Folger SHAKESPEARE LIBRARY
Advancing knowledge & the arts

FOLGER SHAKESPEARE LIBRARY

WEB ARCHIVES

Prepared by:
Jaime McCurry
2013-14 National Digital Stewardship Resident
Folger Shakespeare Library

May 2014

Web Archiving

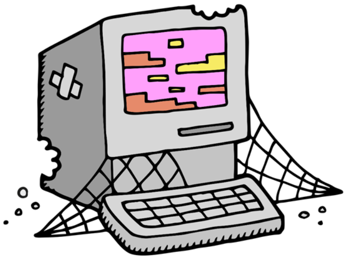
Web archiving is the process of "collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use" (IIPC). Archiving the web allows us to combat the impermanent nature of online content, making future access and use possible.

Web content is harvested through a process in which a web crawler accesses and gathers content from designated URLs through a process referred to as crawling. A web crawler is an internet "bot," or program, that browses the web for indexing purposes. Crawlers access the desired website in a similar way to a web browser and captures all content related to the site, including any necessary information needed to render the site correctly as if it were live on the web: CSS files, etc.

The results of these crawls are captures of web content that can then be archived, described, and curated into digital collections. There are multiple digital resources involved in the capture and harvesting of even just one seed. A seed is an individual URL within a web archive collection. Following a web crawl, the information pertaining to a seed is organized into a WARC preservation file. The WARC file format is able to contain all necessary information and digital resources gathered from a seed during a crawl. It can also be expanded upon to include ancillary metadata elements. Websites archived in the WARC file format can be viewed and interacted with in a web browser using access tools such as the Internet Archive's Wayback Machine.

Advanced manipulation of web archive data can facilitate a number of research techniques: potential uses for web archive collections include textual or link analysis, among others.

Ultimately, web archiving is intended to preserve a realm of cultural history that is increasingly present, and sometimes only present online in digital format. Digital information is very sensitive. Sites are reliant upon a number of external factors in order to be accessed by users: content creators, host domains, web browsers, markup languages, etc. Subsequently, internet content can disappear for a variety of reasons frequently and often without notice.



"Mr. Shakespeare is now retired."

- Users are greeted with an unfortunate message upon arrival at http://shakespeare.palomar.edu

For example, the popular web resource *Mr. Shakespeare and the Internet* (http://shakespeare.palomar.edu) was taken offline in October of 2013. If not saved, the information it contained would have been lost to users. Fortunately, the website was archived in time by the Internet Archive and is made accessible via the Wayback Machine.

Web Archiving at the Folger Shakespeare Library

The Folger began archiving and preserving select websites using the Archive-It subscription service in October of 2011. Collections are administered by the Folger Shakespeare Library: Central Library. They can be accessed here (URL: http://archive-it.org/organizations/576).

The mission of the Folger Shakespeare Library is as follows: "to preserve and enhance our collection; to make our collection accessible to scholars and others who can use it productively; and to advance understanding and appreciation of Shakespeare's writings and the culture of the early modern world."

Developed by Jim Kuhn (Head of Collection Information Services, 2006-2013) and Emily Wahl (Central Library), the Folger web collections were created to address a new update (2010) to the Folger Collection Development mandate which expresses an institutional commitment to digital collecting in Shakespeare-related areas, including born-digital ephemera.

For more on web archiving at the Folger Shakespeare Library:

An Introduction to Web Archiving at the Folger
Web Archiving I Folgerpedia
William Shakespeare: Playwright, Icon, Web Archivist? | The Archive-It Blog
Continuing the Celebration: Preserving Birthday-Related Digital Ephemera

Archive-It Subscription Service

Archive-It, an initiative of the Internet Archive, is the subscription web archiving service that the Folger uses to build and maintain its current web collections. For the most up-to-date service information, including information on upcoming training and webinar opportunities, refer to the Archive-It help documentation. The Archive-It blog is an additional resource for information on the latest service updates and related content.

Updates to the Archive-It web application are frequent. At the time of this writing, 4.9 is the most recent version. Archive-It 5.0, which will be a "major overhaul" of the service, has a tentative release date of Summer 2014. An explanation and a full list of upcoming features can be found here.

Additionally, Archive-It hosts an Annual Partner Meeting where partner institutions and partner specialists meet to discuss the Archive-It service, partner projects, and the current professional environment surrounding the web archival field. The Folger presented at the November 2013 Partner Meeting in Salt Lake City, Utah. The presentation, which can be accessed here, was titled: "'The Short and The Long' of It: Web Archiving at the Folger Shakespeare Library."

Administering the Folger Shakespeare Library Web Collections

Administrator Role

The Web Archive Administrator is responsible for collection creation and management, seed selection, metadata description, crawl activities, quality control, user training, and collection advocacy, both internally at the Folger Shakespeare Library and externally at relevant conferences and events.

Google Drive Shared Workspace and Public Contact

Folger Web Archive Administrator(s) now have access to a shared working environment on Google Drive. This collaborative workspace allows administrators to access, edit, and add documentation relating to the Folger Web Archives from anywhere, for internal use purposes. The Folger Shakespeare Library Seed Nomination Form and user response submissions are stored here along with the code and framework for the 2014 #Shax450 Tweet Archive. Additionally, this account hosts the newly created Web Archive Administrator Public Contact Email: folgerwebarchives@gmail.com, which allows Administrators to consider comments, requests (take-down, etc.), and questions relating to the Folger Web Archives directly from the user audience.

The Web Archive Administrator is responsible for maintaining folgerwebarchives@gmail.com and Google Drive documentation; including, but not limited to: monitoring and adjusting seed nomination forms, evaluating seed nomination form responses for inclusion in current and future collecting efforts, maintaining and creating Archive-It training documentation, and creating new genres of documentation as needed.

Collection Development

Collecting Scope

The Folger Shakespeare Library Web Archives exist to compliment the Library's existing mission to "preserve and enhance our collection; to make our collection accessible to scholars and others who can use it productively; and to advance understanding and appreciation of Shakespeare's writings and the culture of the early modern world."

Folger web collecting activities aim to digitally close gaps in the collecting process and to create new areas of thematic expansion for the Folger Shakespeare Library collections. Each existing collection has an individual collecting scope and all are appropriate for the general collecting mission of the Library.

Collection Development

When considering new collection topics or individual seeds for inclusion in a collection, consider: how could this collection and/or site compliment:

- existing web archives initiatives;
- existing Folger collections;
- future collecting intentions

Seed Selection Criteria

The Web Archive Administrator should verify, to the best of their knowledge, that the website they are considering is:

- Created and/or maintained by a reliable source
- Immediately relevant to the collection scope and theme
- Of potential cultural, historical, and research value to the Folger's user audience and the general public

Acquisition Sources

The Web Archive Administrator will work with information provided by Folger staff and by its user audience to identify additional collecting areas, new collection themes, and ways to improve collecting practices:

Web Archive Administrator

The Web Archive Administrator is responsible for evaluating and selecting seeds for the Folger web collections based on their individual research; the needs, suggestions, and comments of Folger staff and readers; and the needs, suggestions, and comments from the Folger Shakespeare Library general user audience.

Recommending Officers (Institutional Stakeholders)

Folger staff and Folger readers may contact the Web Archive Administrator directly to discuss potential collecting areas and to nominate individual seeds. This includes stakeholders in all departments and working groups, such as the Collection Development Committee and the Online Strategy Council.

Additionally, the Web Archive Administrator will take advantage of internal opportunities to consult with Folger departments such as Central Library, Digital Media and Publications and working groups such as the Collection Development Committee and the Online Strategy Council to obtain general guidance on web collecting activities and avenues of collecting interests.

Folger staff and readers may also contact the Web Archive Administrator via the public contact email: *folgerwebarchives@gmail.com* or they may nominate a website using the *General Nomination Form*.

User Audience and General Public

The broader user audience and general public may interact with the collections on the Folger's Archive-It homepage, learn about the collections on Folger web resources such as the *Collation*, *Folgerpedia*, and Hamnet's *E-Resources* page, and offer their thoughts and suggestions on improvements and expansion via the public contact email *folgerwebarchives@gmail.com* and the General Nomination Form.

Nomination Forms

The nomination forms are created and maintained in the Google Drive shared workspace environment that is utilized by the Web Archive Administrator(s). Forms may be created as needed, may serve a general purpose (see the General Nomination form) or may serve a more specific purpose in enhancing an individual collection (see the Shakespeare's 450th collection form). These forms are created to be shared publicly via Folger resource such as blogs, wikis, and social media announcements.

- General Nomination Form
- Shakespeare's 450th: Nomination Form

Permissions & Copyright Policy —DRAFT—

The Folger Shakespeare Library Web Archives program was created to encourage and support scholarship and research in the arts and humanities disciplines in an accessible manner to contemporary audiences. Collecting as a nonprofit library, archive, and a leading educational resource for educators and scholars, all Folger Shakespeare Library web preservation efforts are intended to be non-commercial in nature and non-intrusive in form. The Web Archive Administrator will remove harvested web content from the archive upon request by site owner(s).

Code of Best Practices in Fair Use for Academic and Research Libraries (ARL: January 2012)
U.S. Copyright law at 17 U.S.C. § 107

Robots.txt Files / Site-Owner Communication

Historically, there has been no communication with site owners and no permissions requested if and when ignoring robots.txt files during the collecting process. Due to the smaller nature of the web archives program at the Folger, it would be problematic to investigate site owners for every individual site for our collections. Our hope is that owners who wish to discuss our collecting practices will feel free to contact us through our newly created folgerwebarchives@gmail.com public contact email and will learn more about our collecting intent and institutional mission through continued outreach and educational opportunities on the web via the *Collation Blog*, *Folgerpedia*, and other pertinent resources.

Description

The Archive-It web application utilizes the Dublin Core Metadata Element Set, Version 1.1 for collection-level, seed-level, and item-level description. These elements are repeatable. Archive-It also allows for the addition of a custom field for further description, if need be. To ensure consistency across all Folger web collections, the following guidelines were created: (Please note: * Indicates mandatory fields). For more information on each individual element, please see the Dublin Core Usage Guide: The Elements documentation.

TITLE*

The name given to the website by the creator or publisher:

- Capitalize the first word of the title and all proper nouns. All other words should be lowercase
- Exclude articles (*the, an, a,* etc.)
- Include all punctuation marks found within the title as it appears on the web

CREATOR*

The person or entity primarily responsible for creating and maintaining the intellectual content of the resource

- Whenever possible, use a Library of Congress Name Authority; otherwise, utilize a standard name authority format

SUBJECT*

The topic of the intellectual content of the resource:

- Use a controlled vocabulary such as the Library of Congress Name Authority and/or Subject Headings when possible: http://authorities.loc.gov/
- Example:
  Subject: Shakespeare, William, 1564-1616.

DESCRIPTION*

A brief, free-text explanation of the resource: its function and what type of intellectual content it contains.

PUBLISHER*

The person or entity primarily responsible for making the resource available.

CONTRIBUTOR

A contributor is any entity which makes intellectual contributions to the resource in partnership with the Creator.

DATE

The specific date of creation of the resource. Only include the date if there is an explicit creation date stated on the site.

TYPE

The nature or genre of the content of the resource:

- Type: Website

FORMAT

The physical or digital manifestation of the resource. Format may include the media-type or dimensions of the resource. Examples of dimensions include size and duration. Additionally, format may be used to describe the software, hardware or other equipment needed to display or operate the resource.

IDENTIFIER

An unambiguous reference to the resource within a given context. Examples of formal identifiers include the Uniform Resource Identifier (URI), the Uniform Resource Locator (URL), the Digital Object Identifier (DOI), and the International Standard Book Number (ISBN).

SOURCE

A pre-existing source from which this resource has derived.

RELATION

Describe, if any, the resource's relation to another, more general resource. This may include digital records and other websites, groups, or entities. Use a controlled vocabulary when possible.

COVERAGE

The coverage element describes the places or locations which are covered in the intellectual content of the resource. Additionally, temporal coverage may be applicable, which refers to the general time period described by the intellectual content of the resource. For example:

- Coverage: United States
- Coverage: 1950-1999

RIGHTS

Information, as available, pertaining to the rights of the resource. This includes Copyright and Property Rights.

COLLECTOR*

Collector: Folger Shakespeare Library. Central Library.

LANGUAGE*

The language in which the intellectual content of the resource is written. Utilize the ISO 639-1 codes for language classification; for example, a resource written in English would be described as:

Language: en.

Collection Management Processes: General Best Practices

Adding Seeds to a Collection

- Be sure to check both active and inactive seeds when adding new seeds to avoid duplication.

Modifying Crawl Scope

- Adding limits allows you to restrict the number of sites crawled. This is a general best practice when crawling a site with a calendar (in order to avoid crawler entrapment). To limit the crawl, block all URLs containing a string of characters unique to the calendar pages, or to the page which you want to exclude from capture.

Edit Metadata

- To edit seed-level metadata, do so from Seed Management, not from Collection management.

Viewing Reports (Key Things to Look For…)

Hosts

- URLs: identify the host responsible for the most number of URLs captured within the crawl and also for those additional URLs which were captured that you did not expect, or which were captured in smaller numbers than you anticipated. If you had expected more URLs to be crawled, then this site likely has a robots.txt file blocking most of the content from capture. Alternatively, if a site has listed too many URLs crawled, it might be worthwhile to consider modifying the crawl scope to limit the crawl.

Data

- Anything under 100 MB is generally not problematic due to the current data allowance in place per our vendor agreement with the Archive-It service. At this time, if you are in doubt, there is no way to remove captured content form a crawl: this is why limiting and test-crawling beforehand is helpful. Future iterations of the Archive-It program may include the option to erase content.

Queues

- If a crawl has run out of time based on the pre-defined time limit, there will likely be uncrawled URLs left in the Queue. Please identify these and patch crawl as necessary.

Out of Scope

- Be sure to check if any portion of the site you have intended to capture isn't included in the Out of Scope list (specific URLs which have been excluded from crawl boundaries). Modify the crawl scope and patch crawl as necessary.

Seed Status

- If a seed is live on the web and able to be crawled, it will be listed as "OK." If the status is unavailable or not OK, be sure to check that the website hasn't been taken down or that you haven't entered the incorrect URL by mistake.
- Occasionally, a 'Redirected' will appear. The best practice is to change the host to whatever URL it redirects to (when appropriate and after verifying the redirect is within scope).

Quality Assurance (QA)

- The Quality Assurance report request takes 12-24 hours to process, and can only be done with full crawls (i.e. not on patch crawls).
- QA reports may be utilized to show how well the sites were captured and to show problematic areas within the crawl and individual pages which weren't captured.

Example Workflow

- Choose a collection theme/topic relevant to institutional collecting intentions
- Describe the collection in Collection Management
- Select individual seeds to be included as pertinent to the collection scope
- Identify a proper crawl schedule for these seeds (quarterly, semi-annual, etc.)
- Run a test crawl on the chosen seeds
- Review test crawl reports, specifically the Seed Source report, weed out unruly seeds
- Review the queue to determine what content was missed, or where the crawler ran into a trap
- Review the Out of Scope report
- Place limits on the crawl as needed
- Run an official crawl
- Run reports on the completed crawl. Pay specific attention to what might have been missed due to blocked robots.txt
- Run a QA report and review
- Patch crawl as needed
- Describe the successfully crawled seeds in Seed Management
- Review the final collection
- Make the collection available to the public
- Re-crawl as scheduled and/or needed

Collection: Folger Shakespeare Library Websites and Social Media

URL: https://archive-it.org/collections/2873

Archived Since: October, 2011

Type: Institutional

Crawl Frequency: Quarterly

# of Seeds: (4/14) 35

Collecting Scope: Websites and social media sites from the Folger Shakespeare Library.

Selection Criteria: This collection is an institutional collection which archives and preserves the Folger Shakespeare Library's web presence over time. Sites included in this collection must be considered to be one or more of the following:

- Folger web domain
- Folger social media profile
- Folger blog or Wordpress site
- "Other" web resource: created and/or maintained by the Folger Shakespeare Library, or to which the Folger has a made a significant intellectual and/or creative contribution (ex.: http://penfaulkner.org)

Descriptive Metadata: This collection utilizes the Dublin Core Metadata Element Set, Version 1.1

Specific Descriptive Metadata: Publisher: the platform the content is published on. For example:

Publisher: Wordpress (Electronic Resource)
Publisher: Twitter.
Publisher: Facebook.

Rights: Copyright © Folger Shakespeare Library ®.

Collection: Shakespeare Theatrical Festivals and Performing Companies

URL: https://archive-it.org/collections/2877

Archived Since: October, 2011

Type: Thematic

Crawl Frequency: Semi-Annual

# of Seeds: (4/14) 285

Description: Websites for drama festivals and theatrical companies with a focus on Shakespeare performance. Scope primarily limited to the United States, but includes some international festivals as well.

Selection Criteria: Accept an explicit statement that the festival or company repertoire is rooted in Shakespeare's corpus, or if Shakespeare is featured consistently as part of their repertory.

Descriptive Metadata: This collection utilizes the Dublin Core Metadata Element Set, Version 1.1

Specific Descriptive Metadata:

Title: The specific title of the company or festival, as it appears on their site.
Creator: Same as title.
Subject: (as applicable)

1. [Title]
2. Shakespeare, William, 1564-1616 — Stage history — [state, province, etc.] — city
3.Repertory theater — [state, province, etc.] — [city]
4. Drama festivals — [country]
5. Theater, Open-air — [state] — [city]
   (Used if their repertory features outdoor productions).
6. Non profit
7. Traveling theater

8. Drama in education — [state]—[city]
   (Used for festivals and companies that are affiliated with a school or university; educational programs run by the company do not count).
9. Charity
10. Amateur theater
11. Improvisation (Acting)

Description: A brief overview of their mission statement. Include a physical address if found.

Date: Only include an end date if the website specifies that the company is no longer active or if they have no record of being active in 5 last years.

Collection: William Shakespeare's 450th Birthday: Celebrations and Commentary

URL: https://archive-it.org/collections/4511

Archived Since: April, 2014

Type: Event-Based

Crawl Frequency: One-Time Only

# of Seeds: (5/1) 17

Description: This collection seeks to document various celebrations, commentary, and events as depicted on the web related to the 450th anniversary of William Shakespeare's birth.

Selection Criteria: Accept official event websites; news articles; media commentary; or any related resource which specifically discusses the 450th anniversary of William Shakespeare's birth.

Descriptive Metadata: This collection utilizes the Dublin Core Metadata Element Set, Version 1.1

Specific Descriptive Metadata:	Type:	Interactive Resource.
	Date:	YYYY-MM-DD
	Subject:	Shakespeare, William, 1564-1616—Anniversaries, etc.

#Shax450 Tweet Archive and Visualization

Created in April 2014 using Martin Hawksey's TAGSExplorer and Google Spreadsheets, the #Shax450 Tweet Archive is an interactive archive and visualization of tweets that have used the hashtag #Shax450 on Twitter to celebrate William Shakespeare's 450th birthday.

Using TAGS (Twitter Archiving Google Spreadsheet) technology, Hawksey's TAGSExplorer tool visualizes Twitter conversations, connections, and user activity which are all related to a specific hashtag. This is achieved by hooking up to Twitter's API and transferring relevant data to a Google Spreadsheet for organization and analysis.

Hashtags are used on social media sites to link various content of a similar nature together by keyword. A hashtag is a word or a phrase prefaced with a #, generally referred to as a "pound sign" or a "hash." A hashtag does not support spaces or punctuation marks. The #Shax450 Tweet Archive has gathered all Twitter activity marked #Shax450 from April 16, 2014, on. This specific hashtag is largely Folger-centric and statistics related to the tag paint an interesting picture of social media interactions surrounding the birthday on Twitter.

The TAGSExplorer program hooks up to Twitter's API and pulls activity related to the hashtag into a spreadsheet, runs analytics on the data, and generates a visualization of conversations and actions. The text of the Tweets is saved in the archive, along with relevant available metadata such as the day and time the Tweet was created, whether it was a stand-alone instance or if it generated interaction and conversation, and more. The Archive is updated for new information on a continual basis. The #Shax450 Tweet Archive can be accessed here, archive analytics here, and the interactive visualization here.

Additional Resources

Vendor Resources

Archive-It

Archive-It Blog

Archive-It Help Documentation

Training and Webinars

Version Release Notes

Overview | Terms, Tools, and Definitions

Archive-It
Glossary of Web Archiving Terms

International Internet Preservation Consortium (IIPC)
Web Archiving: Tools and Software

International Internet Preservation Consortium
Web Archiving: Why Archive the Web?

D-Lib Magazine
An Overview of Web Archiving

Wikipedia
Web Archiving

Best Practices and Guidelines

Archer, J., Fallon, T., Grotke, A., Odell, K. (2011). [PowerPoint Slides].

Creating and Maintaining Web Archives

Bibliographies / Additional Learning Resources

International Internet Preservation Consortium

Web Archiving Bibliography (2012)

Web Archiving Use Cases (2013)

METRO New York: More Podcast, Less Process. Episode 007.

"The Web Archivists Are Present." (2014)

SAA Web Archiving Roundtable
Weekly Web Archiving Roundup

Notable Projects and Initiatives

California Digital Library
Web-At-Risk Project
This project was grant-funded by the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) group and culminated in the Web Archiving Service, a web archiving tool for libraries and archives.

Columbia University
Human Rights Web Archive
This project, operating through Columbia University and made possible with support from the Andrew W. Mellon Foundation, is one of the first fully-described collections of archived websites. The project was stewarded by a knowledgable team and has made use of seed nomination forms and general public contact methods. Additionally, they have created a Permissions and Copyright statement. This project also utilizes the Archive-It service.

Internet Archive
Archive.org
Wayback Machine
The Internet Archive is a 501(c)(3) non-profit that was founded to build an Internet library. To-date, they have saved over 411 Billion web pages. Tools developed by the Internet Archive such as the Wayback Machine, used for rendering archived web content and the Heritrix Web Crawler are widely used in the web archiving field and by Archive-It. Archive-It is a service of the Internet Archive.

Library of Congress
Library of Congress Web Archives (Minerva)
History of Web Archiving at the Library of Congress
K-12 Web Archiving Program
The Library of Congress has been archiving and preserving select websites on behalf of the United States government since 2005. They have an official Web Archiving Team which works closely with the National Digital Information Infrastructure and Preservation Planning (NDIIPP) group to research web archiving trends, develop new tools, and to lend their expertise to the creation of standards and best practices for web archiving.

UK National Archives
UK Government Web Archive
The UK Government Web Archive Project, like Library of Congress web archiving efforts, is another government project aiming to archive the web: their goal is to preserve information specifically related to the UK National Government made available on the World Wide Web. The interface to their collections, as linked above, is notable, along with their use of themed collections and their creation of a Twitter Archive and Video Archive.

Professional Groups and Events

Digital Preservation Coalition (DPC)
Web Archiving and Preservation Task Force

International Conference on Digital Preservation
Annual iPres Conference *Note: while this conference is not solely focused on web archiving, it generates discussion on the topic regularly.

International Internet Preservation Consortium (IIPC)
Annual General Assembly Event

National Digital Stewardship Alliance
NDSA Web Archiving Survey

Society of American Archivists
Web Archiving Roundtable

The images used in this report are from the Digital Preservation Business Case Toolkit; licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License.