CROATIAN WEB ARCHIVE (HAW) :
SPREADING BOUNDARIES

Karolina Holub & Inge Rudomino, National and University Library in Zagreb

BACKGROUND

❑ 1991 - 1st Internet connection in Croatia

❑ 1993 Croatia got ccTLD .hr

❑ 1997 the Law on Libraries (online publications as legal deposit)

❑ 1998 NSK started cataloguing online resources

❑ 2003 project Design of the System for Capturing and Archiving Legal Deposit of Croatian Web Publications

❑ September 2004 first archived web resource

❑ 2008 member of the International Internet Preservation Consortium (IIPC)

❑ 2019 – 2020 new regulations

1997. • Law on Libraries

2019. • Law on Libraries and Library Activities

2020. • Croatian Legal Deposit Regulations by the Ministry of Culture and Media
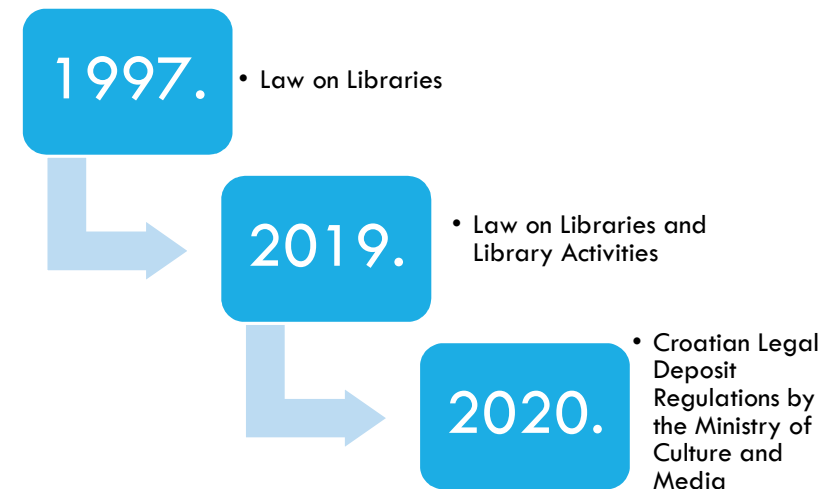
ABOUT HAW

☐ Collection of content archived from the web with the purpose of archiving and permanently storing online legal deposit copy

☐ Developed in 2004 in collaboration with the University Computing Centre University of Zagreb (Srce)

☐ Since 2011 conducting .hr crawls and thematic/event crawls

☐ 2 persons fulltime, 1 part time

2004-2011

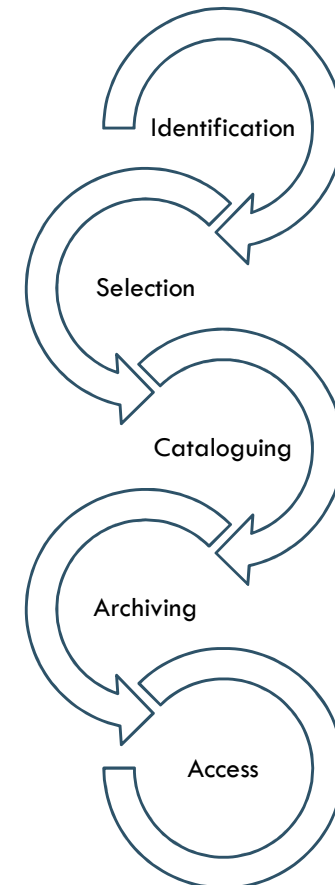2011-2019
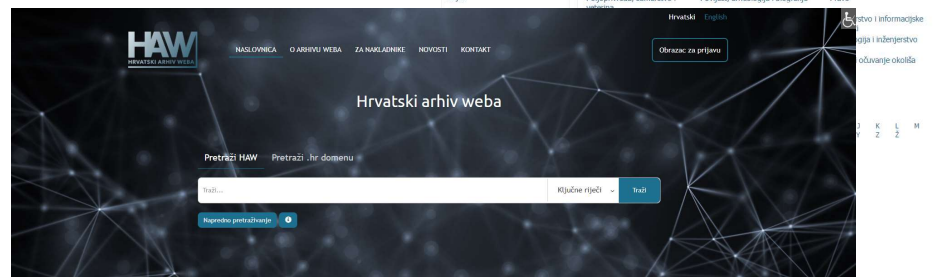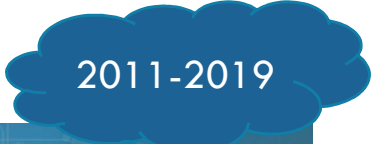
ARCHIVING APPROACHES

SELECTIVE

.HR CRAWL

THEMATIC/EVENT CRAWLS

SELECTIVE ARCHIVING WORKFLOW

❑Archiving system developed by Srce - University Computing Centre
University of Zagreb

❑Selection Criteria https://haw.nsk.hr/en/selection-criteria/

❑Identification (registration form) http://haw.nsk.hr/en/registration_form

❑Cataloguing every title

❑Daily data exchange between ILS and HAW archiving system

❑Access → public/within the NUL

❑Full-text search (basic & advanced), browse by subject

❑ URN-NBN for every title and archived copy

❑OAI-PMH interface for metadata harvesting

Identification

Selection

Cataloguing

Archiving

Access

HOW TO SEARCH/BROWSE SELECTIVE HAW?

Search HAW   Search .hr domain

Search...   Keywords ∨   Search

Title
URL
Keywords

Advanced search

Advanced search

Vjesnik.hr

URN:NBN: urn:nbn:hr:230:344071
ID: 252892
URL: http://www.vjesnik.hr
Other title: Vjesnik on-line, Vjesnik on-line :
Type: web site
Subject: Knowledge and culture, Politics, News portals
Rights: Free access. All rights reserved.
Note: URL not available.
Number of copies: 2386

		Size	Time of archiving	Web address	URN:NBN
1. Copy		23.77 MB	14.06.2012 17:13:14	http://www.vjesnik.hr	urn:nbn:hr:230:646893
2. Copy		23.95 MB	13.06.2012 17:10:33	http://www.vjesnik.hr	urn:nbn:hr:230:570136
3. Copy		24.06 MB	12.06.2012 17:07:35	http://www.vjesnik.hr	urn:nbn:hr:230:275411
4. Copy		24.25 MB	11.06.2012 17:07:29	http://www.vjesnik.hr	urn:nbn:hr:230:533840

Subject categories

Biology, botany and zoology
Number of subcategory (3)

Civilization and culture
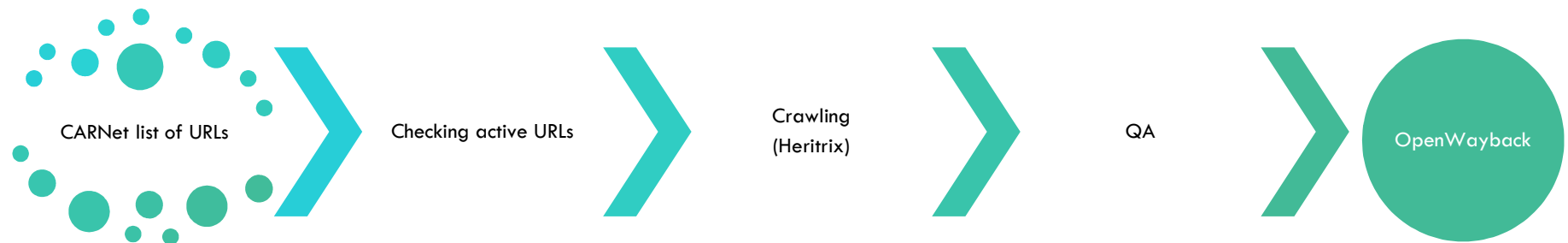Number of subcategory (5)

Society

Economics
Number of subcategory (4)

Ethnology and anthropology
Number of subcategory (2)

Philosophy and religion
Number of subcategory (2)

.HR CRAWLS

❑ Since 2011, annually

❑ 10 .hr crawls

❑ Tools: Heritrix, OpenWayback

CARNet list of URLs → Checking active URLs → Crawling (Heritrix) → QA → OpenWayback

HOW TO SEARCH .HR CRAWLS?

THEMATIC/EVENT CRAWLS

☐ Since 2011

☐ Periodical crawl of websites related to topics and events of national importance

☐ Tools: Web Curator Tool, Heritrix, OpenWayback

☐ Collection level cataloguing in ILS

COVID-19 (3)

The global pandemic of COVID-19 (coronavirus disease 2019), caused by the coronavirus SARS-CoV2 (severe acute respiratory syndrome coronavirus-2), started in December 2019, in Wuhan, the capital of China's Hubei province, wherefrom it rapidly spread to the rest of the world, causing a wide range of global socio-economic consequences.

The collection includes websites and news sites covering issues relating to the origin of the latest coronavirus, infection patterns, experiences of COVID-19 patients, official mortality rates and other COVID-19 cases-related statistics and recommendations from doctors, politicians, researchers and healthcare professionals. Also, the included resources cover the medical, scientific, social, economic and political aspects of the pandemic.

In order to make the collection as comprehensive as possible, the Croatian Web Archive issued a call for submissions of relevant COVID-19-online content. Since the pandemic is still present, online resources are still being collected and the collection will soon be updated.

2021 Local Elections

The elections in Croatia for the members of the representative bodies of the units of local and regional self-government, municipality heads, mayors and county prefects and their deputies are held every four years.

The collection includes all publicly available 2021 elections-related online content, as well as the websites of the participating political parties, webpages of the candidates and web portals covering the elections.

Resources in the collection were harvested as part of a crawl performed between 28 and 30 June 2021, resulting in 448 collected URLs, totalling 277 GB of data.

Earthquake in Sisak, Petrinja, Glina and the surroundings

On 29 December 2020 at 12.19 PM an earthquake measuring 6.2 on the Richter scale with the epicentre at 3 kilometres southwest of the city of Petrinja severely hit Croatia's Sisak-Moslavina County, resulting in human casualties and extensive material damage in Sisak, Petrinja, Glina and other cities, towns and villages in the county.

The earthquake was one of the two strongest recorded earthquakes in the Republic of Croatia since the beginning of the 20th century and apart from all across Croatia it was also felt in Austria, Bosnia and Herzegovina, the Czech Republic, Italy, Hungary, Germany, Montenegro, Romania, Slovakia, Slovenia and Serbia.

Resources in the collection were harvested as part of a crawl performed between 4 and 6 February 2021, resulting in 452 collected URLs, totalling 144 GB of data.
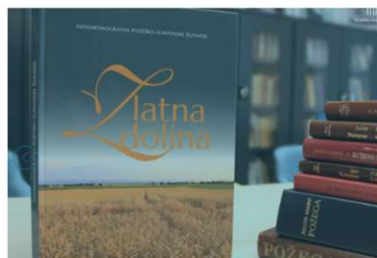
LOCAL HISTORY WEB COLLECTIONS

Spring 2021

Collaboration with public libraries: *training, selection* (URL nomination), *promotion, comprehensive archived web collection*

4 public libraries, 4 new collections

Plans
- Include more public libraries
- Create nomination tool/form
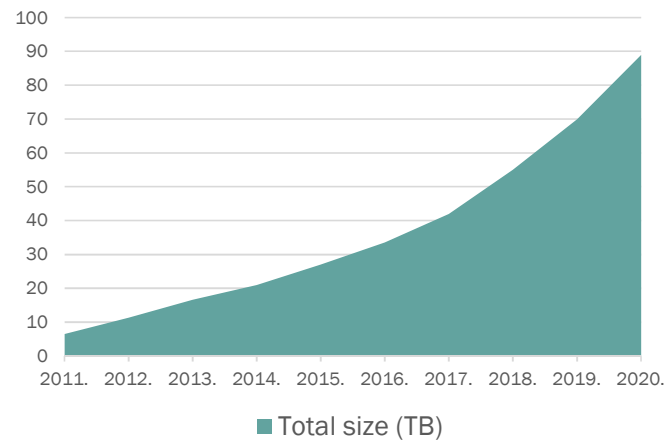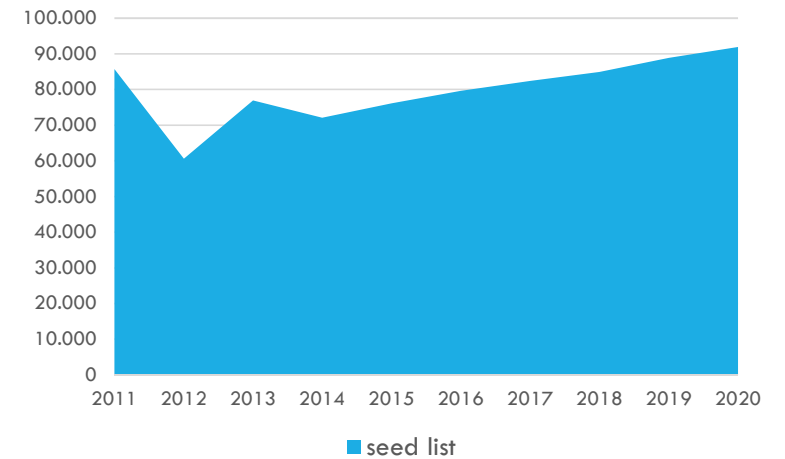- Annually

Local collections

Požega

Mali Lošinj

Cres

View all

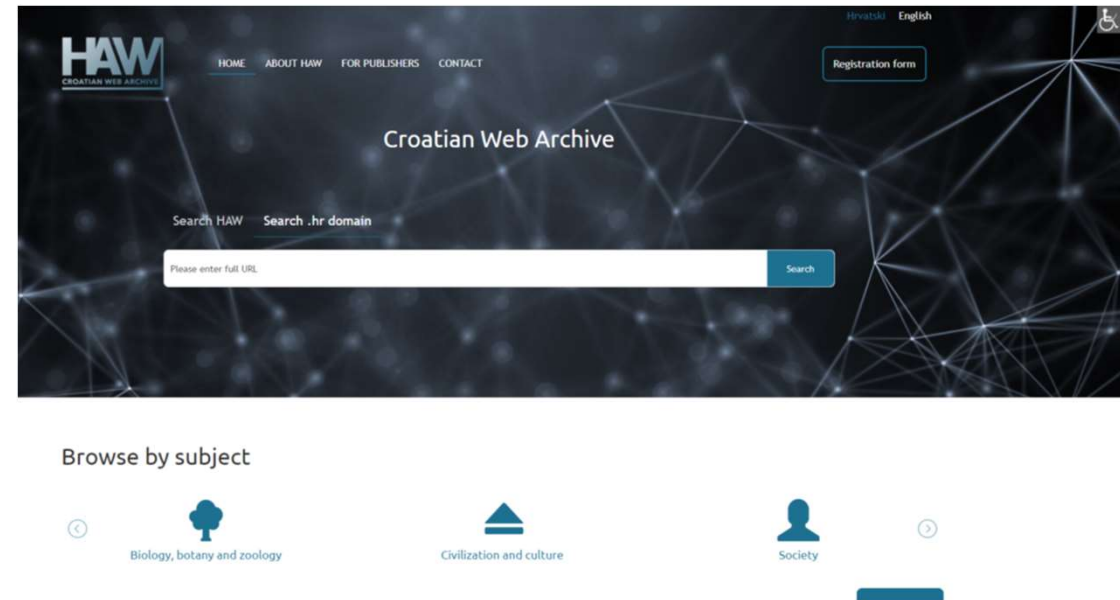# STATISTICS

Over 8200 Archived Titles
87 415 Archived instances

10 domain crawls; over 1 billion files

22 thematic collections;
4 local history collections

seed list

Total size (TB)

NEXT STEPS…

☐ New tool for selective archiving

☐ Archiving social media (starting with Twitter)

☐ Increase the number of .hr crawls

☐ Collaboration with researchers
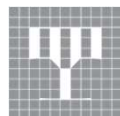
THANK YOU !

http://haw.nsk.hr/en
haw@nsk.hr
kholub@nsk.hr
irudomino@nsk.hr
#HAW_NSK

NATIONAL AND UNIVERSITY LIBRARY IN ZAGREB

HAW
CROATIAN WEB ARCHIVE