Human Scale Web Collecting for Individuals & Institutions: Webrecorder Workshop

IIPC | November 14, 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 get experience with software for accessing web archives (Webrecorder Player)

• Leave with the ability to describe, manage and share their web collections online (via Webrecorder.io)
Part 1 (of 3)

- Introductions, participants share goals/anticipated use cases
- Overview of Webrecorder.io’s core tool sets followed by question time and brief discussion
- Hands on activity to sign up for account, capture sample collection and review sample collection
- Participants check back with group to discuss/reflect on progress
Web archiving is a multi-step process

- Collection development and planning
- Selection
- Permissions
- Harvesting
- Description
- Access
- Long-term preservation
The slides for Web Archiving Fundamentals can be accessed at

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
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
Hands on

1. Sign up for account via Webrecorder.io

2. Capture small sample collection (at least a few pages)

3. Review sample collection
Check back with group to discuss & reflect on progress
Part 2

- Prompts or questions given to help participants envision how they could form meaningful small scale collections or augmentations to larger collections; then start forming those collections (can be done individually or in groups of 2-3)
- Introduction to using pre-configured (remote) browsers and importing from public web archives
- Overview of description, management and sharing features followed by an exercise to use these tools
Sample collections

1. https://webrecorder.io/demo/narrative-archives
2. https://webrecorder.io/imamuseum
5. https://webrecorder.io/sup/enchanting-the-desert
To Do

1. Quick sketch of theme or purpose

2. Establish first URL(s) to capture (list more if applicable)

3. Go! Capture by browsing pages

4. Keep notes on process or if encounter errors
   (if errors seen, bug reports are very much appreciated)
Review capture
Patching
Feedback?
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**
<table>
<thead>
<tr>
<th>INDEX</th>
<th>TIMESTAMP</th>
<th>PAGE TITLE</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>7/27/2018, 1:11:33 PM</td>
<td>Pro-Government Twitter Bots Try to Hush</td>
<td><a href="https://www.wired.com/2015">https://www.wired.com/2015</a>,</td>
</tr>
<tr>
<td>6</td>
<td>7/27/2018, 1:09:20 PM</td>
<td>Packrat: Seven Years of a South American</td>
<td><a href="https://citizenlab.ca/2015,12">https://citizenlab.ca/2015,12</a>,</td>
</tr>
<tr>
<td>7</td>
<td>7/27/2018, 1:08:07 PM</td>
<td>Ecuador Transparente - Godwin</td>
<td><a href="https://ecuadortransparente.o">https://ecuadortransparente.o</a></td>
</tr>
<tr>
<td>13</td>
<td>7/27/2018, 1:02:44 PM</td>
<td>Automated sectarianism and pro-Saudi prc</td>
<td><a href="https://exposingtheinvisible.or">https://exposingtheinvisible.or</a></td>
</tr>
</tbody>
</table>
New list

- Title list for pieces in collection
- What is Code?
- How to Spend Your Bonus
- A Global Guide to State-Sponsored Trolling
Steph Davidson's Excellent Design Work

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!
What is Code?

Paul Ford: What is Code? | Bloomberg
https://www.bloomberg.com/graphics/2015-p
Firefox v49

Engaging and educational!
Steph Davidson's Excellent Design Work

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

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

Sessions

18 Sep Tue 2018

Session Notes
Add notes about this session. Visible only to you.

Session Pages (1)

Correa vs Crudo - Radio Ambulante
http://radioambulante.org/audio/correa-vs-crudo2

1 Pages
1 min
3.66 MB

Patch

1 Pages
2 mins
2.49 MB
Part 3

• Demo of Webrecorder Player; then collections created earlier in the workshop can be downloaded and viewed in Webrecorder Player

• Through discussion explore how components of the practice of web archiving could fit into existing and emerging digital preservation workflows
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
Augmenting existing web archives

NYARC’s capture in Archive-It of a page which the video should pop out from (doesn't play back from this location):

Archive-It Wayback link, where it does play back -- uploaded WARC of capture made with Webrecorder:
http://wayback.archive-it.org/4544/20180618160117/https:/fpdl.vimeocdn.com/vimeo-prod-skyfire-std-us/01/4375/9/246875750/892145210.mp4?token=1529341575-0x5e78ab4726ee291fd215abe5cf7bf89aaeccdfc7
Eve Abrams
New Orleans

Exposed is an audio documentary on Louisiana, the world's incarceration capital, history of the state's sky-high levels of incarceration, and a key hub in the nation's domestic when I'm 16 black men from New Orleans.

Eve, Eve Abrams is an audio documentary with stories hearing to her grandmother's on documenting the stories of her story's audio project. Documented, printed through NPR and WNYC, which tells five new stories.

Exposed received a Gabriele's 2019 documentary award through the Anne Frank Center and received the Edward R. Murrow Award for outstanding work in the category of investigative.

Exposed: writing can be found in the New Orleans Atlas Bookstore, Zephyr Publishing, and Louisiana State University. Her book works can be heard on a host of national programs such as Morning Edition and Newshour. She has been teaching for 35 years and currently teaches for the Society for the Relief of Destitute Orphan Boys.
How used at Library of Congress (USA)

- Primarily used to flesh out selected sites - things wanted but missed by crawler (i.e. due to scoping rules)
- Capturing video – can view video once it is captured in Webrecorder; video replays very nicely, which is hard to do in our huge selective crawls
- Verification of things really want to make sure are included
- Testing things that don’t replay well using Heritrix/Open Wayback
Stewardship

• Storage (e.g. short or medium term) WARCs 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 WARCs into repositories are being developed (including how to describe/assign metadata to these files)
Discussion: How does human scale web archiving fit into your institution’s workflow(s)?
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 me

Anna.Perricci@rhizome.org
@AnnaPerricci
Webrecorder

Thank you