https://doi.org/10.2352/issn.2168-3204.2017.1.0.11

Archiving Websites Containing Streaming Media

Howard Besser; New York University; New York, NY, USA

Abstract

The software most North Americans use to archive websites is notably deficient in capturing streaming media. This paper reports on a project to re-architect the Heritrix web crawler in a newer approach to archiving websites. The project focuses on web content produced by contemporary young composers, and also explores developing relationships with these creators that address other web archiving issues such as copyright and high quality capture

Background Web Archiving Issues

As the web has matured, it has evolved from a platform of static documents, collocated images and media files, and fixed hyperlinks into a more complex publishing platform, increasingly dependent on a mix of dynamic and interactive server-side and browser-based behaviors. Audiovisual material is particularly challenging for the automated processes of web archiving due to the issues posed by embedded third-party playback tools, format and streaming methods, and non-canonical URIs. Web archiving tools and processes are thus required to keep pace with constantly evolving web technologies. [1] [2]

In North America, almost all web archiving employs the Internet Archive's Heritrix web crawler. This web crawler has serious problems capturing streaming content, and when it does capture it, frequently the content is displayed out of context on the resulting web pages.

In the early stages of this New York University (NYU) project we identified and documented a variety of the crawling problems that resulted in incomplete archiving: embedded third-party playback tools, format and streaming methods, video and audio inconsistencies, audio functionality, navigational context, and non-canonical URIs. We also identified problems with particular types of software tools and delivery platforms (e.g. SoundCloud, Vimeo, YouTube) whether they were embedded or not. We examined a variety of other problem sources, including web-building tools like WIX and the use of embedded Google Maps. And we developed a specific plan of work to bring web archiving tools into alignment with contemporary methods of delivering and accessing audiovisual materials on the web.

Background on Music Composers' Websites

The World Wide Web has changed the nature of historical documentation. In the field of music, many activities -- promotion, publishing, and distribution of scores and recordings; critical discourse; correspondence; and even performance -- have moved decisively to the Internet. Websites have thus become an essential component of cultural memory, and will inevitably be tomorrow's historical documents. Yet websites are extremely unstable, fragile, and vulnerable to loss. Anyone with an interest in historical documentation should be concerned about the volume of significant content disappearing daily from the cultural record.

Many music websites have enormous potential research value as primary sources in the history of music. Subjects of study include individual composers, performing artists and ensembles, concert organizations and venues, and recording and publishing companies, all of which have a substantial and growing presence on the web, as does commentary by music critics and by everyday listeners through blogs and social media.

There are many thousands of composers active around the world, and many of them have their own websites with biographical and career information. These sites frequently include (or have links to) audio and video recordings of the composers' music as well as notated musical scores in PDF or another format that represents even richer content than is streamed on the web. These sites constitute an extraordinary wealth of content and context for understanding the music of the early 21st century, but preserving them through standard web archiving practices results in the loss of much of their value for two reasons: 1) the limited quality of the audio and video provided via standard web access, and 2) the limitations of content capture for dynamic content.

The Mellon Composers Project

We developed a formal proposal to build the new web archiving software and test it on a body of web sites created by young composers likely to become prominent in the future. The proposal also included the building of significant relationships between the NYU Library and the composers to assure ongoing archiving of each composer's future work. With generous support from the Andrew W. Mellon Foundation, work on this project began in 2015 and will continue through mid-2018.

The initial set of composers was based upon a 2011 crowd-sourced list of "100 Composers Under 40" commissioned by US National Public Radio. [3] This corpus was later augmented by NYU music librarians. A key goal of the project was to archive how early-career composers represent themselves with a web presence, and to be able to archive how their self-representation might change over time. Ideally this would be achieved by the NYU Library and Archive maintaining an ongoing relationship with each composer.

The primary technical goal of the project was to extend existing web archiving tools and services to not only collect audio and video streams, but also to present the results in proper context.

Another project goal was to also collect and archive the higher-quality content used to create the audio and video streams on the websites, and this is being done by building relationships between NYU and the composers.

Still another goal was to develop language for donor agreements that would allow for deep scholarly use of the archives without impeding on the rights of the composers.

And another aspect of the project involves integration of the new web crawls with existing NYU Library workflows and services. NYU has partnered with Hudson-Molonglo to develop an API for ArchivesSpace that will provide new functionality for dissemination of metadata about audiovisual assets held in NYU's digital repository. With modifications to the Archive-It service by the Internet Archive, the planned outcome of this integration will be the seamless presentation of data about the assets of composers collected by NYU with the versions of their archived websites in Archive-It.

Thus far, all composers contacted have been pleased by the interest shown in their work by an academic archive, and have indicated a willingness to let NYU allow researchers to view and query the archived websites. They have indicated basic agreement with donor language allowing significant scholarly use, and have only suggested minimal language changes to the draft donor language.

Web Archiving Technical Challenges

The near ubiquity and constant evolution of the JavaScript scripting language [4] which powers browser-based user interactivity and dynamic content generation within a browser, has posed specific challenges to current methods of web archiving. Traditional crawlers that support web archiving, such as Heritrix, [5] do not load web pages in a browser, instead acquiring files directly from the host server. This means that URLs or other content or links generated through user interaction may not be discovered, and acquired for archiving, by the crawler.

The ubiquity of JavaScript on webpages has led to dynamically generated URLs that reveal the location of multimedia assets only after a certain user interaction within the browser. Because of these dependencies, the media asset often cannot be captured without the page being opened in a browser and a specific input or action being performed. Moreover, the URLs generated by these interactions are often themselves transient, existing only for a short period of time and no longer valid by the time they are queued for capture by the crawler.

Current tools working to address the demands of JavaScript-induced archiving problems include Umbra [6] the open source tool developed by Internet Archive that mimics user-behaviors within a browser to expose dynamically-generated URLs. Umbra works "alongside" Heritrix, discovering content and URLs and feeding this information to the Heritrix crawler for capture. Headless browsers (essentially browsers without a graphical user interface) and WebKits like PhantomJS also attempt to automate the capture of JavaScript-heavy sites. Tools that work alongside crawlers, aka "helper" tools like Umbra, have necessitated ongoing enhancements to the Heritrix crawler itself, especially in how it queues URLs for acquisition, as well as the development and customization of middleware, such as advanced message queuing protocol tools,[7] that facilitate communication between helper tools and web crawler. As part of the Mellon Composers project, the Internet Archive will develop new tools or significantly enhance existing tools such as Umbra and Heritrix to address these challenges and improve the capture of audiovisual materials on the web.

Another current challenge for web archiving is the proliferation of streaming audiovisual services, such as YouTube, SoundCloud, FlowPlayer, Vimeo, and other third-party web services. The ease of use, low- or no-cost pricing, and embedded player options of these platforms makes them widely utilized. However, content creators uploading and storing content on these services cede control to the providers over how their audiovisual materials are made available (via streaming or download). Technical details such as media asset format, bit rate, and encoding are determined, and often frequently reconfigured, by these third-party platforms and their commercial, rather than preservation, concerns. The method of dynamic generation of URLs used by these services can cause assets to be unavailable for capture or can lead to acquiring unwieldy amounts of duplicate content. At scale, inaccessibility can lead to significant content not being captured, and duplicate captures can overwhelm data storage and quality control capacities and lead to websites being excluded from overall acquisition strategies.

The variety of audio and video formats and resolutions available on the web continues to proliferate. YouTube alone delivers a wide range of asset types, resolutions, and audio encoding methods, in both DASH (Dynamic Adaptive Streaming over HTTP, which facilitates the adaptive bitrate streaming referenced above) and non-DASH distributions. Choosing which version is captured and archived has required ongoing crawler customization and continues to complicate both capture and playback of archived multimedia content.

Current tools and approaches to addressing these challenges have varied, and development work focusing on this set of problems has been minimal. [8] File format challenges have been addressed by utilizing different playback tools, thus allowing a broader range of formats to be captured. Continued development of the Heritrix crawler has aimed to improve its ability to capture streaming content and the variable bitrates at which content is delivered to browsers. Deduplication algorithms and tools have attempted to identify duplicate content accessible through multiple URLs or packaged in different encodings. These approaches, however, have not kept pace with the speed at which web technologies change.

Building a New Web Crawler -- Brozzler

After exploring a number of options, we decided that it would be most useful to replace the Heritrix web crawler with a new crawler. In collaboration with the Internet Archive team responsible for Heritrix we began work on developing a new crawler named Brozzler (the name indicating that it combined features from a browser with the functions of a crawler). The critical development work on Brozzler was carried out by the Internet Archive. Brozzler was created as an open source Python based application.

Brozzler is a distributed, browser-based web crawler that takes advantage of multiple tools and systems to improve the archiving of the dynamic, streaming, and browser action-driven content – especially audiovisual content – that is increasingly common on the current web. Combining browser automation tools, archiving proxy systems, new crawler management tools, and platform-specific libraries, this distributed crawler aims to improve the discovery and capture of audiovisual content, specific HTTP methods, and the dynamically-generated URLs and source content increasingly dependent on user actions within a browser. Brozzler

is intended to supplement (and eventually replace, in some crawling environments) the Internet Archive developed Heritrix web crawler, which has been the standard crawler for web archiving since its release in 1996. Brozzler's features and architecture take advantage of scalability, distribution, and modularity to resolve the limitations of Heritrix in capturing contemporary web content.

Brozzler employs a distributed, modular architecture. It is scalable. And it crawls pages, not URLs.

Collecting Higher Quality Content

Composers' artistic achievements will be the primary determinant of their historical significance. It follows that web archivists should seek to preserve the documents of their music—in particular the audio and video recordings—at the highest quality possible. Composers' websites, however, do not typically offer the highest quality audio and video files, but instead (for practical reasons) use compressed file formats such as MP3 or, alternatively, streaming media services such as YouTube, Vimeo and SoundCloud. Collecting archival sources and highest-quality files will require curators to contact and work directly with composers and site owners -- a major outreach effort and key component of this project.

Additionally, if music files and other elements associated with composers' websites are to be preserved at the highest quality, then as curators we would aim to improve the research experience by developing new tools that provide seamless intercommunication between web crawling services and these "locally" acquired -- that is, obtained directly from the composer or site owner -- preservation-quality files. And at the same time, the preserved sites themselves need to more accurately reflect the original live site experience.

Current State of the Project

Problems with archiving streaming media and presenting it in its proper context have been identified. Brozzler has been designed to solve those problems. The beta version of Brozzler was released in 2016. [9]

As of early 2017, over 100 composers have been contacted, with almost all of them giving an initial indication that they will be involved in the project. Ten composers have carefully reviewed draft intellectual property and donor agreements, and their comments have been incorporated into the language of a standard agreement (which should be finished and released before the end of 2016). These ten composers have also agreed to supply the high quality content used to create the streams on their websites.

In fall 2017 we will begin our evaluation stage where we will examine display and usability factors and iteratively improve the archival display system. We will look at composer comfort level as to how their archived websites are displayed, and examine music researchers' ability to use the archived websites. We will also ask the composers more detailed questions about donor agreements and intellectual property issues.

Implications for the Field

At the end of this project we should have the websites of over 100 young composers available for researchers to study. The public release of the new Brozzler crawler and its incorporation within other web archiving tools such as Archive-It should greatly improve the archiving of websites that contain streaming audio or video. The development of ongoing relationships between NYU Libraries and the composers, and the donor and intellectual property templates we have created in tandem, can serve as models for future collaborations between cultural heritage institutions and creators.

Acknowledgements

This project has involved a team of about a dozen people from NYU Libraries and the Internet Archive. Particular thanks are due to NYU Project Director David Millman, Internet Archive Project Director Jefferson Bailey, and NYU Digital Archivist Don Mennerich.

References

[1] International Internet Preservation Consortium, "IIPC Future of the Web Workshop – Introduction & Overview," 2012 http://netpreserve.org/sites/default/files/resources/OverviewFutureWebWorkshop.pdf *(retrieved 5 March, 2017)*

[2] D. Rosenthal, "Not Your Grandfather's Web Any More", Coalition for Networked Information, 2013 http://www.cni.org/topics/digital-preservation/not-your-grandfathers-web-any-more/ *(retrieved 5 March, 2017)*

[3] National Public Radio Music, "The Mix: 100 Composers Under 40; WQXR's Q2 Presents A Crowdsourced Selection Of Young Composers," 2011 http://www.npr.org/2011/04/23/135473622/the-mix-100-composers-under-40 *(retrieved 5 March, 2017)*

[4] British Library, "Web Archiving in the JavaScript Age," British Library Blog, 2014 http://blogs.bl.uk/webarchive/2014/08/web-archiving-in-the-javascript-age.html *(retrieved 5 March, 2017)*

[5] P. Jack, "Heritrix Wiki", Jira Agile Software Development, 2014 https://webarchive.jira.com/wiki/display/Heritrix/Heritrix *(retrieved 5 March, 2017)*

[6] J. F. Brunelle, M. C. Weigle and M. L. Nelson, "Archiving Deferred Representations Using a Two-Tiered Crawling Approach," in Proceedings of iPRES 2015, 2015.

[7] Wikipedia, "Advanced Message Queuing Protocol" http://en.wikipedia.org/wiki/Advanced_Message_Queuing_Protocol *(retrieved 5 March, 2017)*

[8] G. Truman,. "Web Archiving Environmental Scan", Harvard Library Report, 2016 https://dash.harvard.edu/handle/1/25658314 *(retrieved 5 March, 2017)*

[9] GitHub, 2016 https://github.com/internetarchive/brozzler *(retrieved 5 March, 2017)*

Author Biography

Howard Besser has been involved with digital preservation since the 1990s, has taught classes and dozens of workshops on the subject, and has published numerous articles on it. In 2009 he was named to the Library of Congress' select list of "Pioneers of Digital Preservation". He has also been involved in the creation of several library metadata standards (PREMIS, Dublin Core, METS), and has published more than 50 articles dealing with technology and cultural institutions