Barok, Dušan: For Artist and Community Web Archives. In: Reclaim your Archive! Online article series on reclaim.hypotheses.org, published by Art Doc Archive. Berlin, March 2023.

Abstract

Websites are notoriously fragile and archiving them can be a difficult task. They are closely tied not only to the community that runs them, but also to the infrastructure that hosts them. Creating web archives is becoming an increasingly popular activity, and there are many tools available for web archiving. However, hosting and maintaining web archives is still a relatively rare phenomenon. Many government archives and libraries are faced with the daunting task of ensuring the integrity of each archived website, and the agency of authors, artists, designers, builders and creators of websites can't be maintained or even considered when dealing with large numbers. There are a growing number of smaller-scale initiatives that provide a necessary, though certainly far from perfect, counterpoint. These include net art archives run by nonprofits, netizen initiatives such as the web archives of community servers, and citizen-activist initiatives to save the online cultural heritage of war-torn places. The article discusses some of the challenges of archiving websites using the example of Art Doc Web, a prototype archive of websites of artists based in Berlin.

Author

Dušan Barok's work is concerned with digital culture, memory studies, and activism. He is founding editor of Monoskop, a research platform for the arts and humanities. Together with Peter Gonda, he takes care of the Multiplace server, an independent infrastructure for artistic, cultural and social initiatives in central Europe and elsewhere.

art doc archive

Art Doc Archive is developing a prototype for the centralized archiving of the materials self-documented by Berlin-based artists on websites and in social media. The “Reclaim your Archive!” series of articles accompanies the project and is edited by Bianca Ludewig. All texts can be found on the project's blog: reclaim.hypotheses.org

Art Doc Archive was funded by:

Senatsverwaltung für Kultur und Europa | BERLIN

For Artist and Community Web Archives

Dušan Barok

Over the last thirty years, the web has become a central sphere of social contact. Websites are a fantastic medium for putting things out there into the world, situating them via hyperlinks, inviting participation and facilitating communication. They are alive, while they also store and remember, providing traces and records of our past. But they are also notoriously fragile. A missed domain renewal date, a hard drive failure, a discontinuation of service dependency such as a map or video platform for embedding media, or an update of the web host's infrastructure can render pages partially dysfunctional or make them disappear altogether. The average lifespan of a website is difficult to measure,[i] but we'd be probably not surprised to learn that websites don't usually last longer than three years.[ii] The newer is often viewed as the better so redesigns also often result in the purging of old data and content, because their value might not be immediately apparent. Institutions do it with the change of management, individuals on the occasions of career milestones. The relevance of old pages may later resurface however, when we look back and realise that websites are an important domain of how an organisation, or even a person, remember(s) themselves.

It is not a trivial task for authors and initiatives to keep track of their old webpages. Websites are closely tied not only to the community that runs them, but also to the infrastructure that hosts them and are not as portable as publications in general should be. Therefore, a key question when archiving a website is whether we want to reproduce the original, possibly obsolete infrastructure by means of virtualisation, together with the source code and data, or try to reproduce the website atop a different infrastructure. At the cost of more time and resources, the virtualisation option has the advantage of bringing us closer to the experience of the original site, but even this option can't fix dependencies on no longer supported APIs, RSS feeds and other live channels. The more affordable, one-size-fits-all approach of running a web archive infrastructure raises further questions. When converting an old website to a contemporary format, should we treat the source code and data(base) separately – to preserve the dynamic aspect – or should we combine code and data into static files? Should we retrieve and keep embedded media such as image, video, sound in the archive, or save space and keep only links, hoping that the external sources will last as long as possible?

Websites have a short lifespan, so prolonging access to them through archiving can be a virtue, but it is not easy to achieve. Every website is different. Idiosyncratic menus, structures and behaviours are the rule rather than the exception. Even archives like the Wayback Machine, which have improved their approaches over several decades, have many blind spots that come to the surface when an image or an entire page simply isn't preserved. And it's not just a matter of technology. Bringing disparate, previously unrelated websites onto a shared infrastructure also changes the context in which they are discovered, accessed and experienced. This 'neighbourhood' aspect is ultimately revealed through the archive's interface, inventory, metadata descriptors and search.[iii] There is, for example, a difference between viewing the archived version of a work such as *Female Extension* by Cornelia Sollfrank on the Wayback Machine and in the Net Art Anthology.[iv] Selection, after all, precedes appraisal and access in the archival process.[v]

The barriers to entry into web archiving have been steadily falling. The tools are out there.[vi] Among the most widely used web acquisition tools are heritrix, associated with the Internet Archive and affiliated initiatives, and browsertrix, initiated by Rhizome and developed by Ilya Kramer.[vii] Browsertrix is part of a wider suite of tools and packages aimed at preserving interactive websites in particular, including media-rich portals and social media profiles, called Webrecorder.[viii] Most of the crawlers and replay tools support established formats such as WARC and WACZ.

Creating web archives is becoming a folk activity, but hosting and maintaining them remains a relatively rare phenomenon. Like the Wayback Machine, web archiving initiatives by state archives and libraries face a similarly daunting task, where the cost of ensuring the integrity of each archived website individually is simply too high.[ix] The agency of authors, artists, designers, builders and creators of websites simply can't be maintained or even considered when dealing with such vast quantities. Websites are perceived and treated merely as data. This is worrying also for another reason. The web, however fragile, is not an innocent publishing medium. The centralisation of web presence into a handful of social media platforms and the advent of cloud computing have dragged the web into extractivist terrain, where divisive content is valued because it attracts interactions, which in turn feed advertising revenues.[x] Websites can surveil and they can also deceive and hurt. How can we ensure that archives don't perpetuate the violence that has gripped the web?

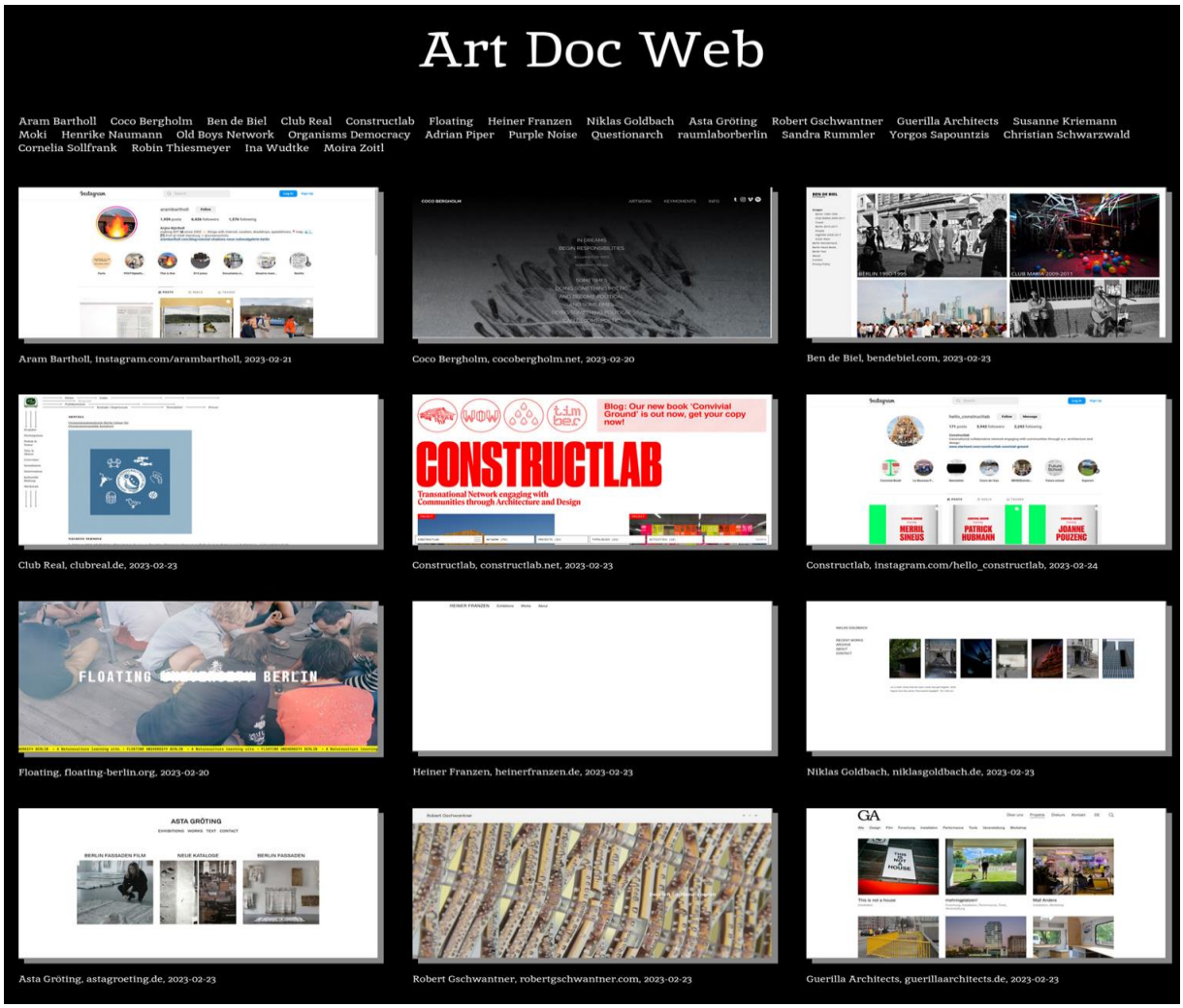
Infrastructures

There is a growing number of smaller-scale initiatives that provide a necessary, though certainly far from perfect, counterpoint. These include net art archives run by nonprofits such as Rhizome and Lima, netizen initiatives such as the web archives of the servers Servus.at and Anarchaserver, or citizen-activist initiatives such as Saving Ukrainian Cultural Heritage Online (SUCHO).[xi] The criteria for selection are specific to each project and not always disclosed or even defined, but many initiatives are tied to projects and initiatives in their physical location, to certain types of practices, or both. For example, the Linz-based Servus.at hosts the archived website of the late Austrian artist Armin Medosch, as well as many historical projects and initiatives.[xii] Anarchaserver "identify interesting communities, websites that are going to be closing soon, and try to make a copy of those websites to have them in the archive in a static form."[xiii] There is no imperative to provide live access to archived websites. Describing their web archive as a "feminist necrocemetery," Anarchaserver "have also the possibility to have [websites] in a zombie form, meaning that if people give us the copy of the database, we can host it for them. If one day other people want to bring these websites alive again, we can put them in contact so they can do it."[xiv]

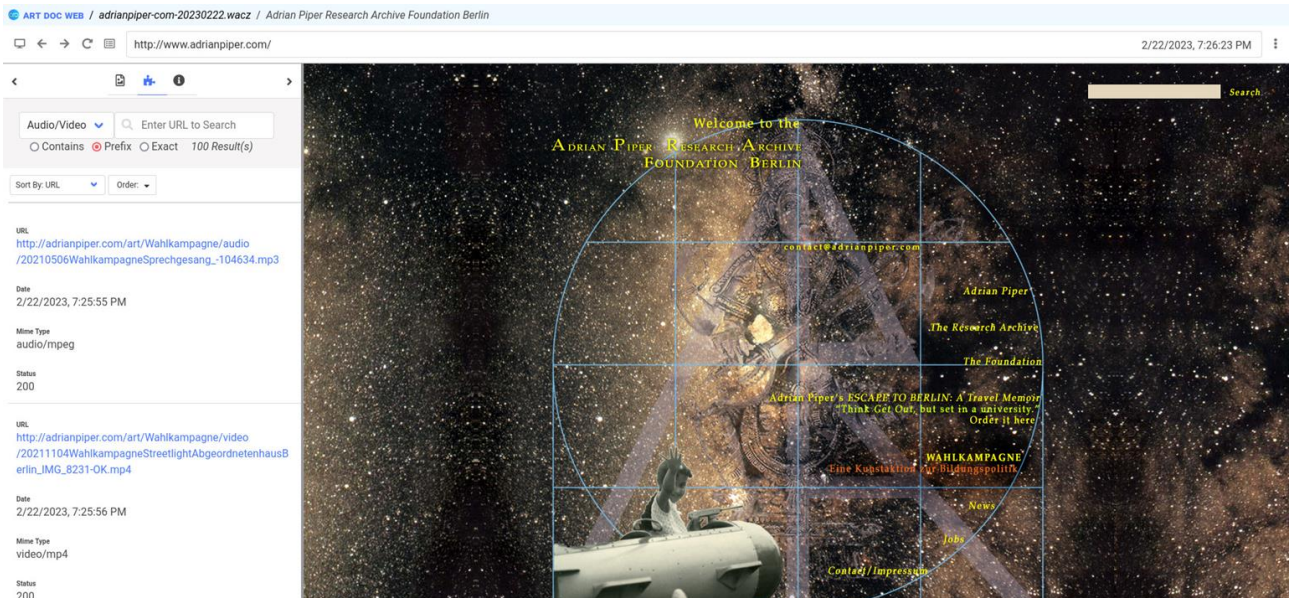
Both Servus.at and Anarchaserver can be described as community servers, a loose alliance of "interdependent, feminist, trans*feminist, free/libre, self-hosted, autonomous, collective, community and art servers"[xv] that aim to provide their social environments with means and tools for online presence and collaboration as well as acting as places to "learn system administration skills, host services and inspire others to do the same."[xvi] Articulated from a feminist perspective, these servers are "run for and by a community that cares enough for [them] in order to make [them] exist."[xvii] This attitude may also inform the criteria for selecting sites to archive.

Multiplace, the server I take care of together with Peter Gonda, subscribes to this alliance.[xviii] We understand the server as an independent infrastructure for artistic, cultural and social initiatives in Central Europe and elsewhere. The server has been providing web hosting, internet services and assistance since its launch on the occasion of Art's Birthday, an annual art-exchange event, when it was held on 17th January 2008. The NGO Multiplace created the server to support the wide range of organisations, artists and activists involved in its annual festival of media arts and culture. Over the years, the Multiplace server has gradually become the main focus of the organisation, going beyond its original scope as the coordinating body of the annual, distributed festival Multiplace. It is now home to over a hundred web domains of artists, writers, spaces, community radios, performance collectives, festivals, magazines, and human rights. The server provides storage, secure shell, mail administration, etherpads, mailing lists and streaming servers, and is committed to free software.

Building an archive



Screenshot of a section of the Art Doc Web archive, 2023, https://webarchive.multiplace.org/artdocweb/



Screenshot of an archived version of Adrian Piper's website, with a sidebar listing the audio and video files it contains

When we were approached to create a prototype web archive for Berlin-based artists, we accepted it as an interesting case study to explore what a web archive could be. The project curators selected a small number of live sites, mainly personal webpages and Instagram accounts of their acquaintances. We hadn't previously hosted any of the selected domains. The artists gave their explicit permission for their work to be preserved in the context of the project. The small scale and the given social setting meant we were not confronted with a mass of data, but rather found ourselves dealing with a collection of works by specific artists and protagonists, entering into relationships. Technically, our server harvested websites hosted on other servers, so these are machine-to-machine relationships, but with people on both ends, aware of each other, even if much remains unsaid, unarticulated and ambiguous. We could enter into a relationship with the artist and their website collaborators, a network of care, participating in the efforts of preservation, which in turn is intertwined with presentation, providing parallel access to their work.[xix]

After trying a number of free software crawlers, including httrack, heritrix and browsertrix, we settled on the latter two as the most appropriate for our context. Browsertrix with its long list of platform-specific crawling "behaviours" proved to be the best tool for archiving social media profiles. We built a simple static-page interface featuring the list of artists and screenshots of websites pointing to a locally hosted instance of replayweb.page to display the websites, and a JSON file listing all the archived sites in WACZ format.[xx]

Although the archive is now operational, it has raised multiple issues for future work.

Future work

The compact WARC and WACZ formats allow a complex website to be packaged into a single file, giving the impression that the site is portable. The original website was openly accessible, but how can we avoid falling into the trap of commodification, as happened with the EPUB format (which

is essentially a website packaged in a file and put on a marketplace)? How can we avoid the risk of degrading archived websites into training data for neural networks? Can we really talk about portability when liveness and external dependencies are cut off?

And not only those. Websites are taken out of their original context and domains change. What can be done to preserve context? Would it be more relevant to aim at creating live mirrors rather than archives? Or should we talk about online archives in terms of documentation rather than reproduction? How much modification of the original code and content is allowed? Can we modify archived websites to improve accessibility, for example by adding missing captions to images? Is it relevant to consider printing or web-to-print to improve accessibility?

What can we do to improve the searchability of web archives? Replayweb.page allows full text and file type searches of individual archived webpages. Would searching the whole archive be desirable?[xxi] And as for web-wide search engines, under what conditions should they be allowed in web archives?[xxii]

Another question is that of sustainability: The web is not exempt from environmental footprint and some websites are quite heavy in terms of size and resources used. Should we only preserve those below a certain ratio?[xxiii] How should this ratio be determined? And finally, what happens when the funded prototype phase is over? How do we distinguish between an experiment we are willing to spend our free time on and a service that involves work?

Web archiving tools are out there, there are many of them to choose from, align and think with. Web archives will always be messy. They thrive on multiplicity, redundancy, mirroring, interdependence. Yet they are also far too scarce and few considering the scale of the task at hand, perhaps we need a web ring to make them a little more visible… Ultimately however, they are always going to depend on communities taking care of their digital infrastructure to survive.

Endnotes

[i]Taylor, Nicholas (2011), "The Average Lifespan of a Webpage", *Library of Congress Blogs*, https://blogs.loc.gov/thesignal/2011/11/the-average-lifespan-of-a-webpage/.

[ii]Agata, Teru, et al. (2014), "Life span of web pages: A survey of 10 million pages collected in 2001", *IEEE/ACM Joint Conference on Digital Libraries*, https://doi.org/10.1109/JCDL.2014.6970226.

[iii]Barok, Dušan (2022), "Collection as a Network Volume", in *Networks of Care: Politiken des (Er)haltens und (Ent)sorgens*, eds. Anna Schäffler, Friederike Schäfer, and Nanne Buurman, Berlin: neue Gesellschaft für bildende Kunst (nGbK), pp. 59-61.

[iv]Compare http://web.archive.org/web/*/www.obn.org/femext/ and https://anthology.rhizome.org/female-extension.

[v]Summers, Ed (2019), "Appraisal Practices in Web Archives", *SocArXiv*, https://doi.org/10.31235/osf.io/75mjp.

[vi]https://github.com/iipc/awesome-web-archiving

[vii]https://github.com/internetarchive/heritrix3, https://github.com/webrecorder/browsertrix-crawler

[viii]https://webrecorder.net/

[ix]See, for example, the UK Government Web Archive at https://www.nationalarchives.gov.uk/webarchive/ and the Australian Web Archive at https://webarchive.nla.gov.au/.

[x]*International Trans★Feminist Digital Depletion Strike*, 2023, https://titipi.org/8m/.

[xi]See https://anthology.rhizome.org/, https://www.arthost.nl/, https://webarchiv.servus.at/, https://nekrocemetery.anarchaserver.org/, https://www.sucho.org/, respectively.

[xii]https://webarchiv.servus.at/arminmedosch.at/, https://webarchiv.servus.at

[xiii]*A Traversal Network of Feminist Servers*, 2023, p. 180, https://hub.vvvvvvaria.org/rosa/ATNOFS.

[xiv]Ibid.

[xv]https://monoskop.org/Community_servers

[xvi]https://systerserver.net/

[xvii]"A Feminist Server Manifesto 0.01", in *Are You Being Served? (Notebooks)*, Brussels: Constant, 2014, https://areyoubeingserved.constantvzw.org/Summit_afterlife.xhtml.

[xviii]Barok, Dušan (2020), "Artist-Run Servers and Community Work", De-platform-ization, Ethics and Alternative Social Media symposium, Prague: Artyčok, https://artycok.tv/en/post/de-platformizace-etika-a-alternativy-socialnich-medii-en#00e5667d-45a9-4de5-9687-1966e0b17894.

[xix]Schäffler, Anna, Friederike Schäfer, and Nanne Buurman, eds. (2022), *Networks of Care: Politiken des (Er)haltens und (Ent)sorgens*, Berlin: neue Gesellschaft für bildende Kunst (nGbK).

[xx]https://webarchive.multiplace.org/artdocweb/

[xxi]Mourão, André, and Daniel Gomes (2023), "Building a Web-Archive Image Search Service at Arquivo.pt", IIPC, https://netpreserveblog.wordpress.com/2023/01/09/building-a-web-archive-image-search-service-at-arquivo-pt/

[xxii]Jackson, Andrew N. (2023), "Letting search engines into the archive", 2023, https://anjackson.net/2023/03/09/letting-search-engines-into-the-archive/.

[xxiii]https://ecograder.com/