Reactive and Proactive Archiving of Crisis

Sawood Alam and Mark Graham
Wayback Machine, Internet Archive

Quinn Dombrowski
Stanford University

December 13, 2022, CNI Fall 2022 Membership Meeting
Saving Ukrainian Cultural Heritage Online (SUCHO)

& ~1,300+ friends and strangers new friends from mostly North America and Western Europe
It all started on Twitter...

Anna Kijas @anna_kijas · Feb 26
Join me for a virtual data rescue session focused on music collections at cultural heritage institutions in Ukraine which may be at risk during the attack by Russia. For more info & to participate go to bit.ly/datarescueforUA (@achdotorg @aseeestudies) Share widely!

docs.google.com
Data Rescue for Ukraine
[Updated on March 1, 2022] We are now combining our efforts in order to work more nimbly and quickly...
It all started on Twitter…

Anna Kijas @anna_kijas · Feb 26
Join me for a virtual data rescue session focused on music collections at cultural heritage institutions in Ukraine which may be at risk during the attack by Russia.

bit.ly/datarescue

@aseeestudies

Sebastian Majstorovic 🇺🇦 @storytracer · Feb 27
Replying to @anna_kijas @achdotorg and @aseeestudies
Maybe the great folks at @webrecorder_io could get involved! Their browsertrix web archiving crawler is amazing and they are testing a cloud version atm: browsertrix.cloud. I’m ready to volunteer anyway I can and will see if we can contribute somehow at the @ACDH_OeAW.
It all started on Twitter…

Anna Kijas @anna_kijas · Feb 26
Join me for a virtual data rescue session focused on music collections at cultural heritage institutions in Ukraine which may be at risk during the attacks by Russian forces. Let's save the past together.

Quinn Dombrowski @quinnanya · Mar 1
Let's hear it for @storytracer who's been keeping extreme hours the last two days keeping his web scrapers 🔥🔥 for archiving Ukrainian cultural heritage! 🙏🏼❤️💙 Find out how you can help at sucho.org

Saving Ukrainian Cultural Heritage Online
Saving Ukrainian Cultural Heritage Online
It all started on Twitter…

Anna Kijas
Join me for cultural heritage!

Quinn Dombrowski @quinnanya · Mar 2
Floored at the response to our effort to archive Ukrainian cultural heritage websites: at the end of day 2 we have 419 people on our Slack, we've sent tens of thousands of first-captures & hundreds of files to the Internet Archives & created many web archives. The work continues.

Quinn Dombrowski @quinnanya · Mar 2
I'm keeping notes to write up later, to help the next people who take on this kind of project, but I don't want to forget the hope and the moments of joy in a dark, dark time. That's what's kept me going, through the all-consuming bubble of trying to co-organize this thing.

Quinn Dombrowski @quinnanya · Mar 2
If you’d like to join us, there’s still plenty to do. Check out the brand new orientation guide for how things are coming together:
Web archiving & collaboration toolkit

**Webrecorder**
*Web archiving for all!*

**Browsertrix Crawler**

**ArchiveWeb.page**

**Browsertrix Cloud**
*Web Archive Collection Zipped (WACZ)*

**INTERNET ARCHIVE**

**WayBack Machine**

**slack**

**Google Sheets**
<table>
<thead>
<tr>
<th>Status</th>
<th>Collection URL</th>
<th>Collection Name</th>
<th>Description</th>
<th>Claimed by</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Done</td>
<td><a href="https://www.km-dp.ua/">https://www.km-dp.ua/</a></td>
<td>College of Radio Electronics</td>
<td>hosted in Kyiv, critical priority</td>
<td>Lynda Howell</td>
<td>domain-level crawl of <a href="http://www.km.dp.ua">www.km.dp.ua</a></td>
</tr>
<tr>
<td></td>
<td><a href="http://rtp-archives.kr.ua/">http://rtp-archives.kr.ua/</a></td>
<td>Irpin Library Archive</td>
<td>hosted in Kyiv, high priority</td>
<td>Andrew B</td>
<td>small site with pages, problem with news page in HTML</td>
</tr>
<tr>
<td>Done</td>
<td><a href="http://www.imek-guselbkin.lviv.ua/">http://www.imek-guselbkin.lviv.ua/</a></td>
<td>Ethnic Museum of Cossack Folk Culture</td>
<td>hosted in Kyiv</td>
<td>Anna</td>
<td>small site with 18 pages, cloud crawl</td>
</tr>
<tr>
<td>Done</td>
<td><a href="http://vinn.ru.ua/">http://vinn.ru.ua/</a></td>
<td>Vinnitsa Historic and Cultural Reserve</td>
<td>hosted in Kyiv</td>
<td>Stefan M</td>
<td>high security risk, outdated, cloud crawl</td>
</tr>
<tr>
<td>In progress</td>
<td><a href="http://den.gov.ua/">http://den.gov.ua/</a></td>
<td>European Union Delegation to Ukraine</td>
<td>hosted in Kyiv</td>
<td>Vitali</td>
<td>cloud crawl, incomplete, trying manual crawl</td>
</tr>
<tr>
<td>Done</td>
<td><a href="https://kpdp.gov.ua/">https://kpdp.gov.ua/</a></td>
<td>Kyiv History Museum</td>
<td>hosted in Kyiv</td>
<td>Lynda Howell</td>
<td>high security risk, outdated, cloud crawl</td>
</tr>
<tr>
<td>Done</td>
<td><a href="http://www.palace.dp.ua/">http://www.palace.dp.ua/</a></td>
<td>Dnipropetrovsk State Art Museum</td>
<td>hosted in Kyiv, security certificate expired earlier this month</td>
<td>Anna</td>
<td>high risk, no malware, cloud crawl</td>
</tr>
<tr>
<td>In progress</td>
<td><a href="http://www.dnpu.dp.ua/">http://www.dnpu.dp.ua/</a></td>
<td>Dnipropetrovsk State Pedagogical University</td>
<td>hosted in Kyiv</td>
<td>Katerina</td>
<td>cloud crawl, incomplete, trying manual crawl</td>
</tr>
<tr>
<td>Done</td>
<td><a href="http://www.vice.dp.ua/">http://www.vice.dp.ua/</a></td>
<td>Dnipropetrovsk State University of Economics and Business</td>
<td>hosted in Kyiv</td>
<td>Dmytro</td>
<td>domain-level</td>
</tr>
<tr>
<td>In progress</td>
<td><a href="http://www.politeknik.dp.ua/">http://www.politeknik.dp.ua/</a></td>
<td>Dnipropetrovsk Politeknik</td>
<td>hosted in Kyiv</td>
<td>Oleksandr</td>
<td>domain-level</td>
</tr>
<tr>
<td>Done</td>
<td><a href="http://www.dnu.dp.ua/">http://www.dnu.dp.ua/</a></td>
<td>Dnipropetrovsk National University of Food Technologies</td>
<td>hosted in Kyiv</td>
<td>Dmytro</td>
<td>domain-level</td>
</tr>
<tr>
<td>Site down</td>
<td><a href="http://www.ukrinfo.com.ua/">http://www.ukrinfo.com.ua/</a></td>
<td>Kyiv National University of Architecture and Civil Engineering</td>
<td>hosted in Kyiv, site down</td>
<td>Dmytro</td>
<td>domain-level</td>
</tr>
<tr>
<td>In progress</td>
<td><a href="http://www.ukrinfo.com.ua/">http://www.ukrinfo.com.ua/</a></td>
<td>Kyiv National University of Architecture and Civil Engineering</td>
<td>hosted in Kyiv</td>
<td>Dmytro</td>
<td>domain-level</td>
</tr>
</tbody>
</table>
● **Discovery**: user submissions, WikiData, domain intelligence, Maps discovery

● **Prioritisation**: important collection? from at-risk area?

● **Triage**: how easy to crawl? needs scraping?

● **Crawling**: automatic, semi-automatic, manual

● **Quality control**: content check by multilingual volunteers
“Digital Dunkirk”
Web archiving is for everyone, from kids to retirees.
Key GLAM infrastructure is poorly adapted for web archiving

web crawler trap 😳
SUCHO gallery & equipment fund

SUCHO Equipment Fund

a network storage with discs to work to preserve the book heritage of Cherkasy and Ukraine.

children of our defenders, children from among internally displaced persons figuratively shared their experiences of the ongoing war on paper.
SUUCHO meme wall
National Digital Library of Ukraine
The situation is worse than ever for Ukrainian internet

Benjamin Schmidt, https://observablehq.com/@bmschmidt/sucho-status
Saving Ukrainian Cultural Heritage Online

SUCHO is an initiative of over 1,500 international volunteers who are collaborating online to digitize and preserve Ukrainian cultural heritage. Since the start of the invasion, SUCHO has web-archived more than 5,000 websites and 50TB of data of Ukrainian cultural institutions, to prevent these websites from going offline. The websites range from national archives to local museums, from 3D tours of churches to children’s art centers.

SUCHO is now moving into a new phase of the project, defined by three goals: Curate, Donate, Educate. In Phase 2, SUCHO will coordinate aid shipments of digitization hardware, exhibit Ukrainian culture online and organize training for Ukrainian cultural workers in digitization methods.

Where can you see the SUCHO data?
We are currently in the process of curating the full 50TB of web archives we have collected. The raw files are publicly accessible from the AWS Open Data registry. Once they have been minimally curated, we will make them viewable in a user-friendly way on our website.
Truth is the first casualty of war
The Internet Archive as a Library...

- Collect
- Preserve
- Lend
- Collaborate
- Access
- Discovery
- Make Useful
SUUCHO Collection on Archive.org
SUCHARO WACZ files ingestion into Petabox

Amazon S3 → 50 TB → Internet Archive
If You See Something...
"An An-26 plane has just been shot down in the Torez area, lying somewhere behind the Progress mine. They warned not to fly in "our sky". And here is a video confirmation of the next "birdfall". The bird fell behind the slagheap, did not catch the residential sector. Peaceful people were not hurt.

And there is also information about the second downed plane, apparently Su."
Ukrainian News (316) and Russian News (947)
Archiving the Ukrainian and Russian Web

Ukrainian: 3 million+
Russian: 24 million+
How 1 seed URL can result in 31,772 archived URLs

1 seed page

338 embeds

174 outlinks

31,259 outlink embeds

Results from cnn.com on June 17, 2021
Archiving Telegram

Telegram

Telegram is a mostly open-source chat system (web and app-based) which acts as a social network due to features like group chats and channels, originally intended as a drop-in replacement for WhatsApp.

Like Facebook pages/groups and Twitter, Telegram chats can contain a significant amount of history (text, audio/video and images) of various online communities, as well as personal digital history.

Contents

1. Structure
   2. Export methods
      2.1 Web data only
      2.1.1 Deprecated
      2.2 All data

3. Notable channels
   3.1 Telegram official
   3.2 Hong Kong pro-democracy movement / 2019–2020 Hong Kong protests
   3.3 QAnon
   3.4 2022 Russian invasion of Ukraine

4. See also

Structure

Channels on Telegram are essentially public streams of posts, similar to Twitter or RSS.

Since 2019-06-01 (update blog post), it is possible to "preview" channels on the web (e.g. @Telegram at [1] or [2]).

Export methods

Web data only
Telegram - Sourcing Channels

**Russia**
- Channels: 426,000
- Groups: 39,600
- Total audience: 2,003,300,000

**Ukraine**
- Channels: 46,000
- Groups: 6,100
- Total audience: 212,450,000

**Belarus**
- Channels: 7,000
- Groups: 1,800
- Total audience: 21,020,000

**Uzbekistan**
- Channels: 106,000
- Groups: 13,300
- Total audience: 575,630,000

**Kazakhstan**
- Channels: 8,000
- Groups: 1,100
- Total audience: 41,680,000

**Kyrgyzstan**
- Channels: 2,000
- Groups: 400
- Total audience: 14,080,000
### Telegram - Progress to Date

**22 Billion Telegram URLs Archived**

<table>
<thead>
<tr>
<th>date</th>
<th>items</th>
<th>count</th>
<th>size</th>
<th>B/uri</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022-11-10</td>
<td>416</td>
<td>103,252,794</td>
<td>4%</td>
<td>4,502,971,462,671</td>
</tr>
<tr>
<td>2022-11-09</td>
<td>463</td>
<td>105,084,491</td>
<td>18%</td>
<td>5,099,244,428,927</td>
</tr>
<tr>
<td>2022-11-08</td>
<td>519</td>
<td>116,564,970</td>
<td>18%</td>
<td>5,616,353,387,213</td>
</tr>
<tr>
<td>2022-11-07</td>
<td>412</td>
<td>102,248,500</td>
<td>18%</td>
<td>4,462,872,974,603</td>
</tr>
<tr>
<td>2022-11-06</td>
<td>335</td>
<td>79,252,715</td>
<td>15%</td>
<td>3,624,432,426,846</td>
</tr>
<tr>
<td>2022-11-05</td>
<td>485</td>
<td>115,839,500</td>
<td>19%</td>
<td>5,240,157,463,761</td>
</tr>
<tr>
<td>2022-11-04</td>
<td>556</td>
<td>147,536,292</td>
<td>12%</td>
<td>6,017,683,705,738</td>
</tr>
<tr>
<td>2022-11-03</td>
<td>380</td>
<td>108,390,626</td>
<td>6%</td>
<td>4,184,759,638,420</td>
</tr>
<tr>
<td>2022-11-02</td>
<td>360</td>
<td>102,359,778</td>
<td>6%</td>
<td>3,887,725,835,588</td>
</tr>
<tr>
<td>2022-11-01</td>
<td>482</td>
<td>117,259,979</td>
<td>7%</td>
<td>5,289,078,689,361</td>
</tr>
</tbody>
</table>

**ALL collections**

<table>
<thead>
<tr>
<th>date</th>
<th>items</th>
<th>count</th>
<th>size</th>
<th>B/uri</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022-11-10</td>
<td>416</td>
<td>103,252,794</td>
<td>4%</td>
<td>4,502,971,462,671</td>
</tr>
<tr>
<td>2022-11-09</td>
<td>463</td>
<td>105,084,491</td>
<td>18%</td>
<td>5,099,244,428,927</td>
</tr>
<tr>
<td>2022-11-08</td>
<td>519</td>
<td>116,564,970</td>
<td>18%</td>
<td>5,616,353,387,213</td>
</tr>
<tr>
<td>2022-11-07</td>
<td>412</td>
<td>102,248,500</td>
<td>18%</td>
<td>4,462,872,974,603</td>
</tr>
<tr>
<td>2022-11-06</td>
<td>335</td>
<td>79,252,715</td>
<td>15%</td>
<td>3,624,432,426,846</td>
</tr>
<tr>
<td>2022-11-05</td>
<td>485</td>
<td>115,839,500</td>
<td>19%</td>
<td>5,240,157,463,761</td>
</tr>
<tr>
<td>2022-11-04</td>
<td>556</td>
<td>147,536,292</td>
<td>12%</td>
<td>6,017,683,705,738</td>
</tr>
<tr>
<td>2022-11-03</td>
<td>380</td>
<td>108,390,626</td>
<td>6%</td>
<td>4,184,759,638,420</td>
</tr>
<tr>
<td>2022-11-02</td>
<td>360</td>
<td>102,359,778</td>
<td>6%</td>
<td>3,887,725,835,588</td>
</tr>
<tr>
<td>2022-11-01</td>
<td>482</td>
<td>117,259,979</td>
<td>7%</td>
<td>5,289,078,689,361</td>
</tr>
</tbody>
</table>
VK.com - Results

- 3K Regional Groups with daily snapshots since May, 2022
- About 5 Million posts
- About 200 Pro-war groups with daily snapshots
Future justice: We set up an online archive to preserve the most important Russian independent media now threatened by dictatorship.

Understanding and awareness: We create a platform for discussion of recent Russian history to better understand the dynamics of authoritarianism and origins of Russia’s political and social catastrophe.

Support of free media: We create and spread tools to help different societies facing attacks on the freedom of press to resist the pressure.
Saving Russian Independent Media (Web)

140 million URLs Archived
Web collections fulltext search index API

In partnership with MediaCloud
Archive-It is supporting multiple partners

---

**Ukraine Conflict**

**Collected by:** Internet Archive Global Events  
**Archived since:** Feb, 2014  
**Description:** This collection seeks to document conflict in Ukraine. Contributions to this collection were made by the Archive-It team and subject matter experts from Stanford University, Library of Congress, Global Investigative News Network, and Harvard Ukrainian Research Institute in the fields of investigative journalism, Russian, and Eurasian studies. Content includes news outlets, social media, blogs, and government websites. Sites are in English, Ukrainian, Russian, and other languages.  
**Subject:** Government, Spontaneous Events

---

**Narrow Your Results**

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Enter search terms here

---

**Group** | **Sort By** | **Count**  
--- | --- | ---  
1 (100) | |  
2 (243) | |  
3 (407) | |  
4 (332) | |  
5 (650) | |  

**Contributor** | **Sort By** | **Count**  
--- | --- | ---  
Library of Congress (979) | |  
Harvard Ukrainian Research Institutes (914) | |  

---

**Page 1 of 67 (6,616 Total Results)**
Adding Archive URLs to Wikipedia

4.8 million to ru.wikipedia.org
2.3 million to uk.wikipedia.org
Adding Archive URLs to Wikipedia
### Adding Archive URLs to Wikipedia

**TARB Insights**

Date Range:

2016/07/13 – 2022/11/11

<table>
<thead>
<tr>
<th>Wikis</th>
<th>Page Edits</th>
<th>Link Edits</th>
</tr>
</thead>
<tbody>
<tr>
<td>269</td>
<td>20,503,277</td>
<td>31,857,145</td>
</tr>
<tr>
<td>↑ 131</td>
<td>↑ 11,821</td>
<td>↑ 14,083</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tagged Links</th>
<th>Fixed Dead Links</th>
<th>Added Archive URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,011,776</td>
<td>15,525,237</td>
<td>28,798,620</td>
</tr>
<tr>
<td>↑ 1,226</td>
<td>↑ 12,128</td>
<td>↑ 12,843</td>
</tr>
</tbody>
</table>
Adding Archive URLs to Wikipedia

Monthly Links Edits on Selected Wikis

Select Wikis to compare:

ruwiki ✗ ukwiki ✗ enwiki ✗

Total Links | Total Edits | Dead Links | Dead Edits | Data

Month: Jul 2016 to Jul 2022
Russian, Belarusian, Ukrainian TV
Russian TV (No Captions)

Информационный канал
1TV November 11, 2022 10:45am-12:01pm MSK

В центре внимания - специальная военная операция, которую вооруженные силы России проводят на Украине. В прямом эфире - свидетельства очевидцев и последние официальные данные о происходящем на Украине

This broadcast did not contain any captioning
Русское ТВ

Транскрипция:

Ну если и о президенте страны даёт отмашку, то, конечно, такие подразделения создаются причин не не смотрели на здоровье на судимости брали грубовато, лишь бы человек был готов ехать на Донбасс и везать все параметры новой киевские власти торопились, юго-восток, страны, отказывался признать результаты переворота и был готов проголосовать на ...

Источник: Google Speech-to-Text API Automatic Transcription
A New Approach To Understanding War Through Television News: Introducing The TV News Visual Explorer & The Belarusian, Russian & Ukrainian TV News Archive

Posted on June 2, 2022 by Kaley Lestaro

For more than 20 years, the Internet Archive’s Television News Archive has monitored television news, preserving more than 9.5 million broadcasts totaling more than 6.6 million hours from across the world, with a continuous archive spanning the past decade. Today just a small sliver of that archive is accessible to journalists and scholars due to the inaccessibility of video at this scale: fast-forwarding through that much television news is simply beyond the ability of any human to make sense of. The small fraction of programs that contain closed captioning, speech recognition transcripts or OCR onscreen text can be keyword searched through the TV Explorer and TV AI Explorer, but for the majority of this global multi-decade archive, there has until now been no way for researchers to assess and understand the narratives of television news at scale, especially the visual landscape that distinguishes television from other forms of media and which is so central to understanding many of the world’s biggest stories from war to pandemics to the economy.

As the TV News Archive enters its third decade, it is increasingly exploring the ways in which it can preserve the domestic and international response to global events as it did with 9/11 two decades ago. As a first step towards this vision, over the last few months the Archive has preserved more than...
Multiple GDELT Project Experiments
TV Rain (Russian Independent TV)
TV Rain (Russian Independent TV)
TV Rain (Russian Independent TV)
TV Rain (Russian Independent TV)
TV Rain Newsroom (English)
Echo of Moscow (Russian Independent Radio)
Collecting Ukrainian Books

Ukrainian Book Drive: Please Contribute
Posted on March 29, 2022 by Breaster Kehoe
Better World Books

The Internet Archive creates wish lists of books based on Wikipedia citations
Adding links to Wikipedia Citations

Helping Ukrainian Scholars, One Book at a Time
Posted on May 2, 2022 by christieapple
**SUUCHO**

- 1,300 volunteers, ~5,000 websites, ~50 TB of web archives
- Distributed approach
- Subject-area focus & team including disciplinary expertise
- Challenges with web archiving sites built on widely-used cultural heritage platforms
- Very little cultural heritage is digitized; need to address that now & for the future
- Volunteering brings new ideas back to your organization

**Internet Archive**

- Supporting SUUCHO.org
- Russian and Ukrainian News Sites
- Russian and Ukrainian Websites
- Archiving Telegram
- Archiving VK.com
- Saving Russian Independent Media (PEN.org & Bard College)
- Many Archive-It projects
- Russian, Ukrainian and Belarusian TV (GDELT Project)
- Independent TV (TV Rain)
- Independent Radio (Echo of Moscow)
- Russian and Ukrainian Wikipedia sites

qad@stanford.edu ● sawood@archive.org ● mark@archive.org