Submitted on: 22.11.2017

2016 Satellite meeting - *News, new roles & preservation advocacy: moving libraries into action*
10 – 11 August 2016
Lexington, Kentucky USA, USA

Wisconsin Model: Capturing Weekly Newspaper Websites

Ron Larson
Serials Resources Librarian
Wisconsin Historical Society
Ronald.Larson@wisconsinhistory.org

Abstract:

In 1845, members of the Wisconsin press led the charge to create a historical society and urged its fellow members to gather their respective newspapers each year and donate them to the society's library.

An editor wrote, "A full collection of Territorial papers, neatly bound and preserved in the State library, we hardly need suggest, would be of the utmost importance for future reference. Come then, brethren, let us arrange ourselves cheek by jowl in some vacant alcove, that we may tell the wise ones that come after us, what happened in our day."

This progressive way of thinking helped to establish the Wisconsin Historical Society and was the beginning of an impressive newspaper collection.

Today, publishers are not as altruistic as their predecessors when it comes to the preservation of newspapers. Local historical societies, public libraries, and the Wisconsin Historical Society, are leading the way in archiving original newspapers, microfilm or digital newspapers.

As for the newspapers' websites, no one in the state has archived or is currently archiving the online version of the various publications, creating a void of born digital material from the past decade.

The Wisconsin Historical Society has started a test project in hopes of reversing this trend. By using Archive-It, the Historical Society is crawling selected weekly newspaper websites once a week for four to six months. This test will provide an idea of the storage space required and the associated storage costs, allowing us to see whether it is a viable project.

An interested player in this test project is the Wisconsin Newspaper Association. Through their involvement, we are hoping publishers will be persuaded to assist with the costs, allowing the project to grow so that we may tell the wise ones that come after us, what happened in our day.

Keywords: Newspapers, Wisconsin, Websites, Archive-It, Wisconsin Historical Society

In 1845, members of the Wisconsin press led the charge to create a historical society and urged fellow members to gather their respective newspapers each year and donate them to the society's library. Chauncey C. Britt is credited with the first mention of creating a historical society in an article on Oct. 22, 1845, in his newspaper, the Mineral Point Democrat, where he called on his "brethren of the press to keep this ball in motion until the object is obtained."

The territorial legislature did not follow up with the idea of creating a historical society, so Britt soon focused his attention on the first constitutional convention that was meeting in Madison in October of 1846. More newspapers were adding steam to the movement. With so much support coming from the territorial newspapers, many of the leading dignitaries attending the convention felt compelled to create the historical society, nearly two years before Wisconsin became a state.

Many Wisconsin editors felt the Historical Society would be a perfect place to preserve their newspapers. One Wisconsin editor, John Y. Smith of the Wisconsin Argus in Madison, wrote, "A full collection of Territorial papers, neatly bound and preserved in the State library, we hardly need suggest, would be of the utmost importance for future reference. Come then, brethren, let us arrange ourselves cheek by jowl in some vacant alcove, that we may tell the wise ones that come after us, what happened in our day."

This progressive way of thinking helped to establish the Wisconsin Historical Society in 1846 and was the beginning of a very impressive newspaper collection, thanks to the newspaper editors who believed in preservation of the news back in the mid-19th century.

Today, publishers are not as altruistic as their predecessors when it comes to the preservation of newspapers. Instead of taking the lead in preserving their legacy, Wisconsin publishers today are looking to local historical societies, public libraries, or the Wisconsin Historical Society, to go it alone when archiving original state newspapers, microfilm or digital newspapers. Some publishers today even refuse to send their newspapers to us, citing they no longer handle mail subscriptions. They would consider mailing the papers to us, however, at a much, much higher subscription rate.

At the Wisconsin Historical Society, I believe there have been three pivotal moments in our history in regards to newspaper preservation. The first was in 1846, at the very beginning, taking the bound volumes of newspapers from the territorial publishers, and keeping them safe for future generations.

The second pivotal moment was in the early 1940s, when the Wisconsin Historical Society became one of the earliest advocates in the country to transfer newspapers to microfilm, ensuring that the old newspaper pages would be preserved for many years to come.

The third pivotal moment in our Wisconsin newspaper collection is today with digital technology. We have been archiving digital pages of over 200 state newspapers since 2005 through an arrangement with the state's press association, the Wisconsin Newspaper Association. And, last year, Wisconsin finally began participating in the National Digital Newspaper Program.

So, clearly, newspapers have been and continue to be of high importance at the Wisconsin Historical Society, especially Wisconsin community newspapers. Even though the Society has an extraordinary range of artefacts and information about North American history, Wisconsin history remains a strength and core to the collections, and newspapers play a vital role in that strength. In fact, newspapers are the most comprehensive record of Wisconsin communities at the Wisconsin Historical Society.

In spite of our tradition of preserving newspapers, we have not ventured into the arena of archiving newspaper websites. In fact, no one that we know of in the state of Wisconsin has archived or is currently archiving newspaper websites, which has created a void of born digital material from the past decade.

When contacted earlier this spring, the director of the Wisconsin Newspaper Association mentioned that her group has discussed archiving newspaper websites but didn't know where to start. And, a staff member at a leading Wisconsin daily said the topic of archiving the newspaper's website has been discussed, but there was little interest in using employee time or paying someone else to archive these pages.

For me, that last comment pretty much sums up the problem with archiving and preserving newspaper material today. Newspaper publishers are not interested in allocating money that will help with archiving – they want someone else to preserve their legacy. This is a far cry from the newspaper editors' attitudes in 1845.

But, more on that later. Let's get back to the idea of archiving Wisconsin newspaper websites.

Archiving newspaper websites is something I have been thinking about for a long time. In fact, back in April 2008, when I was the library director at the Wisconsin State Journal and The Capital Times newspapers, it became very apparent that the websites should be archived somehow. The impetus to this thinking began when the Capital Times, an evening newspaper that published Monday through Saturday, decided to publish a paper copy just once a week and rely on its website for the daily content.

We were archiving the individual articles from the website in our SAVE text archive but the look of the website, the layout, was not being preserved in any way. I continued to wrestle with this dilemma for several months but then I left the problem to someone else as I accepted a buyout in September of 2008.

So, here we are, eight years later and websites of Wisconsin newspapers continue to be disappearing daily and weekly into the ether. Finally, this spring, I accepted the fact that this problem of newspaper websites was squarely back on my shoulders.

The serials department at the Wisconsin Historical Society has been using Archive-It for the past five years to capture pdf's of newsletters, journals and a few newspapers across North America but we had not ventured into the realm of crawling and capturing entire websites. It seemed too large of a project and too costly.

This past spring, however, I decided that we should find out for sure what the cost would be and how involved the process would be for capturing the data using Archive-It.

We began the test project in early April by crawling selected weekly newspaper websites. The plan was to continue the crawls for four to six months. It was our hope that the test period would provide an idea of the amount of data captured and the associated costs, allowing us to see whether it would be a viable project.

As part of the test project, we selected five weekly titles, which are:

The Stoughton Courier Hub:

The Dodgeville Chronicle:

The Peshtigo Times:

The Dodge County Pioneer:

And **The Chetek Alert:**

We began the initial test crawl in early April, using the Dodgeville Chronicle as our guinea pig. I was not sure at all what the results of the crawl would be. I was somewhat expecting the data collected to be very large.

The test results from the Dodgeville Chronicle's first crawl showed we had captured 3.4 gigabytes of data from the newspaper's website. My initial thought was it wasn't a huge number, it didn't sound too high, but then I thought if every newspaper captured over three gigabytes, we would not be able to afford eventually archiving all weekly newspaper websites. If that one number was true for all weeklies, we would be capturing over 500 gigabytes for all weekly newspapers each week, or 26,500 gigabytes for the entire year, or 26 terabytes for the year. The problem with that number is our current Archive-It budget of roughly $10,000 is for just 1 terabyte of data for the entire year.

I felt it was necessary to expand the tests in order to verify if the Dodgeville number was typical or if it was high or low. We set up the weekly crawls for the five newspapers to run for 12 hours every Tuesday, with the Archive-It reports ready to view on Wednesday. All the data collected each week was hidden from the public view.

The results of our expanded crawls have been promising, showing that the weekly gathering of data is not as overwhelming as first feared. The weekly size of the websites captured range from a low of 55 megabytes for the Stoughton Courier-Hub to a high of 5.4 gigabytes for the Chetek Alert. And, after the initial crawls in April for each website, the data gathered for subsequent weeks was smaller across the board and, week-by-week, very consistent. When put together, the average weekly capture for the five newspapers since early April has been 1.24 gigabytes.

When multiplied out by the number of weekly newspapers that we possibly could crawl, which is 150, the average capture becomes 186 gigabytes each week, or 9,700 gigabytes per year, or not quite 10 terabytes, per year. That is a better number but still one that is on the large size for our annual budget.

My confidence in the number of 1.24 gigabytes per newspaper each week was reinforced when the results of a related study at the Wisconsin Historical Society confirmed

that the majority of weekly Wisconsin newspaper websites had basic content that was not updated often. The study was performed by two students who visited 150 newspaper websites to see how often and how much they were updated. And, it's quite possible once we begin looking more closely at these sites, that the data size per paper could go below the 1.24 gigabyte average.

The evaluation of the websites, though not at all scientific, leads me to believe that the 186 gigabytes each week or 10 terabytes each year is fairly accurate, if not a tad high. With that in mind, we will take a closer look at the 150 websites, running tests through Archive-It, and choosing newspaper titles to add to the weekly crawl.

Based on the numbers we have right now, it is very apparent that it is financially impossible for us to capture all 150 newspaper websites. Our budget allows for capturing 1 terabyte of data each year, which costs approximately $10,000. Capturing all 150 websites would produce in the neighbourhood of 10 terabytes of data every year, requiring a budget of $100,000 every year. We just are not able to come up with that amount of money.

I believe we have three options in how to proceed with this project. The first option is to crawl ten percent of the 150 websites, choosing 15 newspapers representing various geographical areas of the state. I believe the ten percent, based on the numbers we have gathered since April, would provide 1 terabyte of data that would be affordable in our budget. The criteria for choosing the 15 newspapers would be based on the amount of data, geographic setting and the quality of the website, reporting and information provided through the website. This careful analysis of the 150 websites would take us through the end of the upcoming semester.

The second option, once analyzing the 150 websites, is to focus on the sites that do not have a large amount of data. Three of the five test sites that we have been crawling since April are quite small, with the average weekly data capture for each site being 88 megabytes, 118 megabytes and 280 megabytes. During our fall analysis of the 150 websites, it will be interesting to find out if sixty percent of the all of the sites fall into this category of having a small amount of data. And, let's say the average weekly capture for these 90 websites is 160 megabytes per site, that would give us a total of 14 gigabytes each week for all 90 sites, or 728 gigabytes for the year, which would meet our annual budget of one terabyte.

The third option is to capture just the home page of every newspaper in the state, thirty dailies along with the 150 weeklies. This would not be my first choice but it would serve a purpose. If the primary goal is to capture the look and feel of the home page, the layout of the page to see what was important on that particular week or day, then capturing the home page will satisfy that goal. For the vast majority of weekly newspapers, there is very little web-only content, most of it is a repeat of what has appeared in the weekly newspaper. And, we are already archiving the digital pages of the weeklies so the news and feature stories are being preserved.

There are 150 possible websites that we could capture but we can only afford one terabyte worth of data. We will need to be selective in the selection of titles and weigh the pros and cons of the three options. Right now, I am leaning towards option 2, capturing the websites of 90 titles and perhaps, doing a little bit of option 3 by capturing the homepage of the remaining 60 titles.

There are a number of reasons we have been slow at looking at the process of crawling and capturing newspaper websites besides budget; there have been the issues of time, staffing and the priorities of other projects. I am the only full-time staff person in the serials department at the Wisconsin Historical Society. The rest of the staff in the department is made up of part-time limited-term employees and work-study students. Since 2009, the staff has always been in flux as the limited-term employees have normally been students from the University of Wisconsin School of Library and Information Studies, which meant they would leave once they graduated and got a permanent professional job.

Last year I changed my approach to limited-term employees and hired non-students. The hope is that they will stay longer than the usual two-year stint of library students. So far, so good. All three LTE's that I hired in early 2015 have stayed well past their one year anniversary.

One of my LTE's has an MLS and has become my primary Archive-It person. With her abilities and stability, I felt we could finally venture into the world of capturing newspaper web sites. Also, another person who has helped to make this project work is my supervisor who is very supportive of archiving born-digital newspaper material.

But, as we all know, nothing stays the same. My supervisor is retiring; his last day is this Friday. Who knows what things his replacement will want me to focus on once he or she begins making a priority list. And my Archive-It in-house expert, who normally works 16-hours a week, is currently on maternity leave. She plans to be back at work by the end of September. And, a third change that will be happening sometime within the next 15 months is my possible retirement.

But, my hope is my new boss will put archiving and preserving born digital newspaper material at or near the top of the list, my Archive-It expert will return from maternity leave raring to go and I will have time before retirement to put this project in motion where it will become part of the weekly routine without a whole lot of thought or effort.

But, besides the personnel and time commitment, adequate financing each year will be required for this project to work. As you all are aware, libraries and historical societies are not usually swimming in money. And, there are always numerous programs and projects that are competing for that same slice of the budget pie. Besides the serials department, there are other departments throughout the Wisconsin Historical Society using Archive-It to capture data for the various collections.

All these departments and projects at the Wisconsin Historical Society share the same Archive-It budget. The maximum amount of data we can capture all together is 1 terabyte which will present a problem as we move forward with capturing more newspaper websites. The bottom line is we could certainly use financial assistance with this project for it to work.

An interested player in this test project has been the Wisconsin Newspaper Association.

The Wisconsin Historical Society and the WNA already have a very strong relationship which began when we started archiving the digital newspaper pages that are produced by the WNA and its members. The unique relationship that has developed in Wisconsin has ensured the preservation and accessibility of Wisconsin daily and weekly digital newspaper pages. The collaboration between the Wisconsin Newspaper Association, the Wisconsin Historical Society and the Wisconsin Department of Public Instruction has provided an opportunity to

preserve the digital pages of Wisconsin newspapers and to make a searchable database of the pages available to all Wisconsin residents.

The searchable database is maintained by the WNA's vendor, Tecnavia, and is searchable through a portal, called BadgerLink, provided by the Department of Public Instruction. All Wisconsin residents have free access to the newspapers. The Department of Public Instruction pays WNA for the access to the database and the Wisconsin Historical Society pays WNA for duplicate copies of every page which is then archived. Our archive of the duplicate digital pages has no public access.

One caveat of the searchable database is there is a 90 day embargo, meaning the most recent page a searcher can find is three months old. Even with the embargo, Gannett has pulled all of its Wisconsin papers from the database in fear that the free searching will impact the revenue flow of the Gannett online archives. Gannett has 11 daily newspapers and a number of weeklies in Wisconsin. Even though the Gannett newspapers are not available through BadgerLink, we still have the digital pages in the Historical Society's dark archive.

The director of the WNA has been a strong advocate of working with the Wisconsin Historical Society since she came to her position as director in 2010. At a meeting in late spring of this year, she reiterated her support and was very encouraged with our current project of capturing newspaper websites. She is, however, looking into working with Tecnavia to see if they can crawl the newspaper websites themselves and add the information to their online collection available through BadgerLink.

Even if that is the case, the Wisconsin Historical Society will continue to work towards crawling and capturing the newspaper websites on our own. We will want to have a collection of the websites that we know will be preserved for the future.

If WNA and Tecnavia do go their own way and capture the websites on their own, the Wisconsin Historical Society will then need to pay for the terabyte of data captured each year. But, if Tecnavia is not able to capture the websites on their own due to technological issues or it is not cost effective, I would hope that WNA would consider assisting with the financing of our project.

The WNA director cannot make the decision by herself to finance our project. It is the decision of the WNA board, comprised of 16 newspaper publishers. And, when they are told they will not benefit monetarily from this project, I have my doubts they will allow the WNA to spend thousands of dollars annually for archiving newspaper websites.

Publishers are notoriously stingy. They want a return on their investment and, I'm afraid, preservation is not considered a return on their investment. Through WNA's involvement, we are hoping publishers will be persuaded to assist with the financing, allowing the project to grow. But, we all know about newspaper owners and publishers who are not always as excited or interested in newspaper preservation as you and I.

Newspaper publishers have been historically lax at preserving their collections, be it news clippings, photographs, bound issues and now the digital content. Publishers' focus has for the most part been on tomorrow's newspaper, not yesterday's, with their eyes on the bottom line.

Working in the newspaper industry for 30 years showed me on several occasions how difficult it was to persuade publishers of the value of preservation. If the project could not justify a return on investment, it rarely was given an okay. And, usually, preservation or archiving plans or projects could not be justified monetarily using ROI. There might be a great need for the project, but if it did not make money or save money, the publishers would not allow it.

So the question remains whether libraries and historical societies on their own can continue to bear the expensive brunt of preserving newspapers and providing access in the digital age while newspaper publishers continue to ignore their role.

Newspaper publishers want someone else to preserve their legacy and pay for it. This protocol is a continuation of a century and a half trend of historical societies and libraries bearing the costs of preserving a private industry's product for the public good.

Are publishers, or will they ever be, committed to long-term storage and the subsequent expenses of saving every web page? Based upon previous experiences during the past 40 to 50 years, when clip files and photo collections have been discarded, the trust that both current and future publishers will understand the historic value of the website archive is of concern.

There are those who will say that newspapers are having a difficult time financially and cannot afford to take part in the cost of preserving newspaper websites. Ad revenues have been on the decline for years, circulation has been dropping and newsrooms are shrinking to almost nothing. But, on the other hand, Gannett can afford to purchase the Milwaukee Journal-Sentinel and the Journal Media Group earlier this year for $280 million and make an offer of $864 million to acquire the Tribune Publishing Company.

But really, how much would it cost a publisher if the Wisconsin Historical Society captured and preserved his or her website? With 10 terabytes of data costing $100,000 to capture and store 150 websites, it would cost each publisher between $600 and $700 a year if the cost was split evenly between the 150 publishers. That seems more do-able than the Wisconsin Historical Society spending $100,000 each year, something I know will not happen.

The early editors and publishers of Wisconsin newspapers were the impetus to the formation of the Wisconsin Historical Society and the seed to the Society's newspaper collection that has grown into a rich and bountiful array of newspapers from all over North America, from 1704 up until today.

I would wish that the present day publishers would see the value of newspaper preservation at the same level as their ancestors in 1845 and actively advocate for the preservation of their respective websites. Perhaps a publisher somewhere in Wisconsin will promote the idea of digitally archiving their websites by stating…

Come then, colleagues, let us arrange ourselves side by side in some Internet cloud, that we may tell the wise ones that come after us, what happened in our day.

Thank you very much.