



DIGITAL LIBRARY OF THE CARIBBEAN

www.dloc.com

Digital Library of the Caribbean
Handbook Series

Introduction to ***WEB ARCHIVING***

by Matthew Kruse

Introduction To Web Archiving: A Handbook of the Digital Library of the Caribbean

Author: Matthew Kruse

Editors: Tania Ríos Marrero and Perry Collins

Copyeditor: Katherine Nguyen

Graphic Designers: Tracy E. MacKay-Ratliff and Ella Terran

© 2024 Matthew Kruse

Published by LibraryPress@UF

Acknowledgements

This guide is part of the Digital Library of the Caribbean Handbook Series.

The project was made possible in part by the Andrew W. Mellon Foundation.

The grant is administered through the University of Florida George A.

Smathers Libraries.



This work is licensed under a Creative Commons Attribution-Noncommercial 4.0 International License (CC BY-NC 4.0). To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc/4.0/>. You are free to copy, distribute, and transmit this work for noncommercial purposes if you credit the author and **the Digital Library of the Caribbean**.



Suggested citation: Kruse, Matthew. *Introduction To Web Archiving: A Handbook of the Digital Library of the Caribbean* (2024), dloc.com/AA00105780/00001

Names: Kruse, Matthew, author. Ríos Marrero, Tania, editor. Collins, Perry, editor. | George A. Smathers Libraries, publisher.

Title: Introduction To web archiving: a handbook of the Digital Library of the Caribbean / by Matthew Kruse ; editors: Tania Ríos Marrero and Perry Collins.

Description: Gainesville, FL : Library Press @ UF, 2024 | Series: Digital Library of the Caribbean handbook series | Summary: Introduction to Web Archiving is designed to provide a brief introduction to web archiving. Included is an overview of key concepts and terms associated with the practice; considerations and strategies for planning and implementing web archiving and social media archiving projects; and a comparison of common tools, resources, and software. Also included is a list of resources for further exploration and learning. Identifiers: ISBN 9781944455330 (digital version)

Subjects: LCSH: Web archiving—Handbooks. | Libraries—Special collections—Web sites. | Libraries—Special collections—Computer network resources. | Web archives—Handbooks.

Classification: LCC Z701.3.W43 K78 2024

About the Digital Library of the Caribbean

The Digital Library of the Caribbean (dLOC, www.dloc.com) preserves and provides open access to cultural, historical, scientific, and research materials from and about the Caribbean.

dLOC was established in 2004 with nine founding partner institutions. Twenty years later, dLOC has grown to over ninety partners and over forty associate partners based in the Caribbean, Canada, Europe, and the United States. Partner institutions are the core of dLOC. They contribute their collections, share governance over the organization, and participate in dLOC's evolution and success.

dLOC operational hubs are at the University of Florida, which leads on the technical and administrative infrastructure, and Florida International University, which leads on outreach in South Florida and collaborates on administrative tasks.

dLOC uses a post-custodial model of archival practice. dLOC and its operational hubs make no claims of ownership to any of the material available on www.dloc.com. Partners and associate partners retain all rights to their materials.

More than a digital repository alone, dLOC serves as a hub for teaching and research initiatives and as a network in support of partners, scholars, educators, students, and the public.

Table of Contents

Purpose of this Guide	01
Introduction	03
What is Web Archiving?	04
Brief History of Web Archiving	05
Web Archiving in the Field	06
Planning a Web Archiving Project or Program	09
Scope of the Project and Selection Process	10
Copyright	12
Ethical Considerations	13
Determining Levels of Access	14
Making the Case for Web Archiving	15
Tools Review and Comparison	17
Key Concepts and Terms	18
Considerations for Scope and Sustainability	20
Select Tools for Traditional Web Archiving	22
Select Tools for Social Media Archiving	27

About the Author

Matt Kruse is the head of collection services at the Department of Special and Area Studies Collections (SASC) at the George A. Smathers Libraries at the University of Florida. He implemented and manages SASC's primary web archiving software, Archive-It. He previously worked as a processing archivist at the University of Florida and as the archivist for the Connie Mack III Political Papers. Prior to that he worked at the College of Charleston and Montana Historical Society (State Archives). He earned his Master's of Library Science from Indiana University at Bloomington.

- Dowling, Stephen. "Why There's so Little Left of the Early Internet." BBC. February 28, 2022. www.bbc.com/future/article/20190401-why-theres-so-little-left-of-the-early-internet
- Duncan, Sumitra, and Amye McCarther. "Sustainable Web Archiving at Scale: An Introduction." Digital Preservation Outreach and Education Network. 2021. https://www.dpoenetwork/wp-content/uploads/2021/07/DPOE-N-Slides_Day-1_June.pdf and https://www.dpoenetwork/wp-content/uploads/2021/07/DPOE-N-Slides-Day-2_June.pdf
- Garrett, Jeffrey. An Evaluation of Web Archiving Programs in the U.S. Relevant to International and Area Studies: The Example of Latin America and the Caribbean. Center for Research Libraries Global Resources Network. 2019. <https://www.crl.edu/reports/web-archiving-programs-latin-america-and-caribbean>
- Prom, Christopher J. Social Feed Manager: Guide for Building Social Media Archives. University of Illinois at Urbana-Champaign. 2017. <https://gwu-libraries.github.io/sfm-ui/resources/SFMReportProm2017.pdf>
- Rackley, Marilyn. "Internet Archive," in Encyclopedia of Library and Information Sciences, 3rd ed., eds. Marcia J. Bates and Mary Niles Maack, 2966–2976. Boca Raton: CRC Press, 2009.
- Rauber, Andreas, Max Kaiser, and Bernhard Wachter. "Ethical Issues in Web Archive Creation and Usage - Towards a Research Agenda." Research Gate, 2009. https://www.researchgate.net/publication/228638059_Ethical_Issues_in_Web_Archive_Creation_and_Usage_-_Towards_a_Research_Agenda
- Stump, Andra and Dory Bower. "Understanding web archives access and use with Google Analytics: Lessons and questions from the Federal Depository Library Program." Archive-It Blog. October 10, 2017, <https://archive-it.org/blog/post/understanding-web-archive-access-and-use-with-google-analytics-lessons-and-questions-from-the-federal-depository-library-program/>
- Wayback Machine Save Page Now. Internet Archive. Accessed November 20, 2023. <https://web.archive.org/save/>

Sharing and Using Web Archives 31

Related Resources 35

Organizations and Tools 36

Blogs, Papers, Guides, and Other 36

Resources on Web Archiving

Examples of Web Archiving Projects 38

Bibliography 41

About the Author 44

About the Digital Library of the Caribbean 45

Purpose of this Guide

Archive-It. “About Us.” Archive-It and Internet Archive. Accessed November 20, 2023. <https://archive-it.org/learn-more/>

Acker, Amelia, Anne C. Loos, and Julia Sufrin. “The Neil deGrasse Tyson Problem: Methods for Exploring Base Memes in Web Archives.” In SMSociety’20: International Conference on Social Media and Society, eds. Gruz, Anatoliy, Philip Mai, Raquel Recuero, Ángel Hernández-García, Chei Sian Lee, James Cook, Jaigris Hodson, Bree McEwan, and Jill Hopke, 255-264. New York: Association for Computing Machinery, 2020. <https://www.ameliaacker.com/wp-content/uploads/2022/05/Neil-Degrasse-Tyson-Problem.pdf>

Association of Research Libraries. Code of Best Practices in Fair Use for Academic and Research Libraries. 2012. <https://www.arl.org/wp-content/uploads/2014/01/code-of-best-practices-fair-use.pdf>

Binder, Matt. “Myspace Lost 12 Years of Music and Photos, Leaving a Sizable Gap in Social Network History.” Mashable. October 29, 2021. <https://mashable.com/article/myspace-data-loss>

Binder, Matt. “Yes, You Can Still Visit Bob Dole’s 1996 Campaign Website. Here Is Its Story.” Mashable. October 29, 2021. <http://mashable.com/article/bob-dole-1996-website>.

Conifer. “About.” Rhizome. Accessed November 20, 2023. <https://conifer.rhizome.org/faq>

Community Webs. “Community Webs.” Archive-It and Internet Archive. Accessed November 20, 2023. <https://communitywebs.archive-it.org/>

DocNow. “Documenting the Now.” DocNow. Accessed November 20, 2023. <https://www.docnow.io/>

Dole Kemp ’96. 4Presidents.org. Accessed November 20, 2023. www.dolekemp96.org/main.htm

Bibliography

Introduction to Web Archiving is part of a handbook series developed for partners of the Digital Library of the Caribbean (dLOC) and the broader community working within libraries, archives, and other heritage and memory organizations.

Over the last twenty years, dLOC has worked to strengthen its partner and practitioner network through training, professional development institutes, and a large range of outreach initiatives. *The Revitalizing dLOC Initiative* (2022-2026) builds on this foundation by prioritizing direct investment in the Caribbean institutions and people who comprise dLOC, enabling higher levels of sustained community and collections growth. *Revitalizing dLOC* scaffolds programmatic support for executing core activities such as digitization, preservation, copyright, and the ethical reuse of dLOC collections in teaching and research.

Introduction to Web Archiving is designed to provide a brief introduction to web archiving. Included is an overview of key concepts and terms associated with the practice; considerations and strategies for planning and implementing web archiving and social media archiving projects; and a comparison of common tools, resources, and software. Also included is a list of resources for further exploration and learning.

Introduction

- Florida State University Coronavirus Archive: <https://archive-it.org/collections/13632>
- The Latin American Government Documents Archive (LAGDA) (University of Texas at Austin): <https://www.archive-it.org/collections/176>
- University of Florida: <https://archive-it.org/home/ufl-sasc>
- University of North Carolina-Chapel Hill Confederate Monument Protests: <https://archive-it.org/collections/10953>

Conifer/Webrecorder

- BLM Collection – Web-Archiving Social Media Activism of the Black Collective: https://conifer.rhizome.org/s_solomon/blm-collection--web-archiving-social-media-activism-of-the-black-collective
- Newark Library: <https://conifer.rhizome.org/NewarkLibrary>

Social Feed Manager

- Coronavirus Tweet IDs (2020), by Daniel Kerchner, Laura Wrubel, and Dolsy Smith: <https://doi.org/10.7910/DVN/LW0BTE>, Harvard Dataverse, V9
- U.S. Capitol Riot and #TrumpRally Tweet IDs (2021), by Daniel Kerchner and Laura Wrubel: <https://doi.org/10.7910/DVN/I5WC7Q>, Harvard Dataverse, V1
- News Outlet Tweet Ids (2017), by Justin Littman, Laura Wrubel, Daniel Kerchner, and Yonah Bromberg Gaber: <https://doi.org/10.7910/DVN/2FIFLH>, Harvard Dataverse, V4

DocNow

- DocNow catalog of Twitter datasets: <https://catalog.docnow.io/>

What is Web Archiving?

Today, information is increasingly found exclusively online. Content that was previously only available in print form is now only available online, and every day huge amounts of new web content is published. However, despite the internet making information more accessible, it is important to understand that this content is in fact often more vulnerable to loss than traditional print records. Online content can just as quickly vanish when a commercial entity deems it no longer valuable, or support no longer exists for the organizations hosting the content. Content is ever-changing as websites receive new updates and redesigns, and older digital archives are simply deleted without a trace. For example, in 2019, millions of original songs, photos, and videos that were uploaded to the social media website MySpace prior to 2015 were completely lost during a failed data migration with no chance for recovery. Prior to Facebook, MySpace was the number one social media website in the United States.¹

Web archiving is the process of gathering data from portions of the World Wide Web to ensure preservation and access for researchers in the future. It consists of capturing and preserving websites containing images, videos, documents, and social media. This process can be done in several ways, using a variety of tools and software, from simple static screen captures to automated systems known as web crawlers that capture data based on parameters set by an archivist or curator.

¹ Binder, Matt, "Myspace Lost 12 Years of Music and Photos, Leaving a Sizable Gap in Social Network History," Mashable, October 29, 2021, <https://mashable.com/article/myspace-data-loss>.

Brief History of Web Archiving

The first web page was created by CERN, the European Organization for Nuclear Research, in 1991. That page in its original form no longer exists, but a copy of the page was archived and made available on the World Wide Web Consortium a year later. Many websites created in the 90s no longer exist or have been dramatically altered from their original form. In 1996, the Internet Archive was founded as a non-profit digital library to help preserve web content. Without this effort, many more websites would be lost to time.²

In 2001, Internet Archive opened its doors to the public with the release of the Wayback Machine, which allows users to search for specific archived web pages by entering their URL in the search fields. Currently the archive is the world's largest public web archive, with all collections freely accessible to the public. In 2005, the Internet Archive launched a service called Archive-It, which is designed for institutions mandated to preserve web content but which lack the IT infrastructure or technical staff necessary to fulfill that mandate. In 2013, the "Save Page Now" feature became available, which allows the public to target a URL to be crawled and included as part of the Internet Archive and Wayback Machine holdings.³

In addition to the Internet Archive and Archive-It, many other programs and software, both proprietary and open-source, have been developed to assist with preserving the web. These include projects like "Webrecorder.io" or Conifer, which pioneered some innovative methods of capturing and preserving web content, as well as programs and projects like Social Feed Manager and DocNow, which are efforts to tackle the difficult-to-capture and rapidly growing wealth of social media web content.

² Dowling, Stephen, "Why There's so Little Left of the Early Internet," BBC, February 28, 2022, <https://www.bbc.com/future/article/20190401-why-theres-so-little-left-of-the-early-internet>.

³ Rackley, Marilyn, "Internet Archive," in Encyclopedia of Library and Information Sciences, 3rd ed., ed. Marcia J. Bates and Mary Niles Maack (Boca Raton: CRC Press, 2009), 2966–2976.

<https://archive-it.org/blog/post/understanding-web-archive-access-and-use-with-google-analytics-lessons-and-questions-from-the-federal-depository-library-program/>

- Preserving Social Media (2016), by Sara Day Thomson: <https://www.dpconline.org/docs/technologywatch-reports/1486-twr16-01/file>
- International Internet Preservation Consortium (IIPC)'s Tools and Software (2023): <https://netpreserve.org/web-archiving/tools-and-software/>
- Library of Congress's Web Archiving Program – Frequently Asked Questions: <https://www.loc.gov/programs/web-archiving/for-site-owners/frequently-asked-questions/>
- Web Archiving @ Yale: Intellectual Property Rights and Web Archiving (2022): <https://guides.library.yale.edu/c.php?g=870243&p=7146684>

Examples of Web Archiving Projects

Archive-It

- The Aanischaukamikw Cree Cultural Institute's web collection: <https://archive-it.org/organizations/2119/>
- Archive-It Public Access Catalog: <https://archive-it.org/>
- The Biblioteca Nacional Aruba (BNA): <https://archive-it.org/home/aruba>
- Community Webs Collections: <https://communitywebs.archive-it.org/collections/>
- The COVID-19 Florida Farmworkers Collection (University of Florida): <https://covid19.archive-it.org/collection/16190/>

- Day 2 Slides: <https://www.dpoe.network/wp-content/uploads/2021/07/DPOE-N-Slides-Day-2-June.pdf>
- Twitter Was Already Difficult To Archive, Now It's Worse! (2020), by Kritika Garg: <https://ws-dl.blogspot.com/2020/07/2020-07-15-twitter-was-already.html>
- An Evaluation of Web Archiving Programs in the U.S. Relevant to International and Area Studies: The Example of Latin America and the Caribbean (2019), by Jeffrey Garrett: <https://www.crl.edu/reports/web-archiving-programs-latin-america-and-caribbean>
- Legal Issues in Web Archiving (2016), by Abbie Grotke: <https://blogs.loc.gov/thesignal/2012/05/legal-issues-in-web-archiving/>
- Ethics and Archiving the Web: How to ethically collect and use web archives (2021), by Muira McCammon, Michael Nelson, and Lorena Ramírez-López: <https://metro.org/events/ethics-and-archiving-web-how-ethically-collect-and-use-web-archives>
- An Overview of Web Archiving (2012), by Jinfang Niu: <https://www.dlib.org/dlib/march12/niu/03niu1.html>
- Social Feed Manager: Guide for Building Social Media Archives (2017), by Christopher J. Prom: <https://gwu-libraries.github.io/sfm-ui/resources/SFMReportProm2017.pdf>
- Ethical Issues in Web Archive Creation and Usage - Towards a Research Agenda (2009), by Andreas Rauber, Max Kaiser, and Bernhard Wachter: https://www.researchgate.net/publication/228638059_Ethical_Issues_in_Web_Archive_Creation_and_Usage_-_Towards_a_Research_Agenda
- Understanding web archives access and use with Google Analytics: Lessons and questions from the Federal Depository Library Program (2017), by Andra Stump and Dory Bower:

Web Archiving in the Field

Case Study: University of Florida–Covid-19 Florida Farmworkers Project

In 2020, the University of Florida (UF) George A. Smathers Libraries received a grant from the UF Center for Arts, Migration, and Entrepreneurship to preserve online content that documented the impact that COVID-19 was having on farmworkers in Florida. Recognizing that many of the web pages, social media, and digital content created by and for farmworker organizations during the pandemic was ephemeral and could be lost over time, Margarita Vargas-Betancort, Librarian for Latin American & Caribbean Special Collections at UF, developed the digital archive *Documenting History: Farmworkers in Florida during COVID-19*. The grant award enabled the project team to hire an undergraduate student to capture digital content from July 2020 to May 2021. The award was later supplemented with special project funding from the UF Department of Special and Area Studies Collections to extend the project through July 2021.

The grant award did not include funding to purchase web archiving software or services, so from the beginning the plan was to simply capture online content in whichever formats would be easiest and most suitable for preservation. For example, the team decided to save websites in PDF format rather than in WARC file format. The disadvantage to saving web pages in PDF is that the pages are static (little more than screenshots or digital printouts of the pages) and the functionality of dynamic content such as video players and drop-down menus would not be preserved. The major advantage to the PDF format is that it is a standard file format used for preserving and providing access to documents in digital libraries, and it is very simple and cheap to create. The team decided that preservation of and access to the informational content of the websites and social media was more important than preserving functionality and interactive elements of websites.

The team also took the same approach with news articles, social media, and UF-produced publications. The majority of this content

was saved in PDF format, while media such as videos were saved in standard media file formats (MP4). All files were saved initially using Google Drive, and then the files were transferred to UF server storage. By the end of the project, the team had preserved over 300 videos, 300 news articles, almost 500 social media posts, and a few dozen research studies and polls, among other document types. Most web content is protected by copyright, and the team determined they would not share these materials in a public digital library like the Digital Library of the Caribbean or UF Digital Collections without permission.

While the project was underway, Matt Kruse, Head of Collection Services at UF, was simultaneously experimenting with Archive-It as a solution for web archiving that could preserve both the content and the interactive elements of websites. The project team successfully made use of Archive-It to crawl the websites of thirteen farmworker organizations. These are now preserved and accessible online at the Internet Archive:
<https://covid19.archive-it.org/collection/16190/>

Guide to the COVID-19 Florida Farmworkers Collection:
findingaids.uflib.ufl.edu/repositories/2/resources/1776

Project documentation:
ufdc.ufl.edu/IR00011204/00001/citation

Case Study: University of Texas at Austin–Latin American Government Documents Archive

The Latin American Government Documents Archive (LAGDA) is a web archiving project of the LLILAS Benson Latin American Studies and Collections at the University of Texas at Austin. As part of their mandate, the LLIAS Benson collects a vast array of physical formats of Latin American government documents. Over time, government documents have shifted from physical formats to born-digital formats stored on government websites.

With support from the Andrew W. Mellon Foundation, LAGDA was founded in 2003 to respond to evolving documentary and web

Organizations and Tools

- Archive-It: www.archive-it.org
- Community Webs: www.communitywebs.archive-it.org
- Conifer: <https://conifer.rhizome.org/>
- Documenting the Now: <http://www.docnow.io/>
- How to download your information from Facebook: www.facebook.com/help/212802592074644
- How to download your X (Twitter) Archive: <https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive>
- Internet Archive: <https://archive.org/>
- Social Feed Manager: <https://gwu-libraries.github.io/sfm-ui/>
- Society of American Archivists: <https://www2.archivists.org/groups/web-archiving-section>
- Webrecorder: <https://webrecorder.net/>

Blogs, Papers, Guides, and Other Resources on Web Archiving

- Descriptive Metadata for Web Archiving: Recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group (2018), by Jackie Dooley and Kate Bowers: <https://www.oclc.org/research/publications/2018/oclcresearch-descriptive-metadata.html>
- Sustainable Web Archiving at Scale: An Introduction (2021), by Sumitra Duncan and Amye McCarther
 - Day 1 Slides: https://www.dpoe.network/wp-content/uploads/2021/07/DPOE-N-Slides_Day-1_June.pdf

Related Resources

archiving models. As part of their initial grant, the project team used Archive-It to crawl, store, and publicly host their collections, which contain websites of approximately 300 government ministries and presidencies and include full-text official documents, audiovisual materials, and more.

Much of the information currently stored in LAGDA no longer exists on the web. For instance, the web presence of the Honduran President Manuel Zelaya would have been lost had they not implemented a web archiving program.

The LAGDA can be accessed through Archive-It:
<https://www.archive-it.org/collections/176>⁴

Case Study: Community Webs

Launched by the Internet Archive and the Institute of Museum and Library Services in 2017, the Community Webs program is designed to assist public libraries and other cultural heritage organizations build web archives that document local histories and underrepresented voices. In 2020, with support from the Andrew W. Mellon Foundation, the program expanded to include numerous public libraries across the US. In 2021, Community Webs opened to international applicants.

More than 150 organizations have joined the program. Each participating organization is equipped with a multiyear subscription to Archive-It, a half terabyte of storage space, training and educational resources, a practitioner network, and more.

Collections by participating organizations are available on the Community Webs webpage:
<https://communitywebs.archive-it.org/collections/>

Make sure to check out the collection from dLOC partner Biblioteca Nacional Aruba: <https://archive-it.org/home/aruba>

⁴ Garrett, Jeffrey, An Evaluation of Web Archiving Programs in the U.S. Relevant to International and Area Studies: The Example of Latin America and the Caribbean, Center for Research Libraries Global Resources Network, 2019, <https://www.crl.edu/reports/web-archiving-programs-latin-america-and-caribbean>.

Planning a Web Archiving Project or Program

Collecting institutions might also consider gathering data that illustrate how web collections are discovered and used. Some web preservation software, like Archive-It, facilitates integration with platforms like Google Analytics. Information can detail the number of sessions, users, average session duration, and page views, as well as detailed technical (but anonymized) information about users, such as which web browsers they use. Beginning in 2015, the U.S. Government Publishing Office started making use of Google Analytics to gain insights into how users were accessing their web collections on Archive-It. Such information may also help justify the continuation and expansion of pilot web archiving projects to administrators and stakeholders.¹⁴

When planning a new web archiving project or program, consider the following:

1. Scope of the project
2. Selection process
3. Copyright
4. Ethics
5. Budget and other resource requirements
6. Levels of access
7. Making the case for web archiving⁵

Scope of the Project and Selection Process

Consider what type of web collection you are interested in creating. Collection types may include:

- **Institutional:** collecting just your institution's web pages.
- **Subject based:** collecting web resources on specific topics.
- **Event based:** collecting on key current events that help strengthen existing collecting priorities (e.g. COVID-19 pandemic, elections, natural disasters, social movements).

Consider if your proposed web collection fits within your organization's collection policies and priorities:

- Would your proposed project fill a gap in your organization's collections?
- Would your proposed project add value to your organization's collections?
- Have your organization's existing collections changed due to the shift to digital?

Consider what content should be included and excluded:

- Does the website document a specific event, policy, creative act, or course of action, as opposed to general subject matter?
- Is the creator of the web content an original source?
- Is the information unique or does it already exist in other forms within your collections?

¹⁴ Stump, Andra and Dory Bower, "Understanding web archives access and use with Google Analytics: Lessons and questions from the Federal Depository Library Program," Archive-It Blog, October 10, 2017, <https://archive-it.org/blog/post/understanding-web-archive-access-and-use-with-google-analytics-lessons-and-questions-from-the-federal-depository-library-program/>.

⁵ Duncan, Sumitra, and Amye McCarther, "Making the case to your administration," from "Sustainable Web Archiving at Scale: An Introduction," Digital Preservation Outreach and Education Network, 2021, pp. 57, https://www.dpoenetwork.org/wp-content/uploads/2021/07/DPOE-N-Slides_Day-1_June.pdf.

- Has this web content or similar content been crawled or archived by another institution? Do some research on web preservation websites like Internet Archive to see if it has already been preserved.
- Consider nominating relevant websites to existing collections maintained by other institutions so you can focus your resources on an untouched area.
- Consider the risk of web pages disappearing. For example, a time-limited event or an organization under threat of closure, as opposed to a longstanding and well-supported organization.

Consider the level of capture:

- Top level domain (single web pages):

Pros:

- Focused.
- Less cost to resources and data budgets.
- Sometimes easier to access, as researchers may only be looking for specific information.

Cons:

- Approach can lack context of the wider website and may be lacking in information.

- An entire website:

Pros:

- Provides full context for the web content.
- More information available for researchers.

Cons:

- More costly on data budgets.
- Can be difficult to provide quality control and description.

Consider the frequency of capture:

- A single, one-time crawl vs. multiple crawls over a set period of time.
- Ask yourself: how frequently is the website likely to add new content?

Web collections are made accessible to researchers in a variety of ways similar to traditional library and archival collections. The metadata and descriptive standards used for web collections often depend on the software, tools, staff expertise, and standards already in place at a given collecting institution.

Websites and web collections can be cataloged through OCLC and a library's integrated library system using traditional RDA and MARC cataloging methods. They can also be incorporated into archival finding aids through DACs (Describing Archives: A Content Standard) and EAD (Encoded Archival Description) standards. For those using Archive-It, metadata is created using the Dublin Core standard, which forms the backbone of the integrated public user interface for Archive-It and the Internet Archive's Wayback Machine. A 2018 report by OCLC, *Descriptive Metadata for Web Archiving*, provides an excellent resource for describing individual websites and web collections through a variety of descriptive standards. Taking the time to create consistent, structured, and useful metadata for web collections will greatly improve their discovery and usefulness for researchers.

Researchers use web collections in a variety of ways. They may simply seek to discover documents that were exclusively created and made available on the web, or they may want to access websites related to political campaigns, organizations, events, and companies that have since disappeared. More and more, researchers are leveraging machine learning and data analysis to enhance the search process and to provide new insights into collections at scale. A 2020 study, "The Neil deGrasse Tyson Problem: Methods for Exploring Base Memes in Web Archives," conducted by researchers at the University of Texas at Austin, used web archival collections and datasets to understand how context and language may imbue internet memes with different meanings as users appropriate and reappropriate them. Collecting institutions and programs should understand web content not just as individual websites, but as datasets.¹³

¹³ Acker, Amelia, Anne C. Loos, and Julia Sufrin, "The Neil deGrasse Tyson Problem: Methods for Exploring Base Memes in Web Archives," In *SMSociety'20: International Conference on Social Media and Society*, eds. Gruzds, Anatoliy, Philip Mai, Raquel Recuero, Ángel Hernández-García, Chei Sian Lee, James Cook, Jaigris Hodson, Bree McEwan, and Jill Hopke, 255-264, New York: Association for Computing Machinery, 2020, <https://www.ameliaacker.com/wp-content/uploads/2022/05/Neil-DeGrasse-Tyson-Problem.pdf>.

Sharing and Using Web Archives

Copyright

Most materials on the web, including social media posts, are protected by the copyright laws of countries in the Caribbean, the US, and most others worldwide. In many countries, copyright law allows for “fair use” or “fair dealing” of copyrighted works without the permission of the author, particularly if it is for the purpose of education and research rather than commercial use. The US-based Association of Research Libraries has issued guidance on fair use,⁶ including strategies for web archives (p. 27). Laws also often include provisions for libraries and archives to engage in activities that support preservation. While the details of copyright law may seem daunting in the context of the internet, where content is constantly reproduced and shared, curators should evaluate legal issues alongside their missions and the needs of their users.

When planning web archiving projects, it is important to decide whether or not to ask permission from the copyright holders of websites and other web content before crawling, preserving, and providing access to that content. There are two main approaches to permissions: an opt-in approach or an opt-out approach. This decision should consider not only copyright law, but also available resources and the scale of the web archiving program.

An **opt-in approach** is one in which the archivist does not crawl or provide public access to any web content without first obtaining permission from the copyright holder. This can be a time-consuming process, as the archivist needs to request permissions and track responses from copyright holders, as well as maintain those permissions for potential ongoing crawls. Additionally, this process often has a low response rate. This method is often best employed when working with a smaller scale web archiving project or program and/or when working with particularly sensitive web content that is likely to have risk of copyright issues.

⁶ Association of Research Libraries, Code of Best Practices in Fair Use for Academic and Research Libraries, 2012, <https://www.arl.org/wp-content/uploads/2014/01/code-of-best-practices-fair-use.pdf>.

An **opt-out approach** is one in which the archivist does not ask for individual permissions before crawling, preserving, and providing access to content, but rather provides creators with the right to ask for the removal of their content. This is commonly referred to as a notice-and-take-down policy. This approach is most beneficial for medium-to large-scale web archiving projects and programs and for web content that is determined to have a low risk of copyright issues. Other methods include implementing embargo periods after a crawl has been completed and making content public after a set date. Organizations might also opt for crawling a website for preservation purposes while restricting access to reading rooms in order to mitigate risk of copyright issues.

In some cases, websites will include robots.txt files that indicate to web crawlers that all or part of a website should not be crawled; however, these files are not always applied.

Social media poses unique challenges. Like other materials, social media posts are automatically protected by copyright law once they are in a “fixed” format, even if they aren’t public, and rights are typically held by the individual or organization that created the content. Additionally, most social media websites (Facebook, X (formerly Twitter), etc.) are run by corporations that determine how or if records can be preserved, and web archivists may encounter both legal and technical barriers.

Ethical Considerations

Archives have the power to determine what is preserved in the historical record. In acknowledging the prevalence of bias in archives and how this shapes the historical record, archivists and curators may want to prioritize selecting and preserving web content created by underrepresented communities and organizations, particularly if this web content is vulnerable to loss. At the same time, archivists need to be conscious and respectful of communities that may not want their web content preserved by collecting institutions. Archivists should always consider privacy and confidentiality, keeping in mind that preserved materials have

Advantages

- Very easy to use and access.
- Captures a complete record of posts and data, beyond API limitations.
- One of few reliable ways to capture difficult-to-crawl social media platforms like Facebook.

Disadvantages

- Requires account holder’s explicit permission, ability, and willingness to go through the request process with the social media company.
- Exports may include some fields, such as direct messages, that should generally not be preserved in archives (note: these can be removed with relative ease after export).
- Reliant on the social media company. Terms may change over time.
- Requires a waiting period for the social media company to retrieve and provide access to the downloadable archive.
- Typically does not export in WARC format.
- May not include embedded or linked resources and certain other data.

and social media. Tools include twarc, a command line tool for collecting post data from X's (formerly Twitter) official API, and Hydrator, a desktop application for turning post ID datasets back into actual post data, and various other tools and trainings for working with social media web content.¹²

Advantages

- Designed for social media.
- Free and open-source.
- Tools useful for integration with Social Feed Manager (SFM).
- Great resource for exploring ethical issues within social media preservation.
- Tools designed to be compliant with X's (formerly Twitter) user intent policy.

Disadvantages

- Relatively new and not yet widely adopted by the web archiving community.
- Reliant on X's (formerly Twitter) API and terms of use restrictions.
- Narrowly focused on the X (formerly Twitter) social media platform.
- Requires some advanced technical skills to operate.

Official Social Media Archive Exports (i.e. X (formerly Twitter) Archive and Facebook Archive)

An alternative approach to archiving social media is to request archive exports directly from the social media companies. This can be an excellent option for preserving social media content that would otherwise be inaccessible. However, there are a number of significant limitations.

the potential to be used in ways that may even threaten the safety of communities or individuals. It is important to be selective and not exhaustive when selecting materials for preservation, to consider potential harm caused by web collecting, and to include community perspectives in web collecting strategies.⁷

Determining Levels of Access

Weighing the considerations of copyright, privacy, access, and resources available at your institution, it is important to create a clear policy delineating the levels of access your institution will provide to your web collections. Some options for this include:

Local access: read-only access is allowed through a moderated process on-site at the repository reading room.

Pros:

- Most secure in terms of copyright and ethical considerations.

Cons:

- Time-consuming to manage.
- Restrictive for researchers.

Online access: content is available and viewable by all, typically through a public interface like Archive-It and/or with WARC's available to download by users.

Pros:

- Requires less resources to maintain.
- Functions at scale.
- Less restrictive to researchers.

Cons:

- Less secure in terms of copyright and ethical considerations.

A combination of local and online access is often required and desirable.

¹² DocNow, "Documenting the Now," DocNow, Accessed November 20, 2023, <https://www.docnow.io/>.

⁷ Rauber, Andreas, Max Kaiser, and Bernhard Wachter, "Ethical Issues in Web Archive Creation and Usage - Towards a Research Agenda," Research Gate, 2009, https://www.researchgate.net/publication/228638059_Ethical_Issues_in_Web_Archive_Creation_and_Usage_-_Towards_a_Research_Agenda.

Making the Case for Web Archiving

Web archiving projects and programs can be resource-intensive. They can consume large amounts of staff time and require sufficient resources to purchase and maintain software tools, subscriptions, and storage space. It is often difficult for archivists to make the case to administrators to allocate resources to web archiving projects and programs.

Strategies for making the case to your organization and administrators include:

- Share examples from peer institutions.
- Share trainings, webinars, and other learning opportunities.
- Invite decision makers to conferences on web archiving.
- Calculate the cost to implement and propose a manageable project given your organization's resources.
- Propose pilot projects and make use of free trials offered by subscription services like Archive-It and Conifer.
- Tie proposals in with existing mandates, collection policies, and as supplements to big non-digital acquisitions.
- Time your proposals during key periods:
 - When your organization is doing a big web redesign.
 - When an organization relevant to your institution's collecting area is closing.
 - During pivotal events relevant to your institution's collection policies.
- Create surveys and track metrics to support your proposal.

Advantages

- Designed for social media.
- Free and open-source.
- Relatively easy to use once installed.
- Very customizable and flexible. Allows archivists to directly manage data without relying on social media platforms' designated archive, which needs to be requested and is limited in the amount of data provided.
- Exports files as WARC, but can also be more flexible in producing modular data sets in CSV, Excel, and JSON formats.
- Ability to harvest from additional services and incorporate new features.
- Useful in limiting risk and privacy concerns. Ability to harvest X (formerly Twitter) IDs only, without preserving the data itself. May be useful for collecting on topics likely to be deleted or removed.

Disadvantages

- Relatively new and not yet widely adopted by the web archiving community.
- Most social media companies place restrictions on data that can be harvested through the API. For example, X (formerly Twitter) only allows harvesting of the 3,200 most recent posts.
- Does not work with all social media platforms. It does not work for Facebook or Instagram.
- Terms and limitations by social media companies are volatile and subject to change dramatically when new opportunities arise to monetize datasets.
- Technical knowledge or support required for setup. Runs best on a dedicated server.

DocNow

DocNow, short for Documenting the Now, was founded in 2015 with support from the Andrew W. Mellon Foundation and Princeton University Library. DocNow develops open-source tools and community-centered practices that support the ethical collection, use, and preservation of publicly available content shared on web

- Some limited automation tools developed.
- Includes public access option and ability to download WARC files.
- Cloud storage options in paid version.

Disadvantages

- Manual process is time-consuming and does not function at scale.
- Public user interface is less robust.
- Difficulty importing WARCs into Archive-It if using as a supplement.
- Local storage solution needed if not using paid version.
- Some browser restrictions.
- Primarily designed for smaller organizations and projects.
- Not as widely supported and adopted as programs like Archive-It.
- Primarily useful for capturing traditional websites. Frequent issues with social media.

Select Tools for Social Media Archiving

Social Feed Manager (SFM)

Social Feed Manager (SFM) is a project by George Washington University Libraries. It is a web application that allows users to create collections of data from specific social media platforms: X (formerly Twitter), Flickr, Tumblr, and Sina Weibo (a Chinese website). It is free and open-source software that connects to a social media platforms' public API to harvest data. It can be used to gather datasets and archive social media activity. Data can be downloaded in CSV, Excel, and JSON formats as well as traditional preservation as WARCs.¹¹

¹¹ Prom, Christopher J. Social Feed Manager: Guide for Building Social Media Archives, University of Illinois at Urbana-Champaign, 2017, <https://gwu-libraries.github.io/sfm-ui/resources/SFMReportProm2017.pdf>.

Tools Review and Comparison

Conifer/Webrecorder

Conifer (known as Webrecorder.io between 2015-2020) was originally developed as part of a multi-year research and development project. The goal was to create a “next generation” web archiving service.

The open-source software components are currently maintained independently by the original Webrecorder project. Rhizome, a non-profit organization that preserves digital art, became steward of the propriety service and relaunched it as Conifer.

Conifer is different from most web archiving tools. Instead of using an automated web crawler that follows links from a seed, it creates high fidelity, interactive copies of web pages. Archivists can manually browse web pages during the process of capture, ensuring that dynamic, interactive elements are included in the capture. This browser-based method has a higher success rate than traditional web archives and allows for a more controlled and focused web crawl.¹⁰

Advantages

- Developed for archiving narrow slices of content.
- Controlled and focused results.
- Creates high fidelity, interactive copies of web pages that are browsed manually during the process of capture.
- Can have a higher success rate on more dynamic web pages than traditional web crawlers.
- Manual crawls reduce need for quality control.
- Manual crawls have less ethical and copyright concerns.
- Reduced duplication.
- Smaller impacts on data budgets and storage.
- Easy to use and low-tech barrier to entry.
- Free options with a limited 5GB storage space. Relatively inexpensive at higher tiers.

¹⁰ Prom, Christopher J. Social Feed Manager: Guide for Building Social Media Archives, University of Illinois at Urbana-Champaign, 2017, <https://gwu-libraries.github.io/sfm-ui/resources/SFMReportProm2017.pdf>.

- Wide array of tools to aid quality control work and metadata creation.
- Robust scoping tools and automation.
- Clear exit strategy (option to stop subscription at any time and still maintain collections in perpetuity, download WARCS).
- Great public interface and search tools. High discoverability.
- Options to restrict access.
- Integration with other major software including ArchivesSpace and OCLC. Ability to upload and ingest WARC files obtained outside of Archive-It.
- Hosted cloud storage solution.
- Widespread documentation, training resources, and technical support.
- Continually updated and maintained.

Disadvantages

- Subscription fees are required and can be expensive. There is no free version available, though a tiered pricing structure is available.
- Automated web crawling can lead to additional quality control work in order to ensure desired content was fully captured.
- Scale and automation can lead to issues related to permissions, ethics, copyright.
- Automated web crawling can lead to excess data usage on limited data budgets.
- Reliant on an outside entity for storage.
- Some cases where replays have not been as functional.
- Occasional difficulty capturing dynamic content or might miss some pages. Typically good enough for most content and seems to be constantly improving.
- No ability to shift crawled seeds between collections once established.
- Primarily useful for capturing traditional websites. Frequent issues with social media. In some cases, such as with Facebook, it is no longer functional.

While there are a wide variety of web archiving tools available, there is currently no single program that does everything perfectly. Many programs work well in tandem, filling gaps and weaknesses in others. In particular, tools that target social media capture are often very specialized and differ from traditional web capturing tools. Web content is a moving target, and new tools and services are constantly being developed and updated to meet ever-changing needs. Web archivists and their institutions must constantly evaluate these new tools and challenges to determine which tools work best given their resources, collections, and users.

Key Concepts and Terms

There are a number of key concepts and terms relevant for understanding how web archiving tools work:

Static vs. Dynamic Web Content and Preservation

Increasingly, websites are no longer made up of static HTML files. Many websites today are made up of a complex, dynamic series of scripts; content is pulled on the fly from databases and pages are generated in real time. Content changes based on the user interacting with it. More and more, the new standard is to preserve websites in a way that captures their full, dynamic nature. This is typically done using automated web crawlers and other tools. However, low-tech strategies such as static screen captures and recordings are still useful, especially in cases where the informational content of websites and social media is more important than preserving functionality and interactive elements of websites, or as supplemental processes.

Crawler/Crawling

Web crawlers (sometimes also referred to as “spiders” or “bots”) are automated agents that work to index and capture web data servers based upon a specified “seed” URL and set of scoped parameters. They move through websites following hyperlinks to discover and download linked web content including web pages,

videos, images, and other data. Web crawlers are designed to capture both dynamic and static websites so they can be preserved and reproduced in their original interactive and dynamic form. One example is Heritrix, the Internet Archive's open-source, archive-quality web crawler project. An alternative to using automated web crawlers like Heritrix are browser-based web crawlers like Webrecorder. These are user-driven web crawlers that allow archivists to manually direct the archiving process through their web browser. Both approaches have several advantages and disadvantages that will be discussed in more detail later in this guide.

Seed/entry-point URLs (EPUs)

A “seed” or “entry-point URLs (EPUs)” is a web address or URL that tells web crawlers where to go on the World Wide Web to initiate crawling and harvesting web data. With many automated crawlers, the archivist will specify scope in addition to the seed, in order to direct web crawlers beyond the initial seed and into subsequent linked pages.

robots.txt

A “robots.txt” file is a file that a website owner can add to their website to prevent web crawlers from accessing all or parts of a website. These instructions can be useful to understand which websites grant permissions to crawl, however, usage of robots.txt files is not uniform and there are many cases where archivists can and will make specifications on web crawlers to ignore the presence of these files.

WARC

A WARC, or web archive file, is the standard and internationally accepted file format for preserving websites from web crawls in their original context. The International Internet Preservation Consortium (IIPC) is responsible for maintaining the WARC file format standard. WARCs are container files, meaning that they function similar to ZIP files—they wrap around the various files that make up a website including HTML, JS, CSS, PDF, MP3, and others. WARCs have a limit to how much data they can hold, so it usually

Disadvantages

- Only available for web sites that allow crawlers.
- Not suitable for capturing larger, more complex websites or web content at scale.
- Not currently effective for social media, though Archive-It is improving their capture technology.
- No control over enhancing metadata.

Archive-It

Founded in 1996 as a non-profit digital library, the Internet Archive works to ensure the availability and accessibility of knowledge by storing digital content from all over the world for anyone to view at any time. The archive is the world's largest public web archive with all collections freely accessible to the public through its Wayback Machine. The automated web crawler Heritrix was developed jointly by the Internet Archive and the Nordic national libraries in 2004 and is currently the main web crawler in use by Internet Archive and Archive-It.

Archive-It is a web application that was created by the Internet Archive in 2006. It has provided web archiving subscription services to over 800 organizations in over 24 countries including libraries and other research institutions. Archive-It provides tools for crawling, preserving, describing, and providing access to web content selected and maintained by member institutions.⁹

Advantages

- Developed for medium- to large-scale web archiving programs.
- Well-established history and widely supported and adopted by many academic organizations and the public at large.
- Low barrier to entry with technical support for setup and hosting.
- Tiered pricing structure with support, cost share, free trials, and discount options available for smaller institutions, including the Community Webs program.

⁹ Archive-It, “About Us,” Archive-It and Internet Archive, Accessed November 20, 2023, <https://archive-it.org/learn-more/>.

Disadvantages

- Static format. Cannot dynamically recreate websites. Non-interactive.
- Only effective for certain types of web pages and content. Not effective at capturing larger, more complex websites, or those containing audiovisual formats.
- Susceptible to falsification and unreliable for authenticity.
- Files cannot be integrated into commonly used web archiving systems like Archive-It, which operate using the standard WARC format.

Internet Archive's "Save Page Now" Feature

The "Save Page Now" feature allows members of the public to freely request for single page URLs to be crawled, preserved, and made accessible in the Internet Archive and Wayback Machine. This option is great for institutions and individuals who need a low-tech, low-cost option to preserve web content.⁸

Advantages

- Simple and easy with little to no extra technology required.
- No cost.
- Minimal training required. Can be performed by volunteers or student workers
- Useful for capturing single web pages.
- Provides a permanent link that can be cited without need for any additional storage or hosting.
- Incorporated into the Internet Archive and the Wayback Machine.
- Preserved as a WARC file, the standard web archiving preservation format, which will capture dynamic content on a single page.
- Resistant to falsification.

⁸ Wayback Machine Save Page Now, Internet Archive, Accessed November 20, 2023, <https://web.archive.org/save/>.

takes multiple WARCs to store a website. WARC files require special technology to replay them, such as the Internet Archive's Wayback Machine and the free and open-source ReplayWeb.page. Because of their standardization, WARCs are flexible. They can usually be transferred between different web archiving tools.

Considerations for Scope and Sustainability

Open-source vs. Proprietary Software

One key consideration when reviewing web archiving tools is the decision to use open-source or proprietary and subscription-based software.

Open-source software can be very beneficial as these programs are free to use. These tools are also flexible and can often do things that some of the proprietary software cannot do. In many cases, the open-source software is the same software that the proprietary services use as their foundation. One concern when using open-source software services is that they can fall out of date if they are no longer being updated or have shifted over to paid versions. They also tend to have less support attached to them and less streamlined interfaces, which can lead to the need for strong internal IT support or tech-savvy staff at your institution.

Proprietary software offers upgrades and features that the free versions do not. While one drawback is the monthly/annual costs and typical data restrictions, opting for proprietary software can be a cost saver when it comes to staff time. Features like automated updates, automated crawls, data deduplication, internal cloud storage, enhanced quality control tools, batch metadata entry tools, built-in public access interfaces, and technical support can often pay for themselves. However, recurring fees can pose budgeting challenges over time. It is therefore important to develop exit strategies detailing how you will get your data out if you need to end your subscription or the service is no longer operational.

Because proprietary software services are designed to appeal to groups and organizations, they tend to lack the flexibility of open-source products. One important thing to remember is that many open-source tools start out free, supported by grants and organizations. But to remain viable and grow, the web archiving community needs to support them. Becoming involved in the larger web archiving community can influence how a software develops over time.

Level of Adoption by Web Archiving Community

Another important factor to consider when selecting tools is the level of community adoption. In general, it is advantageous to choose tools that have been widely adopted. Widely adopted tools are:

- Less likely to be abandoned by the software companies that create and maintain the tools.
- Receive upgrades and new features that keep them current.
- Include software integration with other widely adopted tools.
- Current with shared standards and expertise.
- Include training information, and technical and community support.
- More information for migration and exit strategies.
- Likely to be accepted by decision makers at your organization.
- Consist of public access systems and can be utilized for discovery.

Social Media vs. Traditional Websites

Lastly, a key consideration when selecting tools is determining if you are going to be collecting social media web content as well as traditional websites. While many traditional web archiving tools (Archive-It, Conifer, etc.) can capture social media, they often have limitations and difficulty capturing everything, and their output and public access features are not specifically designed with social media in mind.

Select Tools for Traditional Web Archiving

Tools and software for entry-level, low-tech web archiving.

Screen/video capture

The simplest method is to use screen/video capture software to capture a static image of a webpage or series of webpages. This can then be saved as an image file or in PDF format.

In Windows 10 and 11, this can easily be done using the built-in Snip & Sketch App. Use the keyboard shortcut “Windows key” + “Shift” + “S” to capture a select portion of the screen. You can then save the image. Other keyboard shortcuts such as “Windows key” + “Print Screen” will save your current screen to the “Pictures > Screenshots” folder on your computer. You can then crop and edit with simple image editing software such as Paint.

Mac users can use the keyboard shortcut “Cmd-Shift-3” to screenshot the entire screen. “Cmd-Shift-4” allows for greater manipulation of the screenshot. Additionally, most video recording software, both free and paid, have screen capture options.

Advantages

- Simple and easy with little to no extra technology required.
- Generally free.
- Minimal training required. Can be performed by volunteers or student workers.
- Can be shared and hosted through all common means of hosting digital content. Does not involve special readers and display software like WARC formats.
- Useful for capturing difficult-to-crawl websites, such as social media.
- Best for capturing single pages or posts on websites.