

¡Web Archiving para todos!

Taller de Webrecorder

WET Labs | 15 de febrero de 2019

Anna Perricci

Directora Asociada de Colaboraciones Estratégicas

Webrecorder en Rhizome

Rhizome en the New Museum

Anna.Perricci@rhizome.org

Durante esta sesión, los participantes....

- Aprenderán a utilizar un programa gratis, fácil de utilizar y de software libre libre para recopilar contenido web (Webrecorder.io)
- Obtendrán información acerca de opciones para descargar archivos web (formato WARC) y programas para acceder a estos archivos (Webrecorder Player)
- Aprenderán a describir, manejar y compartir sus colecciones de archivos web

Archivar contenido web es un proceso de múltiples pasos



Certificación de
Especialista en
Archivos Digitales
(Sociedad
Americana de
Archiveros)

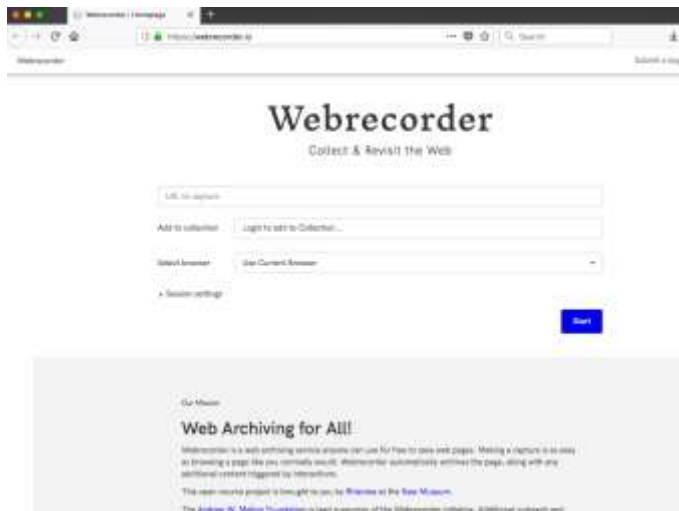
- Desarrollo y planificación de colecciones
- Selección
- Permisos
- Descripción
- Acceso
- Preservación a largo plazo

Puedes acceder a las diapositivas para el taller de introducción a los archivos de contenido web en el siguiente enlace: web archiving fundamentals en el siguiente enlace

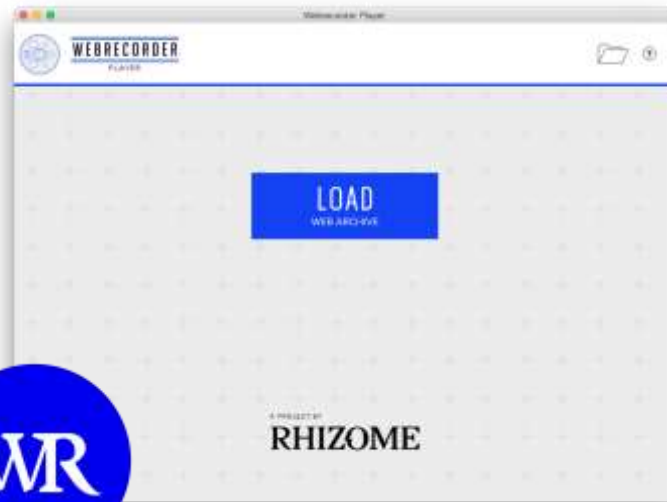
<https://www.slideshare.net/annaperricci/web-archiving-intro-circa-2015>

Acerca de Webrecorder

Crea capturas web interactivas y de alta fidelidad



<http://webrecorder.io>



Webrecorder Player App

Webrecorder: El proyecto

- Herramientas robustas y fáciles de usar
- Gratis
- Software libre
- Se basa en estándares abiertos
- Comunidad de usuarios en crecimiento
- Está desarrollándose rápidamente

A project by

RHIZOME

with generous support from

THE
ANDREW W.
MELLON
FOUNDATION

El equipo Webrecorder



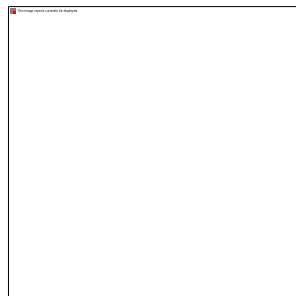
Dragan Espenschied
Director de Preservación,
Rhizome



Ilya Kreymer
Director de Software



Mark Beasley
Desarrollador de Interfaz
Externa



Pat Shiu
Directora Asociada de
Diseño



Anna Perricci
Directora Asociada de
Colaboraciones
Estratégicas



John Berlin
Desarrollador de
Sistema "Back-End"

Archivos web de alta fidelidad

- Fidelidad = semejanza entre el material original y la captura
- Captura cualquier página web en el navegador
- Archiva contenido interactivo (sólo con aportación del usuario)
- Utiliza el mismo sistema para navegar y capturar

Webrecorder crea archivos web de alta fidelidad e **incluye** elementos que a menudo los rastreadores web (“crawlers”) **no logran capturar**; por ejemplo, el contenido interactivo.

Recopilación a escala humana

- La recopilación de datos se lleva a cabo con intención, mediante un navegador y una página a la vez
- Cuenta con capacidades de importación y realce de colecciones creadas por rastreadores web
- Cualquiera puede utilizar Webrecorder. ¡Web archiving para todos!

La recompensa por una captura cuidadosa es una representación fiel al contenido original.

Registrarse es opcional

- No se necesita ser un usuario registrado para utilizar Webrecorder
 - Los usuarios pueden **descargar capturas** al instante (en formato WARC) y **salvarlas localmente**
- Sí se necesita crear una cuenta (gratis) tanto como para obtener **acceso continuo** a contenido archivado en línea además de crear colecciones

Captura / Navegación

- Webrecorder.io se utiliza para crear **capturas** interactivas de páginas web que reflejan lo que el usuario ve al acceder las páginas. Pero, Webrecorder no es un programa para grabar pantallas y reproducir las grabaciones como un video.
- **Navegar** significa que puedes acceder al contenido capturado en el archivo web e interactuar con el archivo del mismo modo que se interactúa con una página web.

Navegar colecciones separadas

- Cada colección es una unidad separada. De modo que sólo se puede navegar contenido dentro en una colección a la vez.
- Esto provee un control curatorial rígido.

Opciones de acceso y divulgación

- Las colecciones y listas creadas por usuarios pueden colocarse en **modo privado o público** por medio de Webrecorder.io
- Las listas y colecciones públicas pueden ser vistas por cualquier persona
 - Actualmente se están considerando otras opciones de acceso más sofisticadas
- También están siendo consideradas opciones de edición compartida

¿Por qué es difícil preservar las páginas web que están en cambio constante? (Ej. Noticieros)

- Algunos sitios web intentan superar los límites de la tecnología (noticieros, artistas)
- ¡Contenido interactivo!
- Mucho contenido sonoro, audiovisual y anuncios
- Rastreadores web tradicionales no son compatibles con Javascript
- Cambios constantes

Apps?

- Webrecorder no es compatible con aplicaciones móviles, las cuales suelen ser navegadores adaptados para alguna página web en particular.
- Webrecorder sí es compatible con bases de datos ricas en medios sonoros y audiovisuales, como páginas de noticias, pero la recopilación a escala humana significa que la captura requiere mucho tiempo.

¿Y las redes sociales?

- Webrecorder puede capturar contenido de redes sociales y funciona particularmente bien con Instagram y Twitter.
- Algunas páginas proveen al usuario contenido **individualizado**
 - Webrecorder puede capturar el contenido que ves **mientras estás conectado a tu perfil social**.
 - *Webrecorder.io no está diseñado para retener credenciales de inicio a tus redes sociales. Pero, puedes crear cuentas en redes sociales específicamente para recopilar archivos web.*

“Patching” (Parches)

ibrecorder / annaperricci / Steph Davidson's Excellent Design Work

Submit a bu

Browsing

Current Browser

https://www.nytimes.com/2015/05/03/magazine/what-happened-when-i-joked-about-the-president-of-ecuador.1

7/4

ection Navigator

l 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.

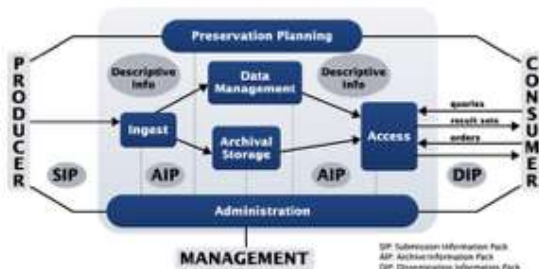
Ejemplos de colecciones

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/>

Navegadores pre-configurados (remotos)

- Utilizar un navegador pre-configurado para capturar y navegar contenido web que quizás **no sea compatible en versiones futuras de navegadores**
 - Ej. Flash
- Acceso a un navegador pre-configurado asegura mayor fidelidad a la apariencia de las páginas web.
- Navegadores que utilizan HTTP proxy = mayor fidelidad

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

Descripción, manejo, y divulgación

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 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)



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/20
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/20
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 Lista nueva| +

- 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

Detalles de la colección

Collection Cover

Public

Nombre de la colección

LISTS (4 Public)

EDIT +

Descripción de la colección

What is Code? (1)

How to Spend Your Bonus (1)

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

INDEX TIMESTAMP PAGE TITLE URL

1 7/27/2018 1:19:24 PM

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

Edit Collection Details

Collection Name:

Steph Davidson's Excellent Design Work

Collection Description:

B *I* U ~~S~~ 

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 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)

Metadata

BOOKMARK 1 OF 1

Paul Ford: What Is Code? | Bloomberg

DESCRIPTION



Engaging and educational

What is Code?

Show Description

REM.	INDEX	TIMESTAMP	BOOKMARK TITLE	URL	CAPTURE
	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 Bloomi

+ 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 Bloom

+ 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 Guid

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="353 707 710 754">Session Notes Add notes about this session. Visible only to you.</div> <div data-bbox="523 768 600 794">EDIT</div> <div data-bbox="227 856 297 875">Delete</div> <div data-bbox="218 904 309 923">Download</div>	<div data-bbox="803 707 919 727">Session Pages (1)</div> <div data-bbox="817 740 1081 760">Correa vs Crudo - Radio Ambulante</div> <div data-bbox="1344 740 1721 760">http://radioambulante.org/audio/correa-vs-crudo2</div>		
<div data-bbox="218 991 285 1011">Patch</div>	1 Pages	2mins	2.49 MB

COLLECTION by [annaperricci](#)
_ Steph Davidson's Excellent Design Work

This collection contains pieces published on Bloomi

+ 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 Alhathk
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

- Aplicación desktop para OSX, Windows y Linux
- Aplicación fácil de utilizar que te permite navegar archivos WARC
- Acceso a archivos sin conexión a internet

Descargar Webrecorder Player

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

Cuidado a largo plazo

- Almacenamientos (ej. a corto o mediano plazo) WARC pueden ser manejados como cualquier otro formato en estandarizado
- Cualquier tipo de migración o dependencias puede introducir riesgos
- Los programas que provean acceso a archivos WARC son un factor importante cuando se planifique el acceso a largo plazo
- Se están desarrollando “workflows” para depositar archivos WARC en repositorios (incluyendo cómo describir y asignar metadatos a estos archivos)

¿Qué es lo próximo para Webrecorder?

- Aquellos que apoyen el proyecto recibirán beneficios tales como más capacidad de almacenamiento y herramientas automatizadas (en el 2019)
- Implementar un plan de sustentabilidad por medio de una estrategia de múltiples facetas para generar ingresos a la vez que se mantenga la herramienta de software libre y gratis (con todas sus capacidades)
- Más opciones para manejar y compartir colecciones

¿Discusión,
preguntas,
ejercicios?

Escribenos a:

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

¡Gracias a los organizadores y traductores de estas diapositivas!



Traducción y coordinación (translation and coordination by) por Cristina Fontáñez Rodríguez (Archivistas en Espanglish)

cristinafontanez@gmail.com

Muchas gracias a Walter Forsberg (WET Labs) y Lorena Ramírez López

Webrecorder

Thank you

A project by

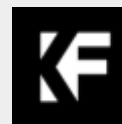
RHIZOME

with generous support from

THE
ANDREW W.

MELLON
FOUNDATION

additional outreach support



**KNIGHT
FOUNDATION**