

Web Collection Report

Assignment - Week 6, CS 891, Fall 2017

Plinio Vargas
Department of Computer Science
Old Dominion University
Norfolk, VA 23529
pvargas@cs.odu.edu



1 INTRODUCTION

The preservation of events has had a great deal of importance in every culture of human civilization. Prior to written form of communication, historic preservation was narrated from generation to generation using songs or stories to account for great battles that provided identity and a sense of unity to a particular social nucleus. With the appearance of written communication, the preservation of historical accounts became more accurate, providing future generations the ability to go into the past, thus making learning and connecting with their ancestors in a way that was not possible using oral communication. Today, the web presents to our current generation a vast amount of information that has never been seen before. It contains social aspects, political accounts, medical advances, etc. However, the information is so dynamic that its contents also disappears at a great speed. Archiving the web is a quest to collect current information appearing on the web for preservation. Our group learned of four tools useful for web archiving. This report is an evaluation to those tools, and my experience using them.

2 THE COLLECTION: LAS VEGAS MASSACRE 2017

At the time this assignment was given, the United States went through the deadliest mass shooting in its history. On October 1, 2017 a shooter opened fire onto a country music concert crowd from the 32nd floor of the Mandalay Bay hotel in Las Vegas. 58 people lost their lives in this incident. The 15 seeds for the collection are:

- <https://newrepublic.com/article/145192/alt-lite-celebrated-las-vegas-massacre>
- <https://www.rawstory.com/2017/10/trump-campaign-exploits-las-vegas-massacre-in-fundraising-email/>
- <http://www.mercurynews.com/2017/10/07/gun-control-debate-ask-las-vegas-massacre-survivors-where-they-stand/>
- <http://www.latimes.com/local/lanow/la-me-vegas-memorial-20171007-story.html>
- <https://www.reuters.com/article/us-lasvegas-shooting/pence-offers-solace-as-las-vegas-police-puzzle-over-shooters-motive-idUSKBN1CA0X6>
- <http://nypost.com/2017/10/07/hero-guards-in-las-vegas-massacre-are-back-to-work/>
- <http://abcnews.go.com/US/las-vegas-shooting-mother-father-dead/story?id=50229707>
- <https://www.yahoo.com/news/isis-las-vegas-shooter-stephen-004312350.html>
- <http://www.aljazeera.com/indepth/opinion/las-vegas-massacre-reveals-troublesome-global-truth-171006054512920.html>
- <http://www.dailynews.com/2017/10/07/family-of-canyon-country-man-slain-in-las-vegas-massacre-seeks-to-freeze-killers-assets/>
- https://www.washingtonpost.com/news/morning-mix/wp/2017/10/02/police-shut-down-part-of-las-vegas-strip-due-to-shooting/?utm_term=.1c25ebc3892e
- <https://www.cbsnews.com/news/victims-of-las-vegas-shooting-list-names-latest-update/>
- <http://abcnews.go.com/US/motive-remains-elusive-days-las-vegas-massacre/story?id=50320760>
- <http://www.cnn.com/2017/10/05/us/las-vegas-shooting-timeline/index.html>
- <http://www.cnn.com/2017/10/05/us/inside-the-las-vegas-massacre/index.html>

3 TOOLS

3.1 Archive-it.org

3.1.1 Advantages

- Archive-it.org has a “seed” feature where the user can drop a list of URLs to archive the collection. Here the user may determine the frequency on which the seed will be archived (One-time, daily, weekly, etc.); seed type, which determines if available links in the html page will also be archived (standard is the default); collection access type (public or private); if it provides information about the last time the seed was crawled, the number of captures and a link of a particular seed to the Wayback machine.

Home / Collections / Las Vegas massacre 2017 / Seeds

Las Vegas massacre 2017

Created: Oct 7, 2017 by pvargas Updated: Oct 10, 2017 by pvargas

Overview **Seeds** Crawls Collection Scope Metadata Wayback QA

Seed List (15 Seeds)

Type to Filter Seeds

[Download Seed List](#)

Run Crawl Edit Settings Add Metadata Edit Groups Delete Seeds [Manage Seed Groups](#) [Add Seeds](#)

Seed URL	Group	Status	Frequency	Type	Access	Last Crawl	Captures	Wayback
<input type="checkbox"/> https://newrepublic.com/article/145192/alt-lite-ce		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> https://www.rawstory.com/2017/10/trump-campi		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> http://www.mercurynews.com/2017/10/07/gun-ci		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> http://www.latimes.com/local/lanow/la-me-vegas		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> https://www.reuters.com/article/us-lasvegas-shoc		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> http://nypost.com/2017/10/07/hero-guards-in-las		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> http://abcnews.go.com/US/las-vegas-shooting-mc		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> https://www.yahoo.com/news/isis-las-vegas-shoo		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> http://www.aljazeera.com/indepth/opinion/las-ve		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> http://www.dailynews.com/2017/10/07/family-of-		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> https://www.washingtonpost.com/news/morning		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> https://www.cbsnews.com/news/victims-of-las-ve		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >
<input type="checkbox"/> http://abcnews.go.com/US/motive-remains-elusiv		Active	One-Time	Standard	Public	Oct 9, 2017	1	Wayback >

- The crawl feature tab is very useful. It provides information on the crawl status to include the starting and completion time, the frequency of the crawls, the collection size (in GB), and the number of documents that have been collected. I used this feature when I created my collection. I used the default **standard** seed type which crawls all the links found in a particular web page and consequent links of the newly discovered pages. As time passed, the number of documents kept growing, so did the collection size, so I decided to stop the crawl process manually.

Home / Collections / Las Vegas massacre 2017 / Crawls

Las Vegas massacre 2017

Created: Oct 7, 2017 by pvargas Updated: Oct 10, 2017 by pvargas

Select two crawls to compare.

Overview Seeds **Crawls** Collection Scope Metadata Wayback QA

Crawl Reports Current Crawls Test Crawls Crawl Frequency

Crawl Report List (1 Crawl Report)

Type to Filter Crawl Reports [Download Crawl Report List](#)

Crawl ID	Started	Completed	Status	Frequency	New Data	Docs
<input type="checkbox"/> 459120	Oct 9, 2017	Oct 10, 2017	Finished: Stopped by User	One-Time	39.2 GB	302,359

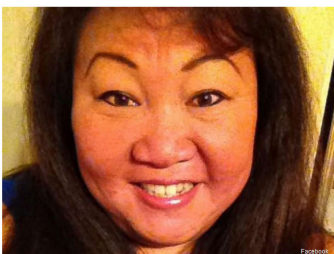
- Although, I did not use the **crawl frequency** feature, exploring the functionality reveals its importance. The previous point described how fast the collection size can increase if the wrong seed option is selected, so in the crawl frequency tab, the user has the ability to edit the crawl limits, including the number of documents, data and size limits.
- I made use of the metadata collection functionality. The user has the ability to upload a picture representative of the collection and provide some description related to its content.
- Archive-it.org appears to do a good job in preserving a time capsule of side articles relative to the time frame when the story was written, which is the desire effect. A snapshot of the archived page shows that during the time when the massacre occurred, there were also wild fires going on in California and the Vice-President Mike Pence in a PR stunt leaving an NFL game just after the national anthem was played.

abc NEWS U.S. International Politics Lifestyle Entertainment Virtual Reality ...

Slideshow: Shooting at Las Vegas concert

Here is what we know about the 58 victims of the massacre:

Laura Shipp



Laura Shipp, one of the people killed in Las Vegas after a gunman opened fire, Oct. 1, 2017, at a country music venue.

Las Vegas resident Laura Shipp, 50, died in the shooting, her brother, Steve Shipp.

Watch Live Shows

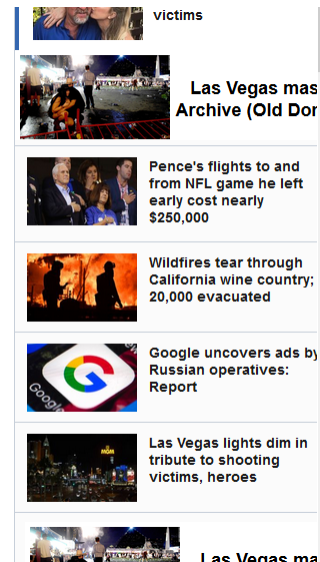
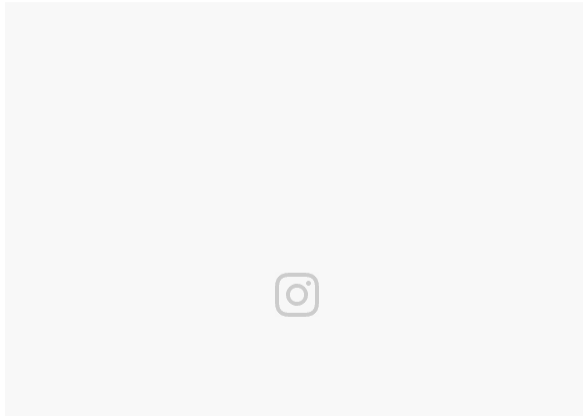
- Las Vegas massacre: Portraits of the 58 victims
- Las Vegas massacre Archive (Old Dominion)
- Pence's flight to and from NFL game he left early cost nearly \$200,000
- Wildfires tear through California wine country; 20,000 evacuated
- Google uncovers ads by Russian operatives: Report
- Las Vegas lights dim in tribute to shooting victims, heroes
- Las Vegas massacre Archive (Old Dominion)

Enter Web Address <http://>

3.1.2 Disadvantages

- There were various instances when a very important part of a page was not shown. In one of the seeds that I selected, my interest was focused on preserving the faces of the people who died in the massacre. As is shown on the capture below, the archived page failed to save the picture of one of the victims of the massacre.

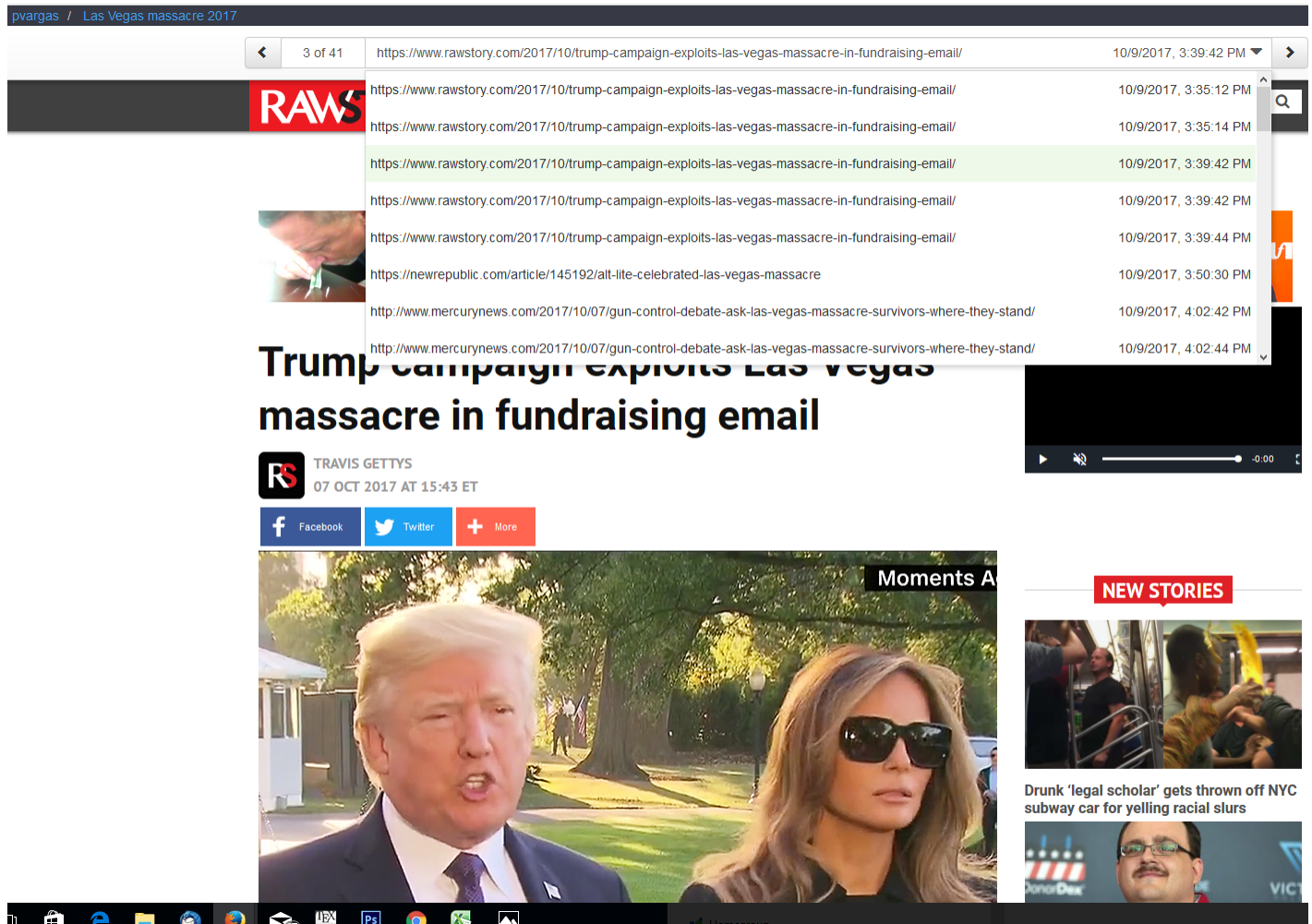
Quinton Robbins



3.2 Webrecorder.io

3.2.1 Advantages

- If the user is not interested in capturing the entire page, but just a particular portion of it, then this is the best tool for that type of requirement.
- A feature not available in **Archive-it.org**, but available in **Webrecorder** is the ability to add WARC files to the collection.
- The user also has the option of downloading the entire collection to the local computer using WARC formatted files.
- When viewing a collection, it can be organized in such a form that it may be replayed in the form of a story. See snapshot below.



- **Webrecorder.io** has a replay option which helps to patch a file that doesn't render properly. This is an option that **Archive-it.org** doesn't have and it proves to be very useful. Using my collection as an example, the archived file containing the pictures of mortally wounded individuals in the massacre is missing the photo of one individual. The ability to patch and fix a very important portion of a file helps the user to preserve the intended content in any particular collection.

3.2.2 Disadvantages

- **Webrecorder.io**'s recording interaction feature with a user may also become a major disadvantage, specially in the case when the user is required to download links of any particular web page. Every time there is user interaction, there is a possibility for error. Therefore, as the number of links in a page increases, so does the possibility that the user may not click on a very important link, and not including it into the collection.
- The interface bar on **Webrecorder.io** keeps flashing recording, but there there is not a way to know if all the resources representation have been loaded. The user has to go and replay the saved resource and verify that all tags render properly.
- Archived seeds get

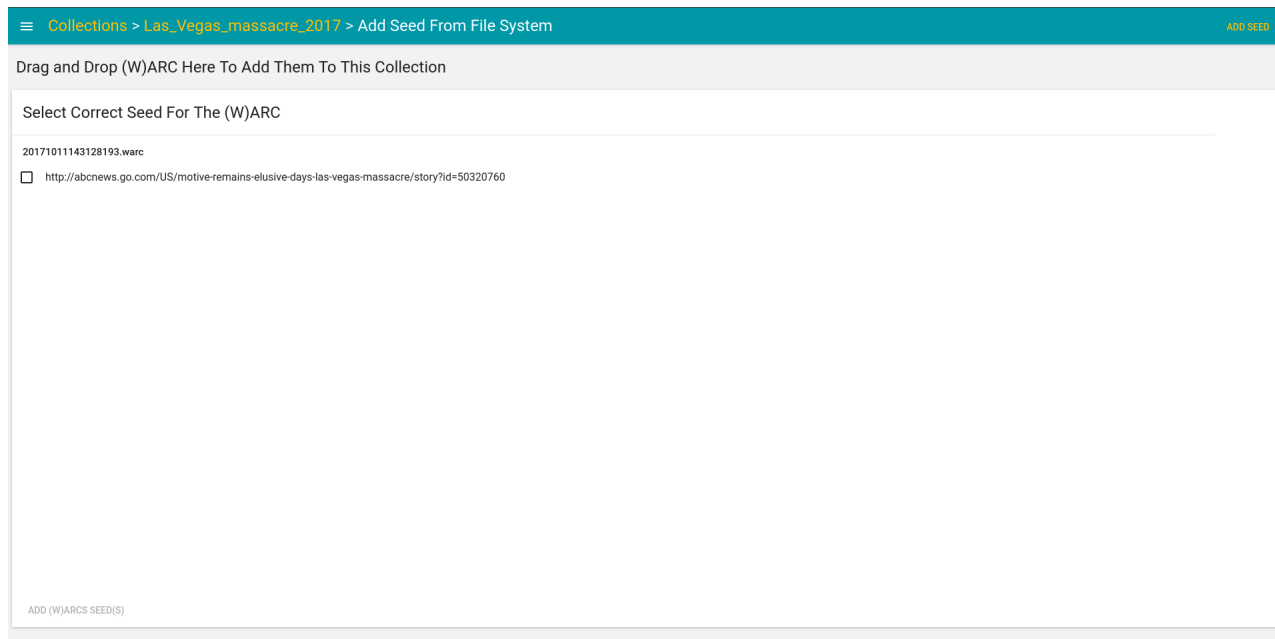
3.3 WAIL

3.3.1 Advantages

- **WARC** files can be dropped into the web collection interface
- The user has the option of archiving an URL directly from the live web.

3.3.2 Disadvantages

- When dragging WARC files into the WAIL interface the user only can drag and drop one file at a time. If additional WARC files were to be added the previous file waiting in the queue to be archived will disappear. A better option will be drag all the desired files into the interface and click on the available select button to add all the WARC seeds files.



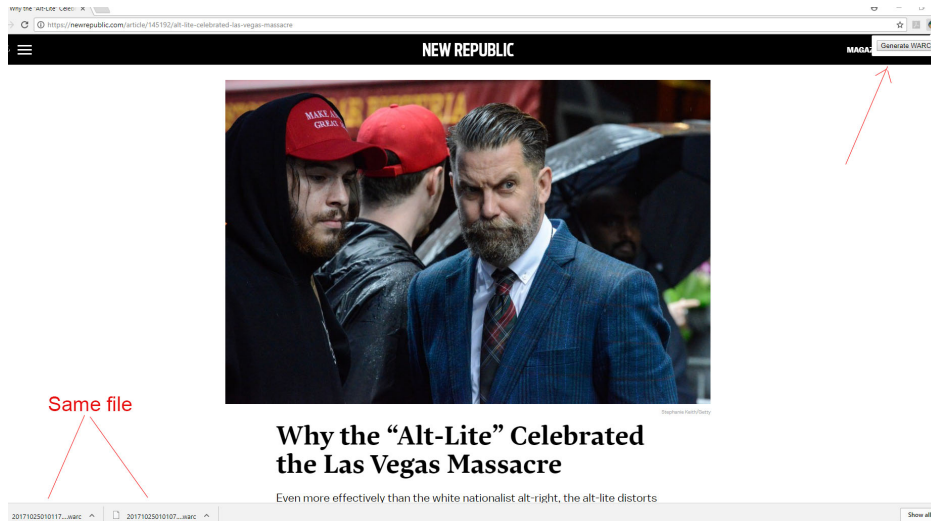
3.4 WARCreate

3.4.1 Advantages

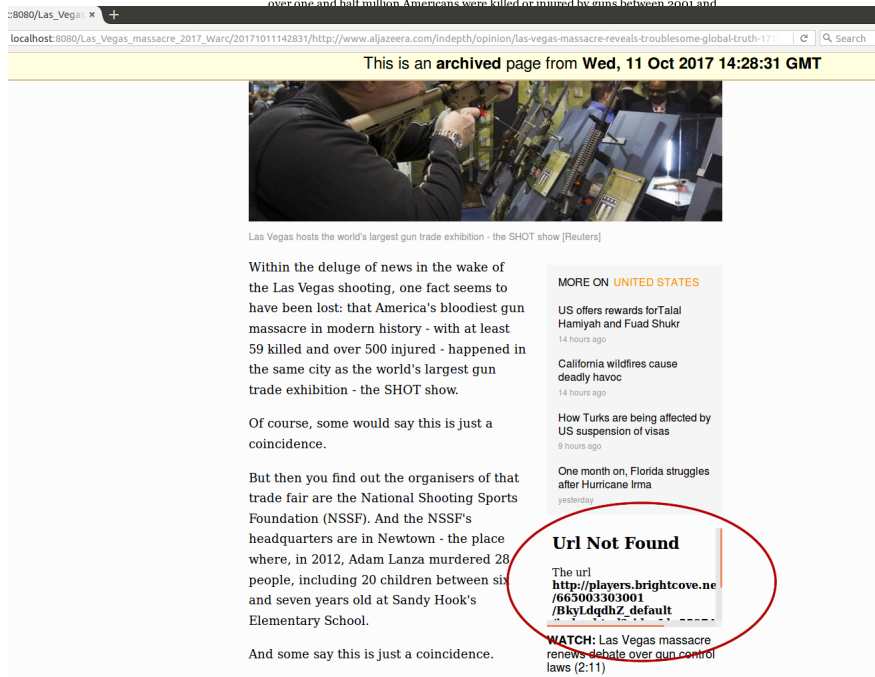
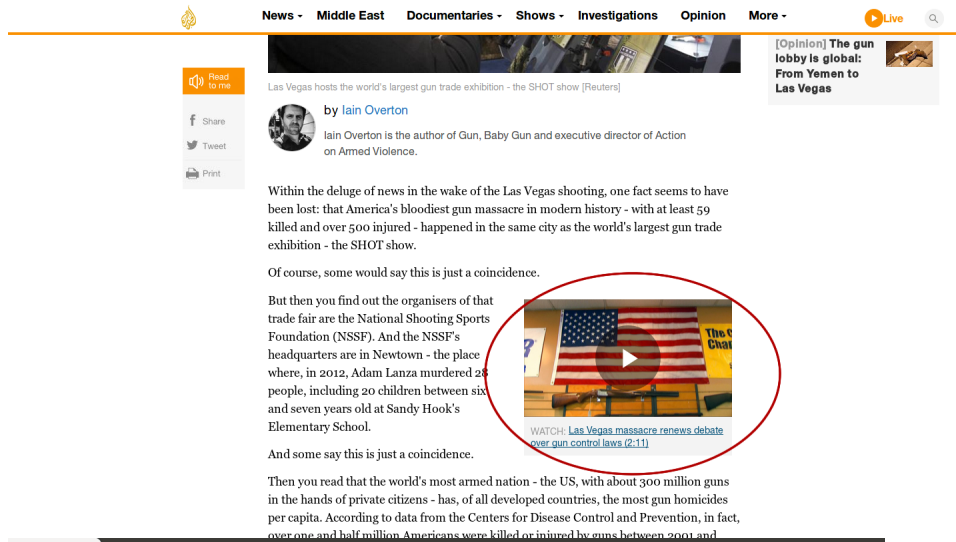
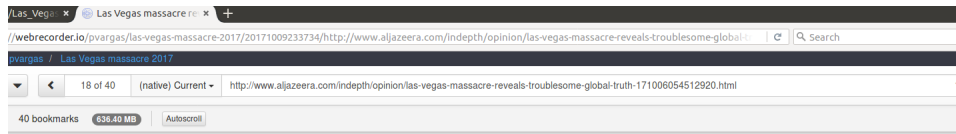
- Representation of a web page resources can be saved into a compressed format. The user does not have to worry about the resource size. If required, **WARCreate** will break the file into various chunks to accommodate size.
- All resources representation are retrieved automatically without any user intervention.
- The steps to create a WARC file is very simplify to the user. The user opens Chrome web browser, types the URL and clicks on WARCreate add-on

3.4.2 Disadvantages

- It is only available in Chrome
- A major disadvantage of **WARCreate** is the lack of interaction with the user during the generation of the WARC file. If the WARC file is not generated quickly when the user presses the “Generate WARC” button, then the user may continue to press the same button several times, leading to the duplication of the same WARC file. This was a common problem while working with **WARCreate**.



- There is no way to patch a particular resource representation into the WARC file. The two screenshots below are taken from **Webrecorder.io** and **WAIL** respectively. WAIL archived a WARC file generated by **WARCcreate**. It can be appreciated that the **Webrecorder.io** representation includes a video related to the massacre while the second representation does not.



4 CONCLUSIONS

I believe that 15 seeds were enough to preserve what transpired in the Las Vegas massacre. Heroism came to light on videos showing first responders providing medical assistance to injured victims. A web page was added to the collection showing the faces of those who were mortally wounded. The conversation about where people stand on gun control was preserved. The healing process was captured when the Vice-President and the President paid visit to the victims of this horrific act. Fake news angle also was preserved on how the Islamic State militant group (ISIS) published information detailing how they were responsible for this attack.

A few weeks after the Las Vegas massacre, the archived page compared with the live page shows some difference in content. The screen-shot below is from the original archived page <http://www.mercurynews.com/2017/10/07/gun-control-debate-ask-las-vegas-massacre-survivors-where-they-stand/>. The section labeled "Related Articles" marks some of the differences between the two represented resources. The

The screenshot shows a web page from The Mercury News. The main headline is "Where do Las Vegas massacre survivors stand on...". Below the headline is a sub-headline: "Like characters in a video game, Valerie Fowler and Kaitlyn Roll were among the targets of the deadliest mass shooting in modern U.S. history, running for their lives from a sniper whose powerful arsenal is redefining America's gun debate." The main body of the article begins with "Many survivors with ties to the Bay Area are only beginning to comprehend what happened when a 64-year-old white man employed his legally modified semi-automatic rifles to rain rapid-fire terror from the 32nd floor of a Las Vegas high rise. But almost a week after the shooter killed 58 of their fellow concert-goers and wounded nearly 500 more, some like Fowler are wishing they, too, had something during the madness Sunday night: a gun of their own." A quote from Valerie Fowler follows: "If I had a gun on me, I would have felt more safe," Fowler remembers telling Roll, one of her old high school friends from Santa Rosa, as they sat on the couch of their Las Vegas rental the next morning, watching the news replaying the traumatic footage.

A red circle highlights the "Related Articles" section on the left side of the page. The articles listed are:

- 'American Horror Story': Vegas tragedy forces edits to scene
- Many in country music mum over gun issues after Vegas deaths
- Where did Las Vegas shooter Stephen Paddock buy his guns?
- At Reno gun show, firearm fans denounce gun control but support bump stock ban
- A 'liberal teacher' became a conservative enemy for a viral Vegas tweet. But does she actually exist?

Below the "Related Articles" section, the article continues with "Roll, who had reached out to help a wounded woman and came home with blood on her hand, quickly disagreed: 'If we were all carrying guns in the crowd, it wouldn't have helped.'" and "At first glance, their disagreement illustrates how America's debate over gun control continues to be one of the country's most divisive and dig-in-your-heels issues. But could we be inching toward a middle ground after more than 20,000 people personally experienced the murder of 58 fellow country music fans — carried out by one man with 12 legally obtained weapons."

On the right side of the page, there is a "Sign up for our newsletters!" section with a "SUBSCRIBE" button, a "Follow Us" section with social media icons for Facebook, Twitter, Instagram, and RSS, and a "MOST POPULAR" section with a list of popular articles:

- 1 Fire maps: Expanded Santa Rosa evacuation area
- 2 Update: Wine Country fires claim 1500 structures — one death reported
- 3 Unhealthy smoke from wine

archived page shown in this section the following articles: American Horror Story - Vegas tragedy forces edits to scene; Many in country music mum over gun issues after Vegas deaths; Where did Las Vegas shooter Stephen Paddock buy his gun?; At Reno gun show, firearms fan denounce gun control but support bump stock ban; A liberal teacher became a conservative enemy for viral Vegas tweet. But does she actually exist?

On the other hand, the live web still shows stories related to the Las Vegas massacre, but they are different: California cops injured in Las Vegas mass shooting heroism denied workers' comp due to state law; father of California woman killed in Las Vegas shooting files wrongful-death lawsuit; Stanford University to study body of Vegas shooter; Las Vegas officials again adjust shooting timeline, GOP punts bump-stock decision to ATF.

www.mercurynews.com/2017/10/07/gun-control-debate-ask-las-vegas-massacre-survivors-where-they

The Mercury News **Where do Las Vegas massacre survivors stand on...**

Like characters in a video game, Valerie Fowler and Kaitlyn Roll were among the targets of the deadliest mass shooting in modern U.S. history, running for their lives from a sniper whose powerful arsenal is redefining America's gun debate.

Related Articles

- California cops injured in Las Vegas mass shooting heroism denied workers' comp due to state law
- Father of California woman killed in Las Vegas shooting files wrongful-death lawsuit
- Stanford University to study body of Vegas shooter
- Las Vegas officials again adjust shooting timeline
- GOP punts 'bump-stock' decision to ATF

Many survivors with ties to the Bay Area are only beginning to comprehend what happened when a 64-year-old white man employed his legally modified semi-automatic rifles to rain rapid-fire terror from the 32nd floor of a Las Vegas high rise. But almost a week after the shooter killed 58 of their fellow concert-goers and wounded nearly 500 more, some like Fowler are wishing they, too, had something during the madness Sunday night: a gun of their own.

"If I had a gun on me, I would have felt more safe," Fowler remembers telling Roll, one of her old high school friends from Santa Rosa, as they sat on the couch of their Las Vegas rental the next morning, watching the news replaying the traumatic footage.

Roll, who had reached out to help a wounded woman and came home with blood on her hand, quickly disagreed: "If we were all carrying guns in the crowd, it wouldn't have helped."

At first glance, their disagreement illustrates how America's debate over gun control continues to be one of the country's most divisive and dig-in-your-heels issues. But could we be inching toward a middle ground after more than 20,000 people personally experienced

MOST POPULAR >

- 1 Pending home sales plunge across Bay Area and state
- 2 This is when Google's village will start to appear in downtown San Jose
- 3 Audio: SFO tower tries to get Air Canada pilot's attention
- 4 \$70 Yosemite entrance fee: Big increases proposed for national parks
- 5 Los Gatos woman killed in crash was DUI suspect's mother
- 6 Again? FAA investigates another Air Canada landing mix-up at SFO
- 7 Jordan Bell's late-game dunk sparks varying reactions among Warriors, Mavericks
- 8 Bay Area hammered by loss of 4,700 jobs

How to build an edible Halloween skeleton platter
Bay Area News Group