starts with the number of newspaper issues added each year and is followed by the percentage that the issues uploaded that year contribute to the total number of newspaper issues currently available on the Portal. The next column shows the number of uses for a given year. As evident in the data, over time we can see a dramatic increase in the number of uses that the Portal receives each year. Interestingly, the most recent years indicate a tapering-off and decline in the number of uses. In the *Use of Year* column, we can see the number of uses that a year's worth of newspaper issues has accumulated in total. For example, the 7,262 newspaper issues uploaded in 2009 have been used a total of 1,210,154 times in the past fourteen years. To better view the overall impact of these numbers, Table 3 below displays them as a percentage of the total next to each column of the raw numbers.

Year	Issues	Mean	Std	Min	Median	Max	Mode	Freq. of Mode
2009	7,262	167	237	10	95	7,178	38	81
2010	44,789	147	253	6	87	30,692	41	452
2011	32,652	183	481	6	112	68,747	52	223
2012	30,933	146	303	0	90	32,042	47	332
2013	49,533	80	208	4	50	30,834	34	818
2014	93,637	78	183	0	45	29,856	25	1,520
2015	56,959	85	235	0	44	23,589	21	986
2016	90,398	59	128	0	31	13,686	19	1,947
2017	62,389	46	101	0	24	3,733	9	1,786
2018	117,840	24	57	0	15	8,811	7	4,824
2019	73,985	18	53	0	12	9,485	6	4,131
2020	59,218	13	32	0	8	5,467	4	4,458
2021	75,334	7	14	0	5	1,271	2	10,030
2022	67,247	6	7	0	5	645	2	8,639
2023	83,933	2	3	0	1	111	0	29,455

Table 3: Descriptive statistics for uses per year.

The data shown in Table 3 represents newspaper issues uploaded in a given year, along with their descriptive statistics based on the uses generated in the time they have been online. This includes the Mean number of usage trending down as the amount of time that the newspaper issue has been online diminishes. This holds true across all of the statistics including Median, Max, and Mode. It is

interesting to note that all newspaper issues uploaded in 2009, 2010, 2011, and 2013 have been used at least once since they have been online. The opposite is present in 2023 where 29,455 of the 83,933 issues (35%) have not yet been used.

Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
2009	26,618	74,701	80,862	133,988	123,196	114,079	77,800	58,679	63,740	80,211	90,875	91,413	65,637	68,379	59,976
2010	-	232,555	358,543	717,380	788,661	574,903	448,365	262,725	364,150	554,352	625,958	564,805	353,710	392,823	364,813
2011	-	-	162,376	585,563	807,220	667,351	504,728	375,633	435,514	392,671	498,064	499,065	367,778	372,086	316,155
2012	-	-	-	203,959	638,183	509,622	432,885	294,423	337,594	307,497	415,350	427,024	336,679	331,605	269,692
2013	-	-	-	-	265,179	531,245	433,977	329,968	371,914	300,151	442,365	418,462	311,829	325,748	242,753
2014	-	-	-	-	-	263,786	646,463	766,579	855,072	651,266	1,033,169	1,003,744	766,265	741,479	609,765
2015	-	-	-	-	-	-	213,225	518,040	638,571	482,396	742,117	716,030	574,115	509,101	444,118
2016	-	-	-	-	-	-	-	258,539	711,466	595,961	882,034	825,519	757,822	714,451	608,545
2017	-	-	-	-	-	-	-	-	266,212	340,225	537,346	485,260	460,078	436,257	353,522
2018	-	-	-	-	-	-	-	-	-	264,437	616,547	553,821	500,301	486,219	404,692
2019	-	-	-	-	-	-	-	-	-	-	188,859	292,872	278,945	302,609	254,957
2020	-	-	-	-	-	-	-	-	-	-	-	132,182	195,282	221,049	193,607
2021	-	-	-	-	-	-	-	-	-	-	-	-	126,889	215,563	185,416
2022	-	-	-	-	-	-	-	-	-	-	-	-	-	219,196	203,131
2023	-	-	-	-	-	-	-	-	-	-	-	-	-	-	191,282
Total	26,618	307,256	601,781	1,640,890	2,622,439	2,660,986	2,757,443	2,864,586	4,044,233	3,969,167	6,072,684	6,010,197	5,095,330	5,336,565	4,702,424

Table 4: Uses of a year of uploaded newspapers by year.

The data shown in Table 4 displays the usage data over time for the newspaper collection in the Texas Digital Newspaper Program on The Portal to Texas History. Looking at the first row of 2009, you can trace the usage for a specific year as you move across the

columns. For example, the 7,262 issues added in 2009 were used 26,628 times in 2009, 77,800 times in 2015, and 59,976 times in 2023.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
2009	100.00%	13.95%	8.57%	6.28%	4.40%	2.81%	2.30%	1.79%	1.55%	1.24%	1.10%	1.01%	0.91%	0.84%	0.77%
2010	-	86.05%	52.88%	38.73%	27.12%	17.31%	14.18%	11.03%	9.56%	7.64%	6.78%	6.22%	5.63%	5.19%	4.73%
2011	-	-	38.55%	28.24%	19.77%	12.62%	10.34%	8.04%	6.97%	5.57%	4.94%	4.54%	4.11%	3.79%	3.45%
2012	-	-	-	26.75%	18.73%	11.95%	9.80%	7.62%	6.60%	5.28%	4.68%	4.30%	3.89%	3.59%	3.27%
2013	-	-	-	-	29.99%	19.14%	15.69%	12.20%	10.57%	8.45%	7.50%	6.88%	6.23%	5.75%	5.24%
2014	-	-	-	-	-	36.18%	29.65%	23.05%	19.98%	15.97%	14.18%	13.01%	11.78%	10.86%	9.90%
2015	-	-	-	-	-	-	18.04%	14.02%	12.16%	9.71%	8.63%	7.92%	7.17%	6.61%	6.02%
2016	-	-	-	-	-	-	-	22.26%	19.29%	15.42%	13.69%	12.56%	11.37%	10.48%	9.55%
2017	-	-	-	-	-	-	-	-	13.32%	10.64%	9.45%	8.67%	7.85%	7.24%	6.59%
2018	-	-	-	-	-	-	-	-	-	20.10%	17.84%	16.38%	14.82%	13.67%	12.46%
2019	-	-	-	-	-	-	-	-	-	-	11.20%	10.28%	9.31%	8.58%	7.82%
2020	-	-	-	-	-	-	-	-	-	-	-	8.23%	7.45%	6.87%	6.26%
2021	-	-	-	-	-	-	-	-	-	-	-	-	9.48%	8.74%	7.96%
2022	-	-	-	-	-	-	-	-	-	-	-	-	-	7.80%	7.11%
2023	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8.87%

Table 5: Percent of items loaded by year.

A helpful way to consider numbers of items added per year is to think in terms of the percentage of the whole number of items available within that year. Table 5 displays items-added data as a percentage of the total newspaper issues on the Portal for that year. Thus, 2009 shows that 100% of the newspaper issues available in the Portal came from that year. In 2010, the issues added in 2009

now only represent 14% of the total online with the remaining 86% of issues being uploaded in 2010. This tracks onward until in 2023, where the content added in 2009 now only accounts for 0.77% of the total newspaper issues in the Portal.

	2009	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
2009	100.00%	24.31%	13.44%	8.17%	4.70%	4.29%	2.82%	2.05%	1.58%	2.02%	1.50%	1.52%	1.29%	1.28%	1.28%
2010	-	75.69%	59.58%	43.72%	30.07%	21.61%	16.26%	9.17%	9.01%	13.97%	10.31%	9.40%	6.94%	7.36%	7.76%
2011	-	-	26.98%	35.69%	30.78%	25.08%	18.30%	13.11%	10.77%	9.89%	8.20%	8.30%	7.22%	6.97%	6.73%
2012	-	-	-	12.43%	24.34%	19.15%	15.70%	10.28%	8.35%	7.75%	6.84%	7.10%	6.61%	6.21%	5.74%
2013	-	-	-	-	10.11%	19.96%	15.74%	11.52%	9.20%	7.56%	7.28%	6.96%	6.12%	6.10%	5.16%
2014	-	-	-	-	-	9.91%	23.44%	26.76%	21.14%	16.41%	17.01%	16.70%	15.04%	13.89%	12.97%
2015	-	-	-	-	-	-	7.73%	18.08%	15.79%	12.15%	12.22%	11.91%	11.27%	9.54%	9.44%
2016	-	-	-	-	-	-	-	9.03%	17.59%	15.01%	14.52%	13.74%	14.87%	13.39%	12.94%
2017	-	-	-	-	-	-	-	-	6.58%	8.57%	8.85%	8.07%	9.03%	8.17%	7.52%
2018	-	-	-	-	-	-	-	-	-	6.66%	10.15%	9.21%	9.82%	9.11%	8.61%
2019	-	-	-	-	-	-	-	-	-	-	3.11%	4.87%	5.47%	5.67%	5.42%
2020	-	-	-	-	-	-	-	-	-	-	-	2.20%	3.83%	4.14%	4.12%
2021	-	-	-	-	-	-	-	-	-	-	-	-	2.49%	4.04%	3.94%
2022	-	-	-	-	-	-	-	-	-	-	-	-	-	4.11%	4.32%
2023	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4.07%

Table 6. Percent of uses by year.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
2009	0.00	0.74	0.57	0.30	0.07	0.53	0.23	0.15	0.02	0.63	0.36	0.51	0.41	0.52	0.66
2010	-	-0.12	0.13	0.13	0.11	0.25	0.15	-0.17	-0.06	0.83	0.52	0.51	0.23	0.42	0.64
2011	-	-	-0.30	0.26	0.56	0.99	0.77	0.63	0.55	0.78	0.66	0.83	0.76	0.84	0.95
2012	-	-	-	-0.54	0.30	0.60	0.60	0.35	0.26	0.47	0.46	0.65	0.70	0.73	0.75
2013	-	-	-	-	-0.66	0.04	0.00	-0.06	-0.13	-0.10	-0.03	0.01	-0.02	0.06	-0.01
2014	-	-	-	-	-	-0.73	-0.21	0.16	0.06	0.03	0.20	0.28	0.28	0.28	0.31
2015	-	-	-	-	-	-	-0.57	0.29	0.30	0.25	0.42	0.51	0.57	0.44	0.57
2016	-	-	-	-	-	-	-	-0.59	-0.09	-0.03	0.06	0.09	0.31	0.28	0.35
2017	-	-	-	-	-	-	-	-	-0.51	-0.19	-0.06	-0.07	0.15	0.13	0.14
2018	-	-	-	-	-	-	-	-	-	-0.67	-0.43	-0.44	-0.34	-0.33	-0.31
2019	-	-	-	-	-	-	-	-	-	-	-0.72	-0.53	-0.41	-0.34	-0.31
2020	-	-	-	-	-	-	-	-	-	-	-	-0.73	-0.49	-0.40	-0.34
2021	-	-	-	-	-	-	-	-	-	-	-	-	-0.74	-0.54	-0.50
2022	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.47	-0.39
2023	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-0.54

Table 7. Heatmap of percent error between items percentage and usage percentage.

The same concept is represented in Table 6, but this time instead of the number of issues or items loaded, we are looking at the usage for the items loaded from a given year. Again looking at 2009, in that year 100% of the usage for the newspapers in the Portal came from content loaded in 2009. In 2010 that lowered to just 24% of the uses coming from content loaded in 2009. This works its way across the row to 2023 where just 1.28% of the total uses come from newspaper issues loaded in 2009.

The heatmap presented in Table 7 represents the percent error between data in Table 5 and Table 6. In this case, we used the data in Table 6 as the observed values and the data in Table 5 as the true values. The thinking for this approach is that if a year's newspaper issues represent 3% of the total collection, then if everything was equal, they would receive 3% of the usage in that year. When a year underperforms, that is when the percentage of newspapers is higher than the resulting usage for that year, the cell is highlighted in a green or blue color. When we have years that over perform such as the issues loaded in 2011 and used in 2013. In that case the 2011 issues in 2013 represented 19.77% of the issues in the TDNP yet made up 30.78% of the usage for that year.

Two related reasons explain a clear pattern in this heatmap. First, if all the newspaper collection usage originated from within The Portal to Texas History alone, we would expect to see usage per year closer to the number of items within a year. However, the nature of newspapers, their individual content, the titles to which the issues belong, and the years issues span, all impact usage. Taking these elements out of the equation allows us to examine other influencing factors on usage for the collection because usage unrelated to specific newspaper characteristics do not indicate users conducting searches within the system; rather, we must consider another source for usage: external search indexing. Search engines like Google, Bing, and DuckDuckGo heavily index newspaper issues in the Portal and present them to users external to the Portal, as part of their search results on those platforms. In the case of Google, these search results are presented in not only their main web search results but also in image search results.

As mentioned above, the length of time content has been online affects how thoroughly it is indexed in search engines, as search engines operate robots or spiders that crawl and index content for retrieval, and they generally do this with the goal of not overwhelming the systems they crawl, so crawling large websites like the Portal takes a very long time. The Table 7 heatmap helps us infer that content takes a while to become part of the search results, demonstrating the lowest percent error for content use for items within the same year they were loaded, because the items had not been fully crawled. Likewise as content has been available in the Portal for longer, it will be better represented within search results, as is visible in the top right of Table 7, showing loaded prior to and including 2012 as having a high usage-to-item percent error rate since 2018.

Another source of usage for Portal newspaper collections comes from outside systems, such as Wikipedia, Facebook, and X (formerly Twitter) linking into the Portal. Based on anecdotal observations of these platforms, some basic notes help indicate the usage origin. For instance, Wikipedia links into the Portal for newspaper content are typically directed to the newspaper title level and not at the item (issue) level. This Wikipedia traffic differs somewhat from that of social media platforms, where users commonly link directly into newspaper issues.

DISCUSSION

This paper and its underlying data present a first step in the analysis of usage data for newspaper collections hosted on The Portal to Texas History. Potential next steps include looking at how different time periods affect the usage of the issues in the collection. To do this, we would most likely begin by dividing the collection into decades based on each issue's publication date, and next generating statistics based on these decades across time.

Another useful inquiry we may pursue is how different geographical origins of news content impact usage, as this could help us understand whether the percentage of use matches the percentage of content from those geographical locations. This could be approached by generating statistics based on the counties that the newspaper issues were published in, which is information readily available within the issue-level metadata records. We would expect that, all other things being equal, the counties with a higher percentage of newspaper issues in the Portal would show higher usage trends. Finally, we may seek to employ this analysis at the title level. This would entail a large amount of data to analyze, examining over 1,700 titles currently available on the Portal (https://texashistory.unt.edu/explore/titles/types/newspaper/).

While this data has not provided any startling analysis beyond what we would generally expect, it does support theories we have long held about of how usage occurs across these large newspaper collections:

1. We have evidence to support our hypothesis that content being online longer leads to a greater number and percentage of uses over newer content.

2. Additionally, we now have research findings suggestive of anecdotal evidence we have experienced over the years that content that has been online longer is better indexed by search engines. This indexed content makes our collections more usable by a wider range of users than those that begin their search sessions directly with our platforms. This reinforces the importance of working to meet the changing requirements of the search engines so that we can make sure that we are represented in their indexes.

3. Finally, it is important to be clear about how our usage data is aggregated, cleared of non-human bot-like uses, and then tabulated and shareable with others. These are goals that usage evaluation practice groups have been striving to achieve for decades, specifically in the scholarly publishing arena. Organizations operating large newspaper collections may benefit from reviewing these requirements and making sure that they represent usage data that meets the needs of this content community.

In closing, the usage patterns of newspaper collections in large cultural heritage collections like those in The Portal to Texas History are long-term investments involving many institutions, over many years. How our users access these collections, and the usage patterns that we can identify about them, lend credence to the ongoing justification for dedicating resources toward sustaining this kind of collection in the long term.

REFERENCES

Allen, L., Sharpe, E., & Whitaker, J. (2022). *Newspapers.* Texas State Historical Association Online. https://www.tshaonline.org/handbook/entries/newspapers

Atmire. (2024). *Content & usage analysis.* Atmire DSpace solutions. https://www.atmire.com/modules/content-usage-analysis-for-dspace

Counter Metrics. (2023). *Counter Code of Practice*. https://www.countermetrics.org/about/strategy/

Krahmer, A. & Phillips, M. (2014). Texas newspaper PDF preservation: A Low-cost solution with tremendous value. *International Federation of Library Associations World Library & Information Congress, 2014.* https://digital.library.unt.edu/ark:/67531/metadc330323/m1/7/

Lyrasis. (2023). *SOLR Statistics DSpace 6 documentation.* LYRASIS Wiki. https://wiki.lyrasis.org/display/DSDOC6x/SOLR+Statistics

OA Book Usage Data Trust. (2024). *OA Book Usage Data Trust.* https://www.oabookusage.org/

Wheeler, J. & Arlitsch, K. (2020). *Repository Analytics and Metrics Portal (RAMP) Workflow Documentation and Data Definition*. https://digitalrepository.unm.edu/ulls_fsp/141