

Web archiving for Artists: Webrecorder Workshop

Artexzte (webinar) | December 13, 2018

Anna Perricci

Associate Director of Strategic Partnerships

Webrecorder at Rhizome

Rhizome at the New Museum

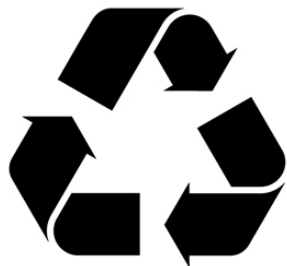
Anna.Perricci@rhizome.org

During this session participants will

- Learn how to employ user-friendly, open source, browser-based software for capturing web content and making collections accessible at no cost (Webrecorder.io)
- Gain information on options for downloading web archives (WARC files) and software for accessing web archives (Webrecorder Player)
- Find out how to describe, manage and share their web collections online (via Webrecorder.io)



Web archiving is a multi-step process



SAA
DAS
course

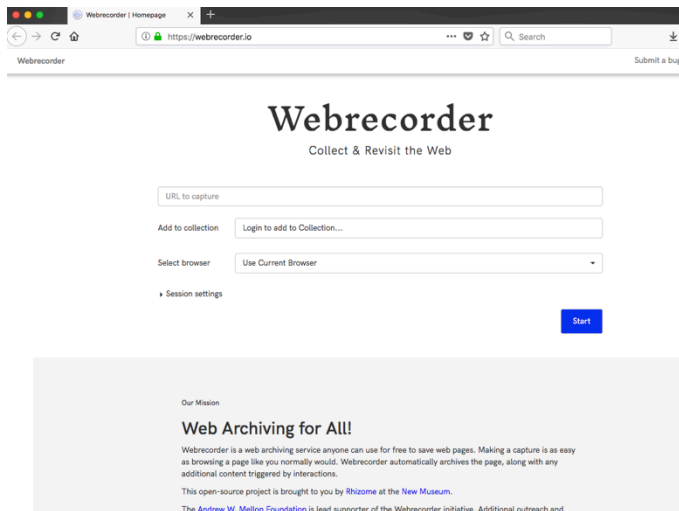
- Collection development and planning
- Selection
- Permissions
- Harvesting
- Description
- Access
- Long-term preservation

The slides for Web Archiving
Fundamentals can be accessed

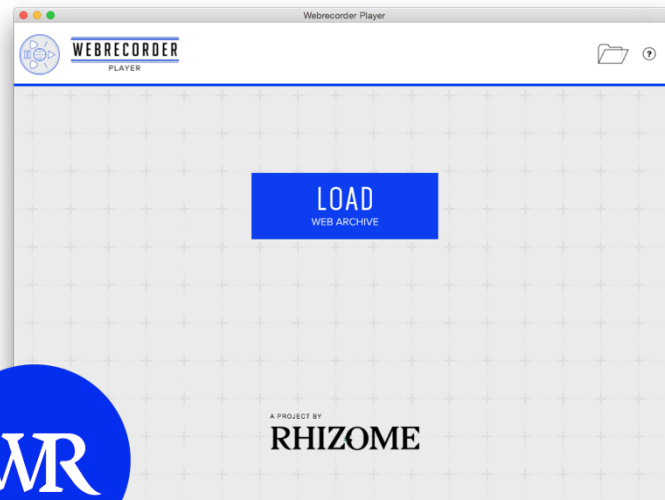
<https://www.slideshare.net/annaperricci/web-archiving-intro-circa-2015>

About Webrecorder

Create high-fidelity, interactive captures of any web pages you browse



<http://webrecorder.io>



Webrecorder Player App

Webrecorder Project

- Easy to use, robust tools
- Free to use
- Fully open source
- Using open standards
- Growing user community
- Quickly evolving

A project by

RHIZOME

with generous support from

THE
ANDREW W.

MELLON
FOUNDATION

Webrecorder Team



Dragan Espenschied
Preservation Director,
Rhizome



Ilya Kreymer
Software Director



Mark Beasley
Senior Front-End
Developer



Pat Shiu
Associate Director
of Design



Anna Perricci
Associate Director,
Strategic Partnerships



John Berlin
Senior Back-End
Developer

High fidelity web archiving

- Fidelity = similarity between original and capture
- Capture any web page loaded in the browser
- Archive interactive content (only available after user input)
- Same system for capture & browsing (web browser)

Webrecorder creates high fidelity web archives **including** elements that crawler based systems **often fail to capture** such as interactive content

Collecting at human scale

- Collecting is done with intent by a person via web browser one page at a time
- Can import and augment collections created by crawlers
- Anyone can use Webrecorder: web archiving for all!

The payoff for careful capture is an accurate representation of the original

Account login is **optional**

- One does not need to login to use Webrecorder to capture web content
 - Users can **download the captures** right away (as a WARC file) & **save them locally**
- For **continued access** to archived content online & to be able to **add to a collection**, one must **log in** to a free account

Capture / Browse

- Webrecorder.io is used to make interactive **captures** of web pages as users see them while collecting but is not a screen recording software that can play recordings back like a video
- **Browse** means you can access the content captured in the web archive and browse it interactively like the live web

Browsing a bound archive

- Each collection is a separate unit so at this time you can only navigate content within one collection at a time
- This gives tight curatorial control though the boundaries of the collection can sometimes be found quickly

Access & sharing options

- User created collections and lists can be kept private or made public through Webrecorder.io
- Public collections and public lists can be viewed by anyone
- Finer access controls are being considered
- Shared editing features are likely to be available in the future

Why is it hard to preserve quickly evolving websites (i.e. news)?

- News sites push the limits of current technology
- Interactive content!
- Lots of video/audio and ad content
- Traditional crawlers can't run Javascript
- Rapidly changing

Apps?

- Webrecorder is not suited for mobile apps, which are often web browsers customized for one or more particular website(s)
- Media rich database-driven websites, including ‘news apps’, are within Webrecorder’s domain though human scale web archiving does mean that capture can be time consuming

What about social media?

- Webrecorder can capture content from social media sites, and works especially well with Instagram and Twitter
- Some websites deliver content **individualized for each user**
 - Webrecorder can capture the content **you see when you are logged in to a social media profile**
 - *Webrecorder.io is designed to not capture or retain your social media login credentials*

Patching

Webrecorder / annaperricci / Steph Davidson's Excellent Design Work

Submit a bug

Browsing

Current Browser

https://www.nytimes.com/2015/05/03/magazine/what-happened-when-i-joked-about-the-president-of-ecuador.f

7/4

Section Navigator

Item lists

[Collection Index](#)

Steph Davidson's Excellent Design Work (94)

The Automation of Sectarianism: .

<https://marcowenjones.wordpress.com/2016/06/2>

Saudi Women Free After 73 Days

<https://www.nytimes.com/2015/02/13/world/mid>

Loujain Alhathloul

<http://www.loujainhathloul.com/>

Pro-Government Twitter Bots Try

<https://www.wired.com/2015/08/pro-government>

What Happened When I Joked At

<https://www.nytimes.com/2015/05/03/magazine/>

Packrat: Seven Years of a South A



Resource not Found

Try to patch

Still missing? Report a bug.

The url <http://radioambulante.org/audio/correa-vs-crudo> was not found in the archive.

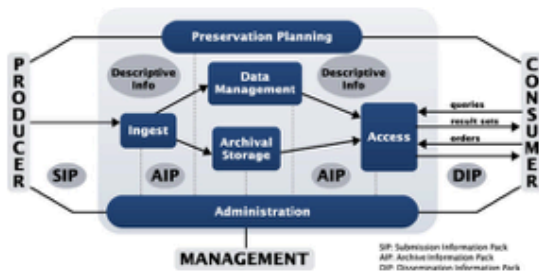
Sample collections

1. <https://webrecorder.io/demo/narrative-archives>
2. <https://webrecorder.io/imamuseum>
3. <https://webrecorder.io/ClarkArtLibrary>
4. <https://webrecorder.io/ivypluslibraries/state-elections-web-archive>
5. <https://webrecorder.io/sup/enchanting-the-desert>
6. <https://webrecorder.io/wrsc/snowfall---ny-times>
7. <https://webrecorder.io/demo/narrative-archives/list/flotus-x-td4w/b9/20170104093624/http://www.td4wbutton.com/>

Preconfigured (remote) browsers

- Using a preconfigured browsers to capture and browse web content **that may not be supported in future web browsers**
 - e.g. Flash
- Access with a preconfigured browser ensures **greater faithfulness to the original look and feel** of web pages
- Browsers use HTTP proxy mode = even better fidelity

Interactive Production : Preservation of Experience



SIP : WARC file grab by Webrecorder (Rhizome)

AIP : Archived using Atempo Digital Archive with all rules to ensure long term access

DIP : WARC with live index and playback with **Webrecorder** (Rhizome) in NFB infrastructure

Development Collaboration between NFB and Rhizome

Description, Management and Sharing

Narrative Archives

We've reached the twilight hours of Barack Obama's presidency.

A typical transition period requires many handovers, but this year, one of these is an altogether new endeavor for a White House: turning the sitting President's social media presence over to the new office holder. These Twitter, Instagram, Vine, Facebook, etc. accounts will be wiped and exported, some to archival accounts—for instance, [Obama's tweets will live on at @POTUS44](#)—and some to a public data archive.

On the occasion of this first social media transition, the White House issued a public call to action, asking Americans to try their hand at breathing life into the huge amount of data that is President Obama's eight years of social media presence. We were thrilled to respond. ([With others, too!](#)) The data archive, vast and disparate as it is, is largely unintelligible; what we proposed was using our built in-house [Webrecorder](#) tool to give it some narrative shape, some

Lists in this Collection

📖 FLOTUS x TD4W

Lil Jon and Michelle Obama's inadvertent collaboration on the First Lady's now-infamous "turnip for what" Vine.

[View list \(14 Bookmarks\) »](#)

📖 Thanks Obama

Exploring the satirization of conservative disappointment with Obama, mapping the *Thanks Obama* meme's progression from its earnest right-wing origins all the way to its self-aware usage by Obama himself.

[View list \(1 Bookmark\) »](#)

📖 LoveWins

Steph Davidson's Excellent Design Work

This collection contains pieces published on Bloomberg.

+ New Session



Collection Cover

Public

Pages (94)



LISTS (4 Public)

EDIT +

Title list for pieces in collection (6)



What is Code? (1)



How to Spend Your Bonus (1)



A Global Guide to State-Sponsored Troll... (1)



Pages

| INDEX | TIMESTAMP | PAGE TITLE | URL |
|-------|-----------------------|--|---|
| 1 | 7/27/2018, 1:19:24 PM | The Automation of Sectarianism: Are Twitt | https://marcowenjones.wordp |
| 2 | 7/27/2018, 1:13:49 PM | Saudi Women Free After 73 Days in Jail fo | https://www.nytimes.com/201 |
| 3 | 7/27/2018, 1:13:39 PM | Loujain Alhathloul | http://www.loujainhathloul.co |
| 4 | 7/27/2018, 1:11:33 PM | Pro-Government Twitter Bots Try to Hush | https://www.wired.com/2015, |
| 5 | 7/27/2018, 1:09:26 PM | What Happened When I Joked About the I | https://www.nytimes.com/201 |
| 6 | 7/27/2018, 1:09:20 PM | Packrat: Seven Years of a South American | https://citizenlab.ca/2015/12, |
| 7 | 7/27/2018, 1:08:07 PM | Ecuador Transparente - Godwin | https://ecuadortransparente.o |
| 8 | 7/27/2018, 1:07:06 PM | Amazon.com: Swati Chaturvedi: Books | https://www.amazon.com/s/re |
| 9 | 7/27/2018, 1:06:36 PM | I am a Troll: Inside the Secret World of the | https://www.amazon.com/Tro |
| 10 | 7/27/2018, 1:05:47 PM | Leaked Documents Show That Ethiopia's R | https://advox.globalvoices.org, |
| 11 | 7/27/2018, 1:05:26 PM | Amazon.com: Customer reviews: I am a Tr | https://www.amazon.com/Tro |
| 12 | 7/27/2018, 1:04:08 PM | I am a Troll: Inside the Secret World of the | https://www.amazon.com/Tro |
| 13 | 7/27/2018, 1:02:44 PM | Automated sectarianism and pro-Saudi prc | https://exposingtheinvisible.or |

+ New Session ...

Collection Cover Public

Pages (94) lock icon

LISTS (4 Public) EDIT +

- Title list for pieces in collection (6) ●
- What is Code? (1) ●
- How to Spend Your Bonus (1) ●
- A Global Guide to State-Sponsored Troll... (1) ●

2 7/27/2018, 1:13:49 PM Saudi Women Free After 73 Days in Jail to

3 7/27/2018, 1:13:39 PM Loujain Alhathloul

4 7/27/2018

5 7/27/2018

6 7/27/2018

7 7/27/2018

8 7/27/2018

9 7/27/2018

10 7/27/2018, 1:05:47 PM Leaked Documents Show That Ethiopia's R

11 7/27/2018, 1:05:26 PM Amazon.com: Customer reviews: I am a Tr

12 7/27/2018, 1:04:08 PM I am a Troll: Inside the Secret World of the

x New list| +

- Title list for pieces in collection
- What is Code?
- How to Spend Your Bonus
- A Global Guide to State-Sponsored Trolling

Done

Steph Davidson's Excellent Design Work

This collection contains pieces published on Bloom

+ New Session



Collection Cover

Public

Pages (94)

LISTS (4 Public)

EDIT +

Title list for pieces in collection (6)

What is Code? (1)

How to Spend Your Bonus (1)

A Global Guide to State-Sponsored Troll... (1)

INDEX

TIMESTAMP

PAGE TITLE

URL

7/27/2018 1:19:24 PM

The Automation of Sectarianism: Are Twitt https://marcove

Edit Collection Details

Collection Name:

Steph Davidson's Excellent Design Work

Collection Description:



This collection contains pieces published on Bloomberg.com that were illustrated by Steph Davidson. This is not a comprehensive collection of her work but these examples are quite excellent!

Cancel

Save

Steph Davidson's Excellent Design Work

This collection contains pieces published on Bloomi

+ New Session ...

Collection Cover

Public

Pages (94)

LISTS (4 Public)

EDIT +

- Title list for pieces in collection (6)
- What is Code? (1)
- How to Spend Your Bonus (1)
- A Global Guide to State-Sponsored Troll... (1)

Metadata

BOOKMARK 1 OF 1

Paul Ford: What Is Code? | Bloomberg


DESCRIPTION

B *I* U ~~S~~ 

Engaging and educational!

What is Code?

Show Description

| REM... | INDEX | TIMESTAMP | BOOKMARK TITLE | URL | CAPTURE B |
|--------|-------|------------------------|--|---|---|
| | 1 | 10/30/2017, 2:36:04 PM | Paul Ford: What Is Code? Bloomberg | https://www.bloomberg.com/graphics/2015-p |  Firefox v49 |

 COLLECTION by [annaperricci](#)

Steph Davidson's Excellent Design Work

This collection contains pieces published on Bloomberg

[+ New Session](#)

...

Collection Cover

 Public

 Pages (94)




LISTS (2 Public)

EDIT +

 Title list for pieces in collection (6)



 What is Code? (1)



 How to Spend Your Bonus (1)

PRIVATE 

 A Global Guide to State-Sponsored Troll... (1)



Steph Davidson's Excellent Design Work

This collection contains pieces published on Bloomberg

+ New Session



New Session

Cover

Manage Sessions

Helped To Collection

Download Collection

Delete Collection

Tour New Features

Help

Collection Cover

Pages (95)

LISTS (4 Public)

Title list for pie

What is Code?

How to Spend

A Global Guide

Pages

| INDEX | TIMESTAMP | PAGE TITLE |
|-------|-----------------------|---|
| 1 | 9/18/2018, 6:22:01 PM | Correa vs Crudo - Radio Amb |
| 2 | 7/27/2018, 1:19:24 PM | The Automation of Sectarianis |
| 3 | 7/27/2018, 1:13:49 PM | Saudi Women Free After 73 D |
| 4 | 7/27/2018, 1:13:39 PM | Loujain Alhathloul |
| 5 | 7/27/2018, 1:11:33 PM | Pro-Government Twitter Bots |
| 6 | 7/27/2018, 1:09:26 PM | What Happened When I Joker |
| 7 | 7/27/2018, 1:09:20 PM | Packrat: Seven Years of a Sout |
| 8 | 7/27/2018, 1:08:07 PM | Ecuador Transparente - Godw |
| 9 | 7/27/2018, 1:07:06 PM | Amazon.com: Swati Chaturver |
| 10 | 7/27/2018, 1:06:36 PM | I am a Troll: Inside the Secret |
| 11 | 7/27/2018, 1:05:47 PM | Leaked Documents Show That |
| 12 | 7/27/2018, 1:05:26 PM | Amazon.com: Customer review |
| 13 | 7/27/2018, 1:04:08 PM | I am a Troll: Inside the Secret |

MANAGE SESSIONS

Remove content from collection by **deleting sessions**. Add content by **uploading warcs**.

Please note that it is not possible to delete individual pages.

OVERVIEW

Steph Davidson's Excellent Design Work

Created on 10/30/2017, 6:35:47 PM

26 sessions over approximately **10months 23days 3hrs 47mins** containing **95 pages**. Total capture time is **2months 13days 34mins** and total data size is **501.91 MB**.



 Delete Entire Collection

 Download Collection as WARC

 Upload WARC to Collection

Sessions

Expand All

| Patch | 1 Pages | 1min | 3.66 MB |
|---|---|-------|---------|
| <div data-bbox="241 707 285 813">18 Sep TUE 2018</div> <div data-bbox="353 707 714 753">Session Notes Add notes about this session. Visible only to you.</div> <div data-bbox="546 773 595 789">EDIT</div> <div data-bbox="227 860 295 876">Delete</div> <div data-bbox="218 909 309 926">Download</div> | <div data-bbox="803 707 923 723">Session Pages (1)</div> <div data-bbox="817 745 1081 762">Correa vs Crudo - Radio Ambulante</div> <div data-bbox="1344 745 1721 762">http://radioambulante.org/audio/correa-vs-crudo2</div> | | |
| <div data-bbox="218 991 285 1007"> </div> Patch | 1 Pages | 2mins | 2.49 MB |

COLLECTION by [annaperricci](#)

_ Steph Davidson's Excellent Design Work

This collection contains pieces published on Bloom:

+ New Session



Collection Cover

New Session

Pages (95)

Cover

LISTS (4 Public)

Manage Sessions

Title list for pie

Upload To Collection

What is Code?

Download Collection

How to Spend

Delete Collection

A Global Guide

Tour New Features

Help

Pages

| INDEX | TIMESTAMP | PAGE TITLE |
|-------|-----------------------|-----------------------------------|
| 1 | 9/18/2018, 6:22:01 PM | Correa vs Crudi |
| 2 | 7/27/2018, 1:19:24 PM | The Automator |
| 3 | 7/27/2018, 1:13:49 PM | Saudi Women F |
| 4 | 7/27/2018, 1:13:39 PM | Loujain Alathlk |
| 5 | 7/27/2018, 1:11:33 PM | Pro-Governmer |
| 6 | 7/27/2018, 1:09:26 PM | What Happene |
| 7 | 7/27/2018, 1:09:20 PM | Packrat: Seven |
| 8 | 7/27/2018, 1:08:07 PM | Ecuador Transp |
| 9 | 7/27/2018, 1:07:06 PM | Amazon.com: S |
| 10 | 7/27/2018, 1:06:36 PM | I am a Troll: Ins |
| 11 | 7/27/2018, 1:05:47 PM | Leaked Docume |
| 12 | 7/27/2018, 1:05:26 PM | Amazon.com: C |
| 13 | 7/27/2018, 1:04:08 PM | I am a Troll: Ins |
| 14 | 7/27/2018, 1:02:44 PM | Automated cont |

Webrecorder Player

- Desktop application for OSX, Window and Linux
- User friendly application to browse any web archive (saved in standard WARC format)
- Can browse web archives offline, no internet connection required!

Download Webrecorder Player

<https://github.com/webrecorder/webrecorderplayer-electron>

Stewardship

- Storage (e.g. short or medium term) WARC files can be treated like files in any widely-adopted standardized file format
- As always any migration or dependencies can introduce risks
- Software that can open and make a WARC browse-able will be important to factor into planning for long-term access
- Workflows to get WARC files into repositories are being developed (including how to describe/assign metadata to these files)

What's next for Webrecorder?

- Subscription based services such as more storage space and automated collecting tools in 2019
- Implement sustainability plan through a multi faceted strategy for income generation while still always maintaining a fully functional, free, open source version
- More options for managing and sharing collections

Wrap up,
further discussions,
Q&A,
lab time?

You are welcome to contact us:

Anna.Perricci@rhizome.org
Support@Webrecorder.io

You can also contact Artexzte:

earthexzte@artexzte.ca

Webrecorder

Thank you

A project by

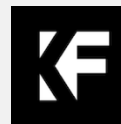
RHIZOME

with generous support from

THE
ANDREW W.

MELLON
FOUNDATION

additional outreach support



**KNIGHT
FOUNDATION**