Journal of Western Archives

Volume 9 | Issue 1

2018

Web Archives for the Analog Archivist: Using Webpages Archived by the Internet Archive to Improve Processing and Description

Aleksandr Gelfand
New-York Historical Society, ag3566@nyu.edu

Follow this and additional works at: https://digitalcommons.usu.edu/westernarchives

Part of the Archival Science Commons

Recommended Citation
Gelfand, Aleksandr (2018) "Web Archives for the Analog Archivist: Using Webpages Archived by the Internet Archive to Improve Processing and Description," Journal of Western Archives: Vol. 9 : Iss. 1 , Article 6.
DOI: https://doi.org/10.26077/4944-b95b
Available at: https://digitalcommons.usu.edu/westernarchives/vol9/iss1/6

This Case Study is brought to you for free and open access by the Journals at DigitalCommons@USU. It has been accepted for inclusion in Journal of Western Archives by an authorized administrator of DigitalCommons@USU. For more information, please contact digitalcommons@usu.edu.
Web Archives for the Analog Archivist: Using Webpages Archived by the Internet Archive to Improve Processing and Description

Aleksandr Gelfand

ABSTRACT

Twenty years ago the Internet Archive was founded with the wide-ranging mission of providing universal access to all knowledge. In the two decades since, that organization has captured and made accessible over 150 billion websites. By incorporating the use of Internet Archive’s Wayback Machine into their workflows, archivists working primarily with analog records may enhance their ability in such tasks as the construction of a processing plan, the creation of more accurate historical descriptions for finding aids, and potentially be able to provide better reference services to their patrons. This essay will look at some of the ways this may be accomplished.

In 1993 the use of the Internet, a system of interconnected computer networks previously used mainly by the defense and education industries, was made more accessible with the introduction of Mosaic, the first popular web browser. Each subsequent year saw the quantity of information put online grow exponentially, from only 130 websites in 1993 to over a million by 1997, making its preservation an increasingly urgent task.

Founded in San Francisco in 1996, the Internet Archive set itself a wide-ranging mission of providing “universal access to all knowledge.” That October it began to


archive and preserve the World Wide Web, an effort that continues to this day. In the two decades since, the Internet Archive has captured and made accessible billions of websites, consistently working to improve the quality and scope of its captures.

Today an increasing number of archivists are engaging in preserving content found on the Web; many institutions have also established their own web archiving programs. However, there is frequently a separation between those engaged in web archiving and those doing more traditional processing, description, and reference. By incorporating the use of web archives into their workflows, archivists working primarily with analog records may be able to enhance such tasks as the constructing a processing plan, creating more accurate historical descriptions for finding aids, and providing better reference services to their patrons.

This essay will look at some of the ways that analog archivists, those working primarily with physical records, may be able to utilize the Wayback Machine, the Internet Archive’s searchable digital archive of captured online content, to inform processing decisions and to fill in the gaps when other sources of information are unavailable.

Using the Wayback Machine to Enhance a Processing Plan and Fill in the Gaps

As a project and processing archivist, I am familiar with the effort that goes into processing and describing a large and complex records collection. If a collection has arrived at the Archive properly boxed and organized by its creator or previous owner, the process of organization and description is a relatively simple task. However the ideal is often far from reality which requires me to impose my own order on the collection during the processing phrase, while at the same time gathering as much historical information as possible throughout the process. In cases such as this, the Internet Archive can be of immense value in providing important contextual information that serves to fill in the gaps and assist in processing and describing the records.


6. Although there are other organizations with large web archive collections that may be utilized for the same purpose, this essay will focus on the Internet Archive as the largest.
The Institute of Public Administration (IPA) Records Case

One such collection confronted me when 700 boxes arrived at the Archive. The records had belonged to the Institute of Public Administration (IPA), a pioneering institution whose efforts contributed to good governance and municipal reform. The boxes encompassed almost a century of materials, from 1906 to 2004, with many having been kept in storage for over a decade.

Based on a preliminary survey, the original order in a large number of the boxes had been disrupted and papers belonging to various unidentified individuals were discovered amongst the records. Composing a full history of the organization added to the challenge. The first few decades of the institution's history were well documented, both within their own records and in widely available published sources. During its heyday, many local and state governments throughout the United States had sought out the assistance of IPA, then called the Bureau of Municipal Research. However, by the end of the 20th century, when the institution had shrunk, information on its history was less readily available. The collection records reflecting that time period were fragmentary, with a paucity of available publications.

An initial search for IPA’s web presence yielded no results. However, a late 1990s memorandum, discovered within the records during processing, helped with its discovery. The office memo announced the launch of a redesigned website of IPA, giving its URL and leading me to a cache of much needed information via the Internet Archive.

The Wayback Machine indicated that the website was initially captured on December 30, 1996, two months after the Internet Archive first began crawling the Web, and was archived dozens of times during the last eight years of IPA’s existence (Figures 1 and 2). Frequently updated, the site featured important elements needed for description: organizational structure, recent accomplishments, and a detailed history of the organization. Some of the pertinent data discovered on the archived IPA webpages was regarding the organization’s later activities, information that was missing from the analog records and which revealed that IPA was much more active in its last years than previously thought. In addition, the archived websites were helpful in identifying the various staff members of IPA and the positions that they held within the organization over time. Thus the presence of personal papers of various individual throughout the collection began to make more sense, providing necessary context that led to an adjustment in the processing plan to accommodate these records.

The IPA website remained on the live Web for seven years after that organization disbanded, until 2011, when it was taken down and the URL purchased by a different company, making its contents inaccessible without the use of the Internet Archive.
The United Nations Compensation Commission Records Case

Although the use of web archiving may be of limited value for information on organizations whose history spans many decades or even centuries, it becomes an increasingly useful source for contextual information when dealing with organizations whose records date from the mid-1990s and later. These types of institutions may potentially have their entire online existence, from birth to death, captured by the Internet Archive, with the result that their evolution may be observed and described with a greater degree of accuracy.

One such case that I encountered was while processing the records of the United Nations Compensation Commission (UNCC). The Commission was established following the First Gulf War (1990-1991) with the mandate to process claims and pay compensation for losses suffered by individuals, businesses, and nations as a result of Iraq's invasion of Kuwait. This gargantuan, multiyear effort involved millions of claims and hundreds of experts and staff, making the process of describing the history of the Commission a challenging task and one where the Internet Archive became an indispensable resource.

At first, some detective work was necessary in order to locate the earliest possible version of the UNCC website. Based on the archived copies in the Internet Archive, the current website of the organization (www.uncc.ch) has only existed at its current location since 2000, leaving a lacuna for all of the 1990s. Making an educated guess that UNCC webpages may have lived on the site of its parent institution, as many organizations tended to do back then, I consulted the archived website versions of the United Nations Office in Geneva (UNOG), the institution where the UNCC headquarters are located. The first capture of that website occurred in December of 1996, just like with IPA, two months after the Internet Archive began archiving the Web. While that archived website yielded almost no information on the UNCC save for an early press release, the next capture, done in February of 1997, was more successful. One of the tabs on this newly updated site featured the U.N. Compensation Commission and directed the user to what was in all likelihood the first website created for the UNCC (Figure 3).

With almost twenty years of UNCC history available on the Internet Archive, it was now possible to sit down and look at this aggregate information at length. Captured over one hundred times, the website had undergone at least three major redesigns, with frequent informational updates. The later versions of the website, while full of information and aesthetically pleasing, tended to focus on the accomplishments of the Commission and its general history (Figure 4). The earlier websites, on the other hand, incorporated greater minutiae on the first years of the organization's existence and progress, elaborating on how UNCC was able to fulfill its mandate during the early 1990s (Figure 5). Had it not been for these earlier versions of the website, information that was invaluable in compiling a historical note for the finding aid would have been left out.

Figure 4. Current iteration of the UNCC website, captured on March 25, 2018, https://www.uncc.ch/.
General Guidelines and Recommended Workflow

When processing a collection, I recommend integrating the use of the Internet Archive as early on in the process as possible. If the lifespan of the records’ creator coincides with the Internet Age (mid-1990s and later), an archivist may gain a broader understanding of the history of the organization or individual by reviewing the various iterations of their archived pages, which may immediately aid in the creation of the processing plan. This source of information may then be referred to throughout processing in cases where additional questions may come up.

As a first step, a basic search of the live Web will have to be made. The website may remain available even if it is no longer being updated or the organization has ceased to function. If found, the archivist may enter the website URL into the Wayback Machine and browse past versions of the site. If the site is no longer available, it may still be possible to find it without knowing the URL beforehand. An educated guess entered into the Wayback Machine, using different spellings or abbreviations may produce successful results.7

7. Making a successful guess may be complicated by the growth in the number of domains in use.
Early websites may sometimes be found under the parent institution, as the case with UNCC demonstrates, or with the employing institution for an individual. If the search does not yield any results, it is always possible that the website may be uncovered while processing the physical records, as in the case of the IPA records.

Web Archives for Reference Services

The Internet Archive may also be utilized to aid archivists when trying to assist patrons by revealing materials that have been lost or otherwise unavailable. In one such case, Washington Square News, the independent student newspaper of New York University, had been storing old issues in their local digital storage until a glitch wiped out a large portion of its archive. Having no backup, the newspaper turned to New York University Archives hoping to reassemble their collection via physical copies kept at that repository. Unfortunately, even with the available analog issues, a gap of more than two years remained. A quick search of the archived pages of the student newspaper via the Wayback Machine revealed that some of the missing issues and articles existed within archived pages and were available for perusal. By assembling a list of archived versions of the newspaper’s website, consisting of links to the discovered missing issues, the student newspaper was able to decrease the gap in their collection and at the same time provide another resource for future researchers who may request to consult back issues of the newspaper.

Mining the Internet Archive for information when these types of inquiries come up is free and does not require the archivist to do any of their own web archiving. A thematic list of archived websites may be compiled and stored in either a text or spreadsheet document for future reference. I created one such reference collection by mining the archived pages of a university’s website, compiling a list of student clubs and their websites from the 1990s and saving them in alphabetical order in an Excel spreadsheet.

If a research request is related to events from mid-1990s and later, archived webpages should be one of the first sources of information consulted. Some of the more common questions may readily be answered by reviewing old press releases, image galleries from public events, institutional publications, reports, and meeting minutes found within archived pages.

Additional Ways to Incorporate the Use of the Internet Archive

Websites containing unique materials not found amongst analog records, or those that the archivist thinks would be helpful for future research, should be considered for possible incorporation into a completed finding aid. Describing Archives: A Content Standard (DACS) gives the option of including a Related Archival Material Element (6.3) that may be utilized in such cases. Many of the publicly available webpages have been captured and are readily found in the Internet Archive. Besides captured pages being easily linked, the Internet Archive provides a free
service where users can capture individual pages to be used for citation purposes, which can be utilized for pages not previously captured (Figure 6).

Figure 6. Internet Archive's Wayback Machine, captured on March 25, 2018, https://archive.org/web/.

Conclusion

Today the Internet Archive is by far the largest institution engaged in preserving the Web. A large and diverse community of users make its existence more sustainable and increases the chances of its longevity. The nonprofit and educational mission of the Internet Archive makes it highly unlikely that the organization would ever start charging for its services, making the types of uses mentioned in this essay a sustainable proposition.

By incorporating the use of the Internet Archive into everyday workflows, archivists stand to benefit not only by becoming part of a growing community but by creating more accurate finding aids and by being able to provide greater assistance to archive users. The type of Internet Archive usage advocated requires minimal investment of resources and the benefits far outweigh the costs. With the passage of time, the online presence of individuals and organizations, whose records will eventually end up in archival repositories, is only going to increase, making pertinent not only their preservation but a need of their consultation if archivists hope to effectively perform their functions in the 21st century.