DIGITAL **AGNES**

Agnes Etherington Art Centre at Queen’s University
agnes.queensu.ca

Archiving the Web

Agnes Bytes
23 October 2025

KEYWORDS

Archives, Web Archives, Cultural Memory, Internet Culture, Browsertrix

SPEAKERS

Danuta Sierhuis, Kaloyan Kolev and Tessa Walsh

TRANSCRIPT

Danuta Sierhuis: Okay. All right, well we'll get started and more people will join as they get from their lunch breaks. Welcome everyone, this is the first talk of the Atmos Bytes series, a new series of lunchtime talks on critical themes in digital art and culture. I am Danuta Sierhuis, the Digital Development Coordinator at Agnes Etherington Art Centre at Queen’s University, and I use she/her pronouns. I am a white woman with shoulder length brown, curly hair. I am wearing a black t-shirt, hot pink squiggly earrings and glasses. And I am sitting in my office, and there is a bookshelf and are some posters and on a white wall behind me.

Thank you all for joining us today on your lunch hour. While we meet online today on Zoom, I would like to begin by acknowledging that Agnes Etherington Art Centre is located in what is known as Ka’tarohkwi, otherwise colonially known as Kingston, on the homelands of the Haudenosaunee, Anishnaabe and Huron-Wendat peoples and that this place is also home to many members of the Metis nation. I understand that this land is subject to the Dish with One Spoon Wampum, a covenant between the Haudenosaunee Confederacy, the Three Fires Confederacy, and other allied nations, including settlers, to peaceably share and care for the land and resources in and around the Great Lakes.

As a white, settler of European ancestry, I understand it is incumbent on me and others like me to recognize the ways in which we are complicit in upholding the status quo of the colonial state, and that we must strive to unlearn behaviors and mindsets, and to uphold the sovereignty of our Indigenous neighbours.

Digital spaces also carry their own histories of harm towards Indigenous peoples, and that technology made under settler-capitalism is never neutral. It is imperative to have critical conversations, to interrogate and to hold accountable our use of digital tools and in the creation of digital spaces. Indeed, we are currently meeting in a digital space (Zoom) that is a helpful tool for facilitating conversations, gatherings and organizing across geographies. However, the technology and infrastructures that Zoom--and upon which the Internet itself--depends on have inherent biases and are built from resource extraction, the appropriation of land and the displacement of Indigenous peoples across Turtle Island.

There are many Indigenous and non-Indigenous artists, scholars, designers and technologists alike that are dreaming and working towards digital and technological futurities that are led and informed by more expansive epistemologies that are grounded in an ethics of care and of multiplicity. As the series develops, I am hoping that each *Agnes Byte* can engage with this mindset to think across worldviews, across disciplines and to think alongside artists to imagine digital futurities that are liberatory, communally responsive and expansively inclusive.

And on that note, I am pleased to host our event today on the theme of "Archiving the Web," where we will discuss how fragile and impermanent websites (and domains) can be, and how web archiving can help to preserve the Internet as we know it and keep it as a resource for social, political and cultural memory and accountability for future generations. This feels particularly important in these days of misinformation and of increasing authoritarianism.

Through their respective practices and experiences working with community-based web archives, our two speakers will touch on the questions: What can we do with web archives when access is provided? And what can we learn from the Internet of the past in relation to the Internet of the present?

So, without further ado, I am thrilled to introduce:

Kaloyan Kolev, a digital artist and researcher based between Bulgaria and New York. His work explores internet culture, nostalgia, and technology in the Balkans, examining how digital tools and online spaces shape collective memory. Working across writing, sound, creative coding, and performance, he tells the stories of defunct websites, obsolete software, and lost music.

And Tessa Walsh, a digital preservationist, archivist, and software developer based in Montreal. Since 2022, she has worked as a senior developer at Webrecorder, where she works on open-source web archiving tools for all, including Browsertrix. She was previously a software developer at Artefactual Systems on the Archivematica and AtoM projects. Prior to switching to software development, Tessa founded digital preservation programs at the Canadian Centre for Architecture and Concordia University Library, both in Montreal. Her research interests include web archiving, environmental sustainability of digital preservation, and preservation of complex objects such digital design records.

Before I hand things off to Kalo to start things off, there's a couple of housekeeping items:

1. Following this conversation, there'll be time for a question-and-answer period. So please save your questions until the end.
2. I'd like to also mention that the session is being recorded, and it's being recorded in speaker mode. So only those of you who speak will be recorded.
3. And finally, I ask that you keep your microphones muted during the presentations, unless you're asking a question during the Q+A period.
4. If you don't feel comfortable being in the recording, you're welcome to keep your video off and/or to submit your question in the Chat Box and I'll read it out loud on your behalf.

Okay. So, I would now like to invite Kalo to start sharing their screen and to kick things off.

Kaloyan Kolev: Hi everyone, I'm Kaloyan Kolev. I was born and raised in Bulgaria. And looking back to how I grew up on the internet, felt much more local for me than it is now. I would check my email on abv.bg. I would watch videos on vbox7.bg. I would play games on kefche.bg, and I'd use bgmaps.bg for navigation and sinoptik.bg for weather.

Bulgaria, and the Balkans as a whole, were early adopters of computers and the Internet. Young programmers in the late '90s and early 2000s had built up a vibrant ecosystem of homegrown directory sites and hosting platforms, and many of them offered free sub-hosting on their servers, sparking an explosion of personal sites, internet art, and alternative culture. Around the 2010s, a lot of these local platforms and these servers started to fade away. Their tech stacks were ageing, they couldn't compete with Facebook and Google, and gradually this Bulgarian Internet began to converge with the rest of the world. Suddenly, Bulgarians were spending their time online in the same places as everyone else. There were a ton of fascinating forums and alternative art communities that kind of disappeared when everyone moved to Facebook groups. And a lot of the people who used to run these sites, they're still around, but maybe because of the lack of institutional support, many have moved on. Maybe they've gotten tired; they've exhausted their resources and energy for maintenance.

Bulgaria still has a very Facebook-heavy internet culture to this day, which makes me very anxious about how much of its post-2010s internet culture will get preserved. So, a lot of the sites you see listed on screen would offer free hosting. They would offer sites under their subdomain. And most of them today are either completely offline or they've gone through so many restructurings that they're completely unrecognizable. I know about them because they've been partially preserved in the Internet Archive—almost by accident. Unlike Western sites like GeoCities, which is probably the most high-profile case of a deleted platform, and one that's been heavily documented, there's very little written about these Bulgarian sites in secondary sources. So, over the last five years, I've been digging through the archives and seeing how these websites could tell their own stories. And I'll quickly show you two examples of sites on these platforms. They're hosted on CodeBG that I have been particularly fascinated by. The first one is bipiem.org. They call themselves "A zone for hyper culture," and they merged forum discussions, image sharing. They had an e-zine. It was very edgy. The web design is so beautiful. It used a lot of Flash, a lot of PHP, and so the archives from this community on the Internet Archive site are kind of semi-broken. You can see a lot of glimpses of some really interesting things happening, and a lot of pages that just don't work. A similar example is experement.org, and this is a so-called network buffer for contemporary art, and it hosted super interesting electronic music and alternative art. And it's like a feed and you can read the descriptions of music from all over the world that people are posting on here. And, some of the downloads, most of the downloads don't work. Some of the things you can find elsewhere; some of the things have just gotten lost.

So, when I think about websites, I really like this essay by writer Laura Maw in the now defunct *Real Life* magazine. I think she does a really nice comparison between neglected architecture and neglected websites. Just like buildings and monuments, websites require constant care to survive, so their existence depends on maintenance. Someone must pay for the server costs. Someone must renew the domain names, and in practice this makes the web a very impermanent medium. So, in this essay, Maw says that fewer than 200 million out of the 1.1 billion websites online today are active, and she uses this term for the rest of them that I really like, she calls them "mostly dead websites." So, mostly dead websites are accessible to the public, but they're no longer maintained. So, just like decaying shopping malls, they were victims of capitalism, victims of changing technologies, of unsustainable business models, and dwindling ad revenue even. *Real Life* magazine, I guess, is an example of a mostly dead website, so it's no longer active. They were funded by Snap, and a couple years ago, Snap stopped the funding, and so the website is still alive, but who knows for how much more. So unlike abandoned buildings, which remain parts of the landscape, this is what Laura Maw says: "Abandoned websites must be visited or stumbled upon. They lurk on the periphery, and you have to remember they exist, and this invisibility makes them very easy to forget." And I think a lot of Bulgarian websites that I showed you maybe suffered from this fate, where people moved on and forgot.

In the Balkans, we love to forget our recent history and to not know what to do with the artifacts of the past. There's this tendency to want to wipe the slate clean, to destroy monuments, and to return to some other greater past, but not the recent past. And this maybe makes sense given the trauma from regime changes, from wars, from closed and shifting borders. On the slide here, you see a photo from the mausoleum of Georgi Dimitrov in downtown Sofia, which was torn down in 1999. One of my favourite Bulgarian writers, Dimiter Kenarov, has this beautiful essay about the event, in which he describes this whole phenomenon quite poetically. He says, “Bulgaria has never felt quite comfortable in its own skin; it has never quite accepted the fact that its history—like any country's history—is messy.” So, when I'm thinking about all of this in regards to websites, to what we might call Bulgarian Internet culture, I tend to get a little sad and a little dumar (sp), because when you're walking around Sofia, you encounter history discarded as trash everywhere. And in a place like this, where even physical heritage can't last, who cares about digital memory?

This is a paper stack sculpture by Serbian artist Aleksandra Domanović. It's distributed in the form of a several thousand-page PDF file. When you print it on standard office paper, using the borderless printing setting, and stack the papers vertically, the ink on the edges reveals images printed on the sides. This sculpture shows smoke from flares set off by the extremist fans of Serbian club Partizan, called Grobari. These fans played a large role in diplomatic conflicts during the dissolution of Yugoslavia, and they continue to do so in present-day Serbia. Domanović has made several of these sculptures, and she calls them monuments to the .yu, internet domain name. .yu was a domain name assigned for Yugoslavia, and it was retired in March 2010, and it outlasted the country it was intended for by nearly three decades. So, the Rhizome Net Art Anthology talks about these paper sculptures by Domanović, and this is what they say. "These works meditate on the role of the digital and the formation and fragmentation of national identity, while drawing attention to the labour and bureaucratic effort required to make such constructs coherent. They also serve as proposals for monuments to the digital traces of a nation that no longer exists."

So, when I first discovered this project three years ago, something clicked for me. I think all these questions that I was just mentioning, that I keep asking about the web, about “who gets to remember?,” they come here in a really interesting way. And, it also led me down this really big rabbit hole that I'll tell you briefly about now, about country code domain names, about how borders are encoded into the fabric of the Internet, and particularly about the .yu domain and the crazy story about the fate of the internet of Yugoslavia. So, I've written about the story of .yu in detail, and there's also a fantastic documentary by Domanović on the subject titled, *From .yu to me*, so you can watch that. But I wanted to give you a high-level version here. The story of .yu, it starts in 1989 when computer scientists Borka Jerman Blažič and her team in Ljubljana decide to connect Yugoslavia to the internet. So, after two years of fighting bureaucratic, political, and technical challenges, just a couple of months before the connection was set to go in live in 1991, Slovenia declares independence, and Yugoslavia starts to fall apart. So, when a country no longer exists, that two-letter domain name is supposed to get retired. So, the life of .yu should have ended when Slovenia declared independence and Yugoslavia started to fall apart. So, when Slovenia gained independence and joined the United Nations a year later in 1992, it received a new domain, .si. But the old domain stuck around, and it was eventually transferred to Serbian scientists who managed it until 2006. And then after 2006, when Montenegro declared independence from Serbia, and the two states also got their new separate domains, .rs and .me, .yu finally felt obsolete and ICANN, the organization which manages domains, decided it's time to retire it. So, in 2010, after a number of bureaucratic delays, the domain was finally shut down. And when it was shut down, it was like the internet was effectively rerouted around it. You could type out .yu website and it just wouldn't work. So, with this shutdown, over 4,000 websites, some of the earliest examples of internet culture from the region, suddenly became inaccessible via their original address. And as Domanović says with these paper sculptures, the deletion of .yu represented the final loss of the former country, the erasure of its digital identity.

The Internet Archive has preserved some new websites, and I wanted to show you two of my favourites that I found while digging through the archives. It's a little difficult to find these websites, just because you have to know the name of a website most times to look for it in the Wayback Machine. But, I was digging through the archives of the Belgrade University servers, and I found these two that I really like. This 1996 page, which reports on the student protests, and it seems the student protest ended in victory. And it's a real historical document, and it also has a really cute design. And then this page from this user's personal website that talks about a really tender dream that they had.

So, as I said, it's hard to search for .yu pages on the Internet Archive, but there's some really good scholarship on the subject that reconstruct the domain and talk about how it changed over the years. And if you ask me for my take on the whole story, I think country code domains are a little weird, and in hindsight, I'm not sure they were the best idea. They have essentially baked borders into the internet. So, even if ICANN, the protocols and the UN, they claim to be neutral, but there's always political implications. And I think there's a strong argument that you could make in favour of keeping .yu alive. After all, the domain contained first-hand accounts of the NATO bombings, of the Kosovo War, mailing lists from scientists where they updated their colleagues in the West on how the war was going. There's nostalgic forums and weird experiments. And .yu could have become a hub for collective memory for Yugoslavia, some sort of memorial domain. But the unstable borders and the lack of a preservation-focussed policy from ICANN, prevented the domain from becoming a sanctuary. And this issue of deleting country domain names, I think Yugoslavia maybe was a little too early for it. But my prediction is that we'll be hearing more and more about questions like these, especially because of global warming and all the threatened island states. For example, Tuvalu owns the lucrative domain .tv, and they're extremely vulnerable to shifting borders and shifting institutions due to global warming.

So, the story gets even juicier. On the other hand, there's the domain .su for the Soviet Union. And that one is still active today, and it's managed by Russian organizations that have found loopholes around ICANN's termination proposals. The .su domain has become both a symbol of Russia's war narrative and a haven for cyber criminals, and it shows kind of how keeping these defunct country domains might not be the best idea. So if yu had been kept alive, sometimes I think maybe it would have been used as a propaganda machine by people like the Grobari shown in Domanović's paper sculpture, and not as some utopian museum that I keep dreaming about. This blog, by the way, the *Domain Insight*, is really fun to read, and they have really good coverage on all sorts of domain news, including the .su question, which has been going on for decades at this point, and it seems .su will be around for a lot longer. And it is alive because there's a stakeholder that's interested in keeping it alive, and there's people who want to keep it going.

So, my takeaway from all these stories that I told you is that everything online maybe feels a little too fragile, and everyone forgets about websites because they probably have more important things to worry about. And even at the highest institutional levels, there's no guarantee that things will end up being fair. The internet's infrastructure is deeply entangled with questions of national identity, of historical preservation, and who gets to control our digital past. And, unfortunately, I often feel a little bit powerless when I watch platforms consolidate and shut down, and I lose data and culture that I thought maybe I could revisit it one day to remember. And I love the Internet Archive and their utopian mission, but I'm worried that, as we see with Bulgarian websites, it has blind spots, that one day maybe -- it's a single point of failure, it could go down. And maybe there's something genetically Bulgarian in me that makes me believe less in institutions or just feel like I need to be doing more about this issue myself, because I don't know if a library will come and save my things for me.

So, to end, I wanted to share some naive ideas that I've been thinking about. What if we all watched over our favourite websites? Websites and archives die because people stop caring for them. So beyond any technical solution or innovation, the main goal as archivists, as preservationists, as educators, should be to get just a lot more people to care about the web, and to give them access to the past, to talk about it more, and to find ways to show it in ways that make it carry the same fidelity, the same weight, and life as the present. I made this diagram, it's a little goofy, and this kind of shows -- like, at the top is our daily lives, and then that kind of interfaces with our personal memories, our collective memories, and then the archives at the very bottom. And for the lack of a better word, I've been thinking about this question mark thing that sits between an archive and our daily lives, and I've been calling it a “wrapper for an archive,” for lack of a better word. So, some tool or intervention that takes the data and brings it back to the present in a way that more people can engage with. So that can be an amateur curator project. It could be an exhibition or even an Instagram account, a Tumblr account. It can be a tool or a browser extension. It can be a bot. It's basically anything within the realm of art. It can be your own personal archive where you download the sites you care about; you share them with the world. It can be a novel search engine. I think if you're an artist, programmer or writer, there's so much unexplored material on the web to play with. And there's so many great examples of projects like this, and I can't mention them all. But I think some of the most important work in this realm is being done by organizations like Rhizome, Webrecorder, and the Internet Archive, and I think the deletion of GeoCities, which I mentioned earlier, has inspired many really good examples that we can follow. So, here's three quick ones. I have to mention artists and archivists Olia Lialina and Dragan Espenschied, whose entire practice has been hugely influential in this field. This is one of my favourite projects of theirs. It's a Tumblr bot which automatically generates screenshots of GeoCities pages every 20 minutes using period appropriate browsers. I think I really like the temporal aspect of this project. It's been going since 2013 and to this day it keeps uncovering real gems from these archives. And I like the idea of bots that pick things at random from an archive and that you can set up and keep running. It feels like one of the few ways for the past to coexist with the present and to keep up with the algorithmic feeds and this need that -- when we're on Instagram, Twitter, TikTok, we need to see only new stuff. GifCities.org is another project from the Internet Archive that I use like every week. It's a visual search engine that lets you browse through millions of GIFs from GeoCities. I think images are just very powerful. They're very informationally dense and so it helps you discover stuff much faster. And same thing here, Cameron's World. This is a web collage made by artist Cameron Askin which arranges GIFs, texts, and elements from archived GeoCities pages into thematic rows for infinite exploration.

And then I'm going to end you on this note about this other Bulgarian practice of necrologues. So, these are street obituaries that announce or commemorate the passing of a member of the local community. And if you go to a Bulgarian village, you'll see them everywhere on bulletin boards, in entryways of buildings, just on trees. And they are a simple A4 sheet of paper with a small photo of the deceased and it says their name and there's often a generic poem under. And I think they're so beautiful. The idea is that they keep the dead along with the living and as long as the paper doesn't fade and there's someone to stop and read it, the memory is kept alive. And so, this is a little tongue-in-cheek but I've been thinking about what a necrologue for a website would look like. Maybe it's an ASCII graphic I can put on my website, or post in a comment, or maybe a necrologue is just a direct embed of the archived version of the site because after all it's all still HTML. So, if we all started commemorating our deleted websites in this way, would they feel less dead? Would we feel like we have more ownership over them or like they belong to us more? I'll leave you on that note. Rest in peace bipiem.org. Thank you.

Danuta: Thank you so much, Kalo. Tessa, would you like to start your presentation?

Tessa: Yeah, absolutely. Just get set up here. Okay. You're seeing my screen?

Danuta: Yes.

Tessa: Great. Good afternoon, everybody. And thank you, Kalo, for that really interesting -- I was taking so many notes. I'm really looking forward to the Q&A after these presentations. I'm going to be speaking a little bit about what we can actually do with web archives in current practice, both like how you can create them yourself and also plant some seeds for how we can think about ameliorating some of the problems that Kalo was talking about and generally like improving the fragility of the web. Taking a step back, web archiving as a practice is a few decades on now. It really started in earnest in the 1990s with the development of the Internet Archive and web archiving programs at national libraries. These were a lot of the early drivers of web archiving kind of tools and practices, file formats, things like that. People are generally aware of the Internet Archive. I find less so of the national web archiving efforts. But in Europe and many other places there are a number of national libraries that archive, essentially, their entire country's domain on a regular basis, often as part of an extension of legal deposit laws, the same laws that would have copies of books that are published within a country archived within those national libraries. In the decades since, as I was saying, this is really -- web archiving has evolved into a widespread practice. You know, there are best practices. There are standards. There are international communities like the IAPC, which hold conferences and, you know, push for ISO standardization of file formats, things like that. And there are, these days, a number of open-source, free web-archiving tools, hosted web-archiving platforms, basically tools that you can easily use to go create web archives of websites that you care about without necessarily having to rely on these very centralized surfaces like the Internet Archive—although, I'm a big fan and I would encourage you to use that as well. At a high level, the way that web archives work is by recording the HTTP traffic between a client—so either a bot or your browser—and a web server. So, we basically request these websites with a request, the same way that any request for a website would work, and we archive both the response that we get, but often the parameters of that request as well. And we write that traffic into files, typically using the ISO standardized work file format. And then we can access these web archives later by basically using tools that replay content from the works to recreate the original website. So, we're not necessarily scraping everything exactly as it existed and then, you know, writing a directory that has all the HTML and everything, especially with the more JavaScript heavy and kind of app-heavy modern web. It turns out to be a lot better to use tools that can replay this content from works and kind of rewrite on the fly things like links so that they point to the archive instead of the live web.

I work for an organization called Webrecorder. Our motto is "Web Archiving for All", and what we do is we create and maintain free and open-source, web-archiving tools. And, in particular, we focus on creating high-fidelity web archives using browser-based tools. So, what that means, functionally, is that our tools either use your browser or browsers that we spin up to archive the web, rather than being the kind of older model of like a bot that is not a full browser environment and doesn't necessarily know how to execute JavaScript and all those kinds of things. And what that allows us to do is to create web archives that are virtually identical to what you as a user would see. We also developed the openly-specified WACZ file format, which stands for Web Archive Collection Zipped, which is an increasingly widely-used format, considered a preservation format by institutions like the Library of Congress and a few national libraries, and basically that wraps those ISO standardized work files with like indexes and page lists and all other kinds of things in, essentially, a specific version of a zip file, which means that your web archive can be one file that's easy to move around and share, and also it can be much more efficient to replay so that the actual use of the web archives for -- reloading a page, for example -- is much faster than with older models of web archiving.

We maintain a number of tools. Some of the ones I want to point out that I'm not going to talk about in more detail today, but that I would encourage you to look into, are Archive Webpage, which is our Chrome or Chromium-based browser extension to create archives as you browse the web. So, this will work with any Chromium-based browser like Chrome or Brave or Edge. And basically, it's just a browser extension that you press a record button as you're browsing, and it will record exactly your interaction with that website and create a web archive from it. We also have a tool called replayweb.page that's an in-browser replay engine for web archives. We embed this in a lot of our other tools, but it also exists on the web at replayweb.page. So if you have a work file or a WACZ file, you can go upload it there and view your web archive, and it's all done entirely within your browser. It doesn't send the data anywhere else. We also maintain a tool called pywb that is a Wayback Machine implementation. So, it's basically the only well-maintained, open-source implementation of what you would see on the Internet Archive, so a kind of calendar-based view for interacting with web archives that powers many national archives' public web programs.

The main tool that we focus on these days is called Browsertrix. This is an end-to-end web archiving system to create, schedule, manage, and share web archives with a kind of modern user-friendly UI. So, it allows you to capture and replay websites. And because we use browser-based tools and we have certain features like being able to configure and save a browser profile and then apply that to your crawl, Browsertrix is one of the only tools out there that can crawl behind paywalls, can crawl social media, can crawl interactive websites. We even have like interactive dashboards and maps. It also allows you to view the live progress of your crawls as they go. We have some kind of machine-assisted quality assurance tools, as well as features for curating and publishing web archive collections. This is what Browsertrix looks like, but I'm actually going to switch over to a very quick live demo here just to show you what it looks like.

So, this is Browsertrix. This is an organization that we've set up within Browsertrix just for demos. And if we want to create a crawl, the simplest thing we can do is we go create new and click on crawl workflow. This is basically the configuration for how you want your crawl to go -- like what sites you want it to visit, what kind of is in scope, what links it should follow or not, and so on. We're going to very quickly archive a few pages from the Agnes website. So put the URL in here. I pick what I want my crawl scope to be. If I did single page, it would just capture that first page. I can also do a list of pages. I can do anything on the site and all of its subdomains. For this demo, we're just going to do pages on the same domain. And let me make this a little bit bigger so it's easier for you to see. So, this is basically saying capture this whole website. There are a lot of options here, as you can see, but I'm just going to point out a few ones that might be useful for us. I'm going to set it to crawl only 20 pages because we want this to be quick since this is a demo. Under the page behaviour setting tab, we have a lot of options for what automated behaviours we want on the page. So, if we want it to automatically click on certain elements, this will just use auto-scroll, which will just scroll all pages as far as it can. Because we're using real browsers, we have settings for those real browsers. So, for example, if we wanted to crawl something behind a login or a paywall and we had configured a browser profile, we could select that here to use for this crawl. One thing I am going to do is I'm going to use a proxy based in Canada, based in Montreal, so that our web traffic going to the Agnes site is coming from a Canadian IP, which should help it not get blocked. We'll use two browser windows for this crawl. And then I think that's it. We're going to use just the default settings here, although there are other options like scheduling a crawl if you want it to run daily or monthly, something like that. I'm going to click run crawl, and what's going to happen now is that in the background, it's spinning up instances of Browsertrix crawler, which is our free and open-source crawler that kind of underlies Browsertrix. You can also use that completely for free on your own. If you're familiar and comfortable with command line tools, I'd highly encourage you to check it out. It runs in a Docker container. So, in our Browsertrix application, it's spinning up these machines in the cloud on the fly that will run the crawler, and then once that starts, we'll see live what all of our real browser windows are doing. Usually, it takes about a minute to start up. Yeah, here we go. So, we've configured it to have a limit of 20 pages, so it's only going to crawl 20 pages. We can see our list of what's about to be crawled. If we didn't want any of these to be crawled, we could add exclusions on the fly to say that might be a crawler trap, or I don't care about that, so ignore it. And we can see exactly what our browsers are doing in real time, which really helps you get a sense of what's happening. So, this is probably going to take about one more minute, I would guess, so we can just kind of let it run. While it's running, I want to say like all of our tools, Browsertrix is free and open source, so if you're kind of like a sysadmin-y kind of person, or you're familiar with things like Docker and Kubernetes, you can spin this entire thing up yourself for free on your own infrastructure or your own personal computer. We also offer this as a hosted service, so that's what we're using here. We have plans for individuals and organizations. But I also want to point out that we have a seven-day free trial. So, I would encourage you, whether you think that you would want to use it on an ongoing basis or not, sign up for a free trial, try it, maybe cancel it before it turns into not a free trial. But if you want to get a sense of, like, how easy it is to create web archives, you can go and get started basically right now. So now our crawl is finished, and what we see is -- this is the replay of the website, and this is entirely powered from the web archive. So, there's nothing pulled from the live web here. And that should be pretty easy to see because if we try to go to a page that wasn't captured in our very small 20-page crawl, yeah, we'll get this message saying it's not found in the web archive. So in addition to seeing it in Browsertrix here, we could also publish this as a collection and have a public URL that we could point anyone to so that they could see our web archive. There are tools for being able to embed this web archive into other websites, and you can also get out your crawl files. So, if I click into the crawl here, go to WACZ files. See this is a list of the -- there might be more than one if we were running more browser windows or if we had paused and resumed the crawl, but for this crawl we just have a single WACZ file. I'm going to go ahead and download that. It should take just a second. Yeah, and then this WACZ file we could share with anyone. If we, you know, didn't want to publish something on the internet, but say it was like a family thing or you just want to share with a friend, you could just send them this file and they could go to replayweb.page on their own computer and load this file in and interact with the web archive there. The file format is also, like I said, just a specific format of a zip file. So just like any other zip file, we can unpack it and look at the data there. So, we have the logs for the crawl. We have indexes that help with the replay, as well as the actual work files that also contain things like extracted text and screenshots. So really our idea is that we want to make it as easy as possible for people to create web archives and then do whatever they want with them completely under their own control, which we think helps create a more decentralized web in terms of web archiving, where we don't have to rely on just one institution. Really anybody can do this, whether you're working as part of a community org or just for your own personal use or, you know, you're part of a library. Like, whatever makes sense to you.

I'm going to go back to the presentation for just a minute and I'm going to very quickly talk about what -- so what can we actually do with web archives? This is not an exhaustive list by any means, but I just want to plant some seeds for things that are currently possible so that maybe we can spark a little bit of imagination about new forms of what we can do with web archives that I haven't seen. So, one good example is we can embed web archives as sources to improve trust in news. So, this is an example of an AP News article from a few years ago that we worked on as a collaboration with the “Starling Lab for Data Integrity,” which is itself a collaboration between the USC Shoah Foundation and Stanford Engineering. And basically, as part of the journalistic process here, a web archive was created of this Twitter, with this tweet embed, and then the web archive was actually embedded into the news article. And it has this nice little banner that says, "This embed is part of a web archive. Click here to learn more." Says a little bit about what that means, allows you to download the web archive and has some information including a cryptographic signature, which can help prove that the archive is authentic. What this means is that if that tweet was deleted or changed, you know, this evidence will still be here in this news article. And there's also some additional information that allows people to interact with it themselves and have a little bit more faith that it actually is what it claims to be. We can preserve at-risk scientific data and websites. So, this is a screen grab of govarchive.us, which I would encourage you to take a look at. And this is basically selected archives from our Web Reporter US Government web archive that was created as a part of the End-of-Term Web Archiving Project in collaboration with the Common Crawl Foundation, Internet Archive, a number of university libraries in the United States, and a number of data rescue organizations, formal and informal, large and small. The idea was around the change of presidential administration. The partners of the End-of-Term Web Archive went and web archived a number of US government sites that hosted scientific data, information about LGBTQ protections, labour protections, things that have since been erased or modified. And we provide multiple ways for you to view these web archives so that that information is still easily and publicly accessible. And on this website, you'll notice there are two options for every selected web archive that kind of highlight other examples of what we can do. So one is we can share these as public collections in Browsertrix, where because we use Browsertrix to crawl these and create a public collection to make them public, anyone can go to app.browsertrix.com slash explore slash us gov archive and go click into and view all these collections there and even download the collection locally if they want to have their own copy. We can also publish mirrors of the deleted websites. So this is the other option from GovArchive, where for a number of these archives, we've also re-hosted the website at a slightly different domain with this nice banner saying, "You're viewing an archived version of this page so that you can interact with the website," not just as an embed or in Browsertrix, but really get the full original experience. So what we did here is -- because we registered govarchive.us, the original domain was cdc.gov. Basically, you can put in the original domain for a number of these US government sites and then just add archive.us, and it will take you to a mirror of the website as it was before it was deleted or modified for political reasons. The technology that we used to do this is all open source, and we have other people using it. So, you could also -- if you owned a website and you were like, I don't want to deal with these dependencies in this database anymore, you could create a web archive of that site. And since you still own that domain, you could even just transparently replace a website that's hard to maintain with a much easier to maintain web archive. There are a number of places that do this for things like simplifying maintenance of digital publications and digital humanities projects. So, one example that I like to use a lot is the Stanford University Press, who have a number of digital publications that have been created by scholars over the years. And for each of them, they've created a web archive so that you can go see the publication as it originally was. But there's also a web archive that was created initially as a fallback. And now I believe they're talking about even replacing the original publications with the web archives, since the fidelity is kind of the same, and it makes the maintenance of those things a lot easier to just be, basically, an HTML page with a script that points to a web archive file somewhere. And we can recreate older experiences of the web. So, this is oldweb.today, which is another project by Ilya Kreymer, our founder and lead engineer at Webrecorder. That was created when Webrecorder was actually a part of Rhizome, which Kalo had talked about before. And what this does is it uses in-browser emulation so that you can interact with websites from web archive captures from the Internet Archive, but within an emulation of the browsers that would have been used at the time. So, there's Netscape Navigator; there's Mosaic. There's early versions of Internet Explorer. Various of them have capability for Flash and things that we don't really have capability for in our modern browsers anymore. And this allows us to, not only see the web archive, but kind of see it in its original context, including like, you know, very slow browser load times and, you know, the kind of support for features that our browsers would have had at the time that those archives were created, or that the sites were live rather.

That's all for me. I want to make sure we still have time for Q&A. So I'll just say, if you want to get started or you want to learn more, go ahead and visit webrecorder.net. You can also feel free to contact me personally at tessa@webrecorder.net or tessa@webrecorder.org. If you're curious about what's possible, want more information, whatever, we'd be very happy to talk. Thank you.

Danuta Sierhuis: Amazing. Thank you so much, both Kalo and Tessa, for your presentations. That was such a great mix of kind of the conceptual ideas of web archives, and also the practicalities of, like, how to do this yourself and, you know, how you can use these archives today. I'd like to open it up for any questions. Feel free to unmute yourself or raise your hand or to put things into the chat and I can read them out for you.

Audience Member: Hello. I have a question for Tessa, and later, for Kaloyan. Tessa, you mentioned that there is a possibility for websites to be archived. I'm still a bit -- I'm actually super new to this, so I don't know anything. But, from what I understand, a website can be archived, and it can still be accessed and that would be kind of like easier to maintain because there is no maintaining required if you're using, like, what the Stanford University did with the papers and whatever. What I want to ask, I guess, is this going to become more widespread? If it's, like, a good solution to preserving websites? I imagine it would be better for websites which are not as interactive and are just basically there for information and data. That was my first question. Thank you.

Tessa: Yeah, thank you for the question. I hope it becomes a more widespread practice. We're certainly trying to evangelize around it because I think it's not a thing that most people have any idea is a possibility. And it's not to say that replacing the web archive makes it completely maintenance-free because basically the site, the new site, is a very simple HTML page that points to a JavaScript service worker that actually loads and replays the web archive, so that service worker still needs to be maintained, although that's a pretty core part of what we do and we use it a lot of places. But yeah, I think there are a lot of use cases where this could be really helpful. We have had some kind of preliminary talks with some hosting providers as, like, maybe this could be a cool service that hosting providers could provide for people that, like, when they've decided a site is --the content is done, it's not going to change anymore. Maybe the services that we're using to host our websites could even have this as a feature baked in for us. But yeah, I think in a lot of cases, any time where -- it doesn't have to be a simple website, like, because our tools will handle interactive JavaScript pretty well. But the main thing is, like, you want to make sure that it's not a site where content is continuing to be added, right? But yeah, right now there are a handful of examples. I really hope that in the coming years, we'll see a lot more of that because I think it offers a lot of promise.

Audience Member: I guess I would like to expand by asking, for example, if there would be a website that is about to be shut down because cost of maintaining is too big? Is it simple enough to just archive and leave it there as a work? Or I'm not sure exactly what the --

Tessa: Yeah. Generally, yes. The actual archiving of the website is in most cases pretty simple. If it's something that requires a lot of interactivity, sometimes you get to write kind of custom behaviours, but for the majority of websites, first attempt at crawling them is pretty good. I think the things that you would have to think about then are, you know, making sure that the domain continues to get paid for because, you know, if that goes away, then -- I mean, you can put it up at a different domain, you can put it up on GitHub pages or something, but ultimately you're still going to be reliant on somebody for that domain being up and kind of in your name. And at this point, it does still require somebody with, like, at least fairly minimal developer chops because we have, like, open-source tooling to be able to create this replacement site that uses the web archive. But, you know, it's definitely something we've used a handful of times, like everything else we've published at Open Source, but it's not quite, like, click and play yet. But if you had a developer as a friend at your institution, whatever, it's certainly something that doesn't take a lot of effort at this point. And it's also something that we, as Webrecorder, sometimes we'll work with people so that we can kind of lend our expertise to help them out.

Audience Member: Thank you. Does anyone want to ask a question before I ask a second one? Then I'll ask it. I want to ask Kaloyan about the Bulgarian web culture that is kind of disappearing, as he mentioned, in terms of websites being shut down. Vbox7 was hard for us. But what I want to ask is, we are Bulgarian online culture is a lot on Facebook and there's, like, many Facebook pages about everything. Is there, like, websites that have very clearly migrated to Facebook groups, where kind of the same interactivity and space is created? And can you kind of see traces of how that website worked or this online community and space was, I don't know, on a different platform like Facebook or, I don't know, Instagram? Of course, artists like you do work to kind of preserve that history. But my question is, I guess, is there, like, very clear examples of websites migrating to Facebook and kind of creating their hub there that you know of?

Kaloyan: I can't think of anything off the top of my head, but I think what you're asking is really important and it's really heartbreaking. I guess even going back to, like, 2014, it's so hard to search Facebook. It's so hard to go to a Facebook group and see things that have been posted a long time ago. Things get deleted. And I think Meta has been especially notorious for making it very difficult for any sort of scraping, any sort of archiving practice. I actually wanted to redirect the question a little bit to Tessa to kind of hear if Webrecorder has some Facebook solutions, or just kind of what are your thoughts on that particular platform, and is there work being done there -- whether it's the fact that, like, I'm so worried everything's going to just get lost.

Tessa: Yeah, it's a good question. Meta is, as you said, very hostile to people having any agency over what they create on Meta platforms. So, certainly, they don't want people archiving things using tools like ours. Nonetheless, people do. We're pretty successful being able to archive stuff on Facebook and Instagram, but it definitely takes a little bit of effort. I think that's one of our focuses over the winter and spring is to try to put together, like, guidebooks or easier tools in Browsertrix to be like, okay. How do I configure a crawl for Facebook so that it doesn't get blocked and it actually works? Because you have to -- They very aggressively rate limit. They very aggressively IP ban, so we often have to use proxies and things. You know, you have to be logged in for everything, and those logins expire just like a session cookie would expire in your own browser, so you have to refresh those over time. But we have partners that we work with who archive, like, thousands of social media posts a week. So, it's definitely possible, it just takes a lot of trial and error and kind of finesse at this point. And I'm hoping it'll get easier, but I'm also kind of aware that it's possible in part because we're such a small fish that they're probably not even aware of us. So, if we became bigger, we could be a victim of our own success in that way, where if we get on their radar, they might more aggressively try to block us. But I think that's also why it's really important to have decentralized tools, why it's really important that our tools are free and open source and you can run them on your own infrastructure from any IP, because even if they block our hosted service, that doesn't mean you can't use our tools on your own to go archive stuff still.

Danuta Sierhuis: Unfortunately, we are at time. So, I would just like to say thank you so, so much to everybody, to both Kalo and Tessa and to everybody that was able to join us on their lunch break. And this has just been such an incredible conversation. I'm glad this question about Facebook and social media got asked because I was going to ask it regardless, particularly in relation to Kalo's essay on the confiscation of digital memory, which is on their website. So, we can maybe put that into the accompanying PDF resource that will eventually be accompanying the recording of this talk that will go on Agnes's website in the coming weeks. So, and I also just wanted to quickly just say a thank you to our funders, the Canada Council for the Arts, and then also the George Taylor Richardson Memorial funding at Queen's University. And if you would lie to stay in touch and to learn about any other of the talks int eh Agnes Bites Series, I encourage you to follow our newsletter, sign up for that, and/or follow us on social media and we'll be able to get more information about this recording and the coming talks to you soon. So, thank you again everybody and I hope you all have a really wonderful afternoon. Thank you so much.