Getty

WEB ARCHIVES COLLECTING POLICY[1]

The J. Paul Getty Trust ("Getty") is dedicated to the presentation, conservation, and interpretation of the world's artistic legacy, serving both the general interested public and a wide range of professional communities, in order to promote a vital civil society through an understanding of the visual arts. A crucially important part of this work is the presentation and dissemination of freely available cultural heritage activities and information to the world through the use of websites, social media, and other online platforms. In keeping with its mission, Getty has instituted a web archiving program to permanently document both its own contributions to the world's artistic legacy and those of closely related partners, projects, and collaborators, as presented on the web.

Scope

The goal of the web archiving program is to preserve records that deserve to be retained for the long term in accordance with Getty's information management policies and perceived historical interest. Primarily, this includes collecting the web content produced by the Getty itself, but also, increasingly, the websites of organizations whose papers are held by the Getty Research Institute's (GRI) Special Collections, and select sites meeting the research and documentation needs of all Getty departments and programs. The purpose of this archival capture is to attempt to preserve the content as well as the look-and-feel of these resources as they existed at particular points in time. For captures of content that have a limited period of access, the underlying text, multimedia, and the governing contracts and agreements should be maintained outside of the websites and transferred to the Archives as appropriate.

Methodology

To capture Trust web content, in 2017, Getty's Institutional Records and Archives department chose to partner with Archive-It, a subscription web archiving service from the Internet Archive that helps organizations harvest, build, and preserve collections of web content. Archive-It enables its partners to collect, catalog, and manage their collections of archived web content with 24/7 access and full-text searchability. While Archive-It is the primary tool used for web archiving, Getty also employs tools, such as Conifer (previously Webrecorder), WinHTTrack, and PDF as needed for specific applications.

Fidelity

As a visual arts institution, we strive to capture the visual content as well as the textual content of the pages we collect. This is not always possible, however, given the limitations of the crawlers and rendering technologies to preserve the exact form, functionality, and content of sites as they appear on the live web. The following types of content present significant issues for capture or display:

- Dynamic scripts or applications such as JavaScript or Adobe Flash
- Streaming media players with video or audio content

[1] *This policy was finalized in July 2021 by a cross-department team from Digital and Institutional Records and Archives, and will be reviewed approximately every two years to address any changes in the field or technology.*

- Password protected material (we do not collect any web-content which requires a password to access)
- Forms or database-driven content (i.e. anything with a search box) that requires interaction with the site
- Exclusions specified in robots.txt files
- Certain content management systems/platforms are known to be problematic for web archiving. See Archive-It's website for more details.

Note: If Getty-owned images and media featured on webpages are part of Museum or GRI Collection material, those are already being preserved through normal repository processes. Any other images and media produced by Getty must be deposited in Institutional Archives for ongoing preservation and access.

Access

Archived web content is accessible through our Archive-It organization page, where access is provided to public-facing collections through a version of the Wayback Machine. This content is not currently accessible via Rosetta, the digital preservation system used to deliver other digital content from Institutional Archives and the GRI's holdings.

Restricted collections are only accessible on the Getty's internal network by logging into the Archive-It account. A future consideration is whether we will also provide access to certain web archives through Rosetta or another offline playback alternative for more controlled access.

Metadata

At this time, very minimal metadata is being added In Archive-It or captured as part of the process. This is something that will be enhanced in the future as additional standards for describing web archives are further developed. OCLC and the web archiving community at large are working on these guidelines and we will be following their progress.

Data Storage

Captured digital content is hosted and stored at the Internet Archive data centers. The service's storage and preservation policy can be found here: https://support.archive-it.org/hc/en-us/articles/208117536-Archive-It-Storage-and-Preservation-Policy . An evaluation of the need/desire to export the WARCs for preservation in Rosetta is underway.

Web Archive Collections

The Getty's web archive collections are divided into three primary categories:

1. Getty content on Getty-developed web properties
2. Content contributed by the Getty to third-party platforms
3. Third-party content captured by the Getty as part of acquisition processes or to document external cultural heritage information

Collection details: 1. Getty content on Getty-developed web properties

This category includes Getty.edu and its subdomains, web content developed for the Getty by outside contractors, through institutional partnerships, or captures of partner institution's web pages.

- Unless specifically prohibited by contractual agreements, content in this collection is made publicly available for reference purposes only.
- Any use or reuse of media and images contained in Getty web pages is subject to the Getty's rights and reproductions policies.
- Captures of this collection material occur periodically with additional captures being scheduled as needed. For example, during website redesign phases or when digital art history projects are sunsetted.
- Efforts are made to ensure the look-and-feel of Getty generated pages are maintained to the greatest possible extent. Full fidelity, however, may not be attainable. Content owners will work with the web archiving team to periodically review captures and point out where more robust crawls might be needed.
- It is recognized that third-parties may be crawling Getty-generated web pages and providing independent access to them. The Getty takes no responsibility for this activity, though the Getty's web properties use robots.txt minimally to restrict crawling.

Collection details: 2. Content contributed by the Getty to third-party platforms

This category includes the Getty's social media sites, such as its Facebook, Instagram, Flickr, Soundcloud, Twitter, Google Arts & Culture, and YouTube pages.

- The Getty captures snapshots of the ways in which the Getty represents itself on third-party public platforms, such as profile pages and periodic samples of content.
- Because of potential issues with licensing of content that has been posted on these pages, the social media captures are not available to the public. Captures are for internal use only.
- Captures are not, and are not intended to be, comprehensive, as there are known issues with crawling social media sites.
- It is recognized that third parties may be crawling Getty-generated content on social media sites and providing independent access to it, the Getty takes no responsibility for this activity.

Collection details: 3. Third-party content captured by the Getty as part of acquisition processes or to document external cultural heritage information

This category might include the capture of the website of a gallery whose papers have been acquired by the Getty Research Institute, captures of online auction catalogs for purposes of provenance research, captures of other institution's web pages for a Getty traveling exhibition, or a Twitter feed of a topic related to art history.

- If the content is in the form of a freely and publicly accessible PDF, staff should submit it themselves to the public Wayback Machine (click "Save Page Now"). After submitting the URL, access to the PDF can be provided using the generated link. No further archiving is necessary. This archives the content in the general Wayback Machine, not the Getty's Archive-It collection.

- Captures of more complex sites must be requested, as needed, by a Getty staff member who will act as sponsor and point person for the project. The request will go through an approval process starting with the web archiving team. Please contact Institutional Records and Archives to discuss this kind of capture.
- The degree of comprehensiveness and success of the crawl will be guided by the project sponsor and the web archiving staff member, based on project needs, crawler technical abilities, and the press of work.
- If the capture results in a significant amount of data we may need to evaluate the timing of capture and the subscription size.
- We will obey website-specific instructions concerning archiving, whether expressed in a machine-readable format using the robots exclusion standard or in reasonably discoverable human-readable text. In cases where these directives would prevent the archiving of content, the project sponsor will seek written permission from the content owner before proceeding.

Out of Scope

The Getty's internal wiki (share.getty.edu) and intranet (GO) are not captured with Archive-It. These resources will need to be captured using different methods.

Responsibilities

Capture/archiving team (i.e. staff who initiate captures)

- Manage the Archive-It subscription.
- Organize and manage archived collections of institutional content in Archive-It, entering appropriate descriptive metadata at the collection level and at seed level if warranted
- Monitor the Conifer account (free version)
- Run crawls to archive web content and troubleshoot issues internally and with the Archive-It support staff - more advanced technical issues are escalated to the Digital Preservation Manager
- Monitor enhancements and changes to the tools; leveraging new capabilities as appropriate
- Responsible for keeping Web Archives Collecting Policy (this policy) current, working with Digital Preservation Manager

Digital Preservation Manager, Collection and Content Management Systems (CCMS), GDi

- Evaluate services (i.e. Archive-It, Conifer) from a long-term preservation perspective and monitor changes in this area
- Recommend best practices for making sites archivable (up to content producers whether or not they want to follow guidelines)
- Work with Rosetta Systems Administrator to configure Rosetta to handle WARC files as needed
- Responsible for keeping Web Archives Collecting Policy (this policy) current, working with capture/archiving team

Digital Content Strategy, Communications

- All Getty website and blog content will be crawled and captured several times a year on a schedule to be worked out with Digital Content Strategy.
- Digital Content Strategy is responsible for all Getty-wide web content standards, structures and core guidelines.

Getty Content Owners

- If significant changes are made to Getty web content or legacy Teamsite pages are taken offline, content owners must notify the Getty's web archivist two weeks in advance to allow time for additional captures and any needed review and troubleshooting. See the Web Content Retirement Guidelines for additional information
- Review the results of the regularly scheduled crawls at least annually

Staff Requesting Capture of External Sites

- Identify, appraise, and select external websites that reflect the mission and collecting interests of the Getty
- Provide descriptions and contextual information for material
- If crawls are scheduled regularly, review the results at least annually
- Mediate access (via metadata, catalog records, and an access interface) to associate sites with current holdings of the GRI Library and Special Collections
- Respect the intellectual property rights of owners and ensure compliance with all applicable laws and policies
- Reach out to the Interpretive Software Development team lead when employing new website designs, configurations, or technologies to evaluate potential future capture issues.

Specific Technical Choices

- Keep videos within scope for the Getty Website collection even though we are also crawling YouTube separately (because otherwise, videos won't play in context when embedded in pages)
- For the Collection pages, only the first page needs to be captured to show look and feel. Data management for collections information is done in TMS.
- Contentstack, the CMS used for getty.edu, is headless and does not contain any media assets or presentation code. There are plans to create a regular data export routine.
- Code for the presentation layer will be backed up as well as part of our github data management.
- Scope out external news media
- Scope out search/filter/calendar features

Glossary

Conifer

Conifer is a user-driven platform allowing users to create, curate, and share their own collections of web materials. This can even include items that would be only revealed after logging in or performing complicated actions on a web site. On the technical side, Conifer focuses on “high fidelity” web archiving. Items relying on complex scripting, such as embedded videos, fancy navigation, or 3D graphics have a much higher success rate for capture with Conifer than with traditional web archives. [Source: https://conifer.rhizome.org/_faq]

Archive-It

Since 2006, Internet Archive’s Archive-it has provided web archiving services to over 800 organizations in over 24 countries, including libraries, cultural memory and research institutions, social impact and community groups, and educational and open knowledge initiatives. Archive-it users have preserved over 40 billion born-digital, web-published records, totaling petabytes of data.

Archive-It provides tools, training, and technical support for capturing and preserving dynamic web materials, as well as a platform for partners to share their collections, with multiple search, discovery, and access capabilities. Material archived via Archive-It is stored in not-for-profit data centers independently owned and operated by the Internet Archive, and is available for users to download themselves for additional preservation and sharing. Together, Archive-It users and the Internet Archive are furthering the shared ethos of ensuring perpetual access to diverse, cultural, and historically-relevant digital collections from around the world. [Source: https://archive-it.org/blog/learn-more/]

WARC

WARC is a file format for the long term preservation of digital data. It stores web pages and other digital resources including images and meta information in their original source code.The WARC format is an international ISO standard https://www.iso.org/standard/68004.html

Crawl

Website Crawling is the automated fetching of web pages by a software process, the purpose of which is to index the content of websites so they can be searched. The crawler analyzes the content of a page looking for links to the next pages to fetch and index.

Site crawls are an attempt to crawl an entire site at one time, starting with the home page. It will grab links from that page, to continue crawling the site to other content of the site. This is often called “Spidering”.

Page crawls, which are the attempt by a crawler to crawl a single page or blog post. [Source: https://www.sovrn.com/blog/website-crawling-information/]

Seed

A Seed URL in web crawling is a url from which a web crawler will begin to traverse a site. Once a crawler is on a seed URL it will extra data from the page and look for all links

to additional pages. If a crawler is set to crawl an entire domain it will systematically follow each link on every page, extracting data from each ensuing page. Paths from a seed URL are often influenced by a websites Robots.txt file, which dictates how the site owner would like bots to traverse the site. [Source: https://blog.diffbot.com/knowledge-graph-glossary/seed-url/]