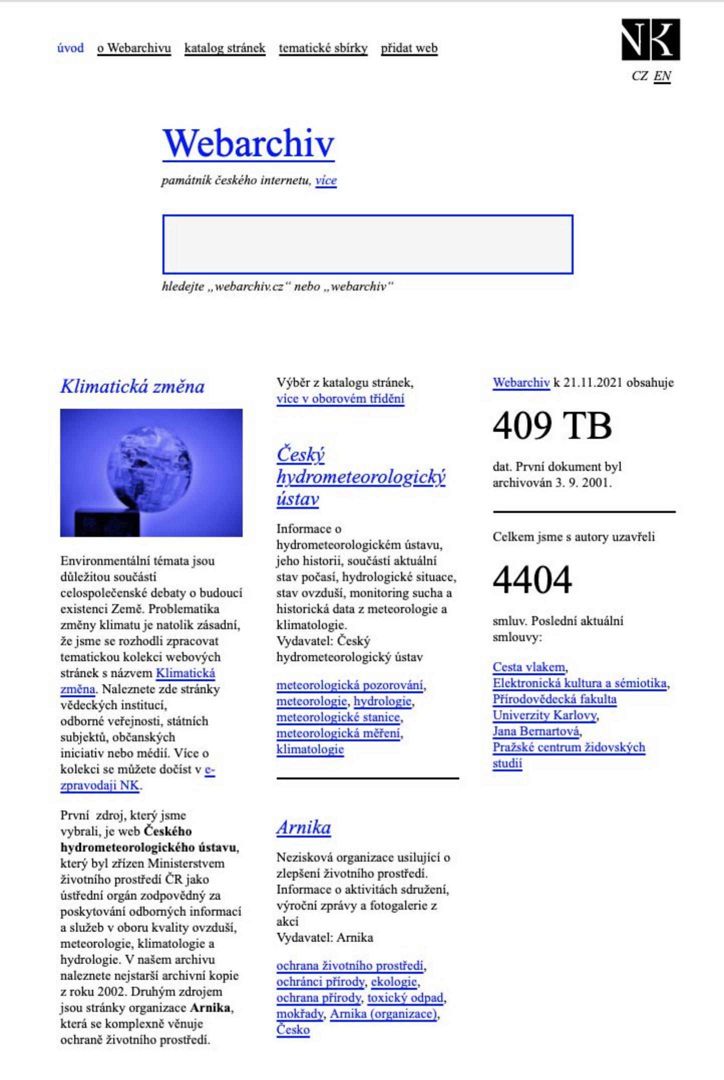
webarchiv.cz

Web archive of National Library of the Czech Republic

Marie Haškovcová, Zdenko Vozár

webarchiv.cz today

- digital library which preserves websites for future generations
- Czech web resources (territory, language, authorship or topic/content) not only within the Czech domain
- more than 400 TB of data
- 4 people + IT support



Webarchiv

památník českého internetu, vice

hledejte „webarchiv.cz" nebo „webarchiv"

Klimatická změna

Environmentální témata jsou důležitou součástí celospolečenské debaty o budoucí existenci Země. Problematika změny klimatu je natolik zásadní, že jsme se rozhodli zpracovat tematickou kolekci webových stránek s názvem Klimatická změna. Naleznete zde stránky vědeckých institucí, odborné veřejnosti, státních subjektů, občanských iniciativ nebo médií. Více o kolekci se můžete dočíst v e-zpravodaji NK.

První zdroj, který jsme vybrali, je web **Českého hydrometeorologického ústavu**, který byl zřízen Ministerstvem životního prostředí ČR jako ústřední orgán zodpovědný za poskytování odborných informací a služeb v oboru kvality ovzduší, meteorologie, klimatologie a hydrologie. V našem archivu naleznete nejstarší archivní kopie z roku 2002. Druhým zdrojem jsou stránky organizace **Arnika**, která se komplexně věnuje ochraně životního prostředí.

Výběr z katalogu stránek,
vice v oborovém třídění

Český hydrometeorologický ústav

Informace o hydrometeorologickém ústavu, jeho historii, součástí aktuální stav počasí, hydrologické situace, stav ovzduší, monitoring sucha a historická data z meteorologie a klimatologie.
Vydavatel: Český hydrometeorologický ústav

meteorologická pozorování, meteorologie, hydrologie, meteorologické stanice, meteorologická měření, klimatologie

Arnika

Nezisková organizace usilující o zlepšení životního prostředí. Informace o aktivitách sdružení, výroční zprávy a fotogalerie z akcí
Vydavatel: Arnika

ochrana životního prostředí, ochránci přírody, ekologie, ochrana přírody, toxický odpad, mokřady, Arnika (organizace), Česko

Webarchiv k 21.11.2021 obsahuje

409 TB

dat. První dokument byl archivován 3. 9. 2001.

Celkem jsme s autory uzavřeli

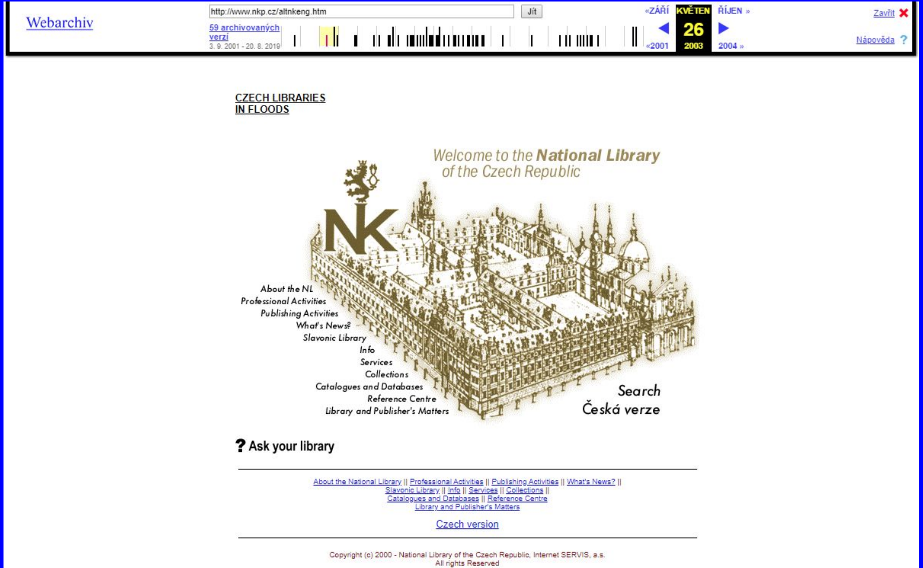
4404

smluv. Poslední aktuální smlouvy:

Cesta vlakem,
Elektronická kultura a sémiotika,
Přírodovědecká fakulta Univerzity Karlovy,
Jana Bernartová,
Pražské centrum židovských studií

webarchiv.cz
history

- 2000 – project of National Library of the CR, Moravian Library and Masaryk University
- 2001 – first archived website
- 2005 – regular harvesting of content
- 2007 – joining the IIPC – International Internet Preservation Consortium
- 2020 – 20 years anniversary



CZECH LIBRARIES
IN FLOODS

? Ask your library

About the National Library || Professional Activities || Publishing Activities || What's News? ||
Slavonic Library || Info || Services || Collections ||
Catalogues and Databases || Reference Centre
Library and Publisher's Matters

Czech version

Copyright (c) 2000 - National Library of the Czech Republic, Internet SERVIS, a.s.
All rights Reserved

webarchiv.cz

legal issues

- **Copyright act** – Library License allows the National Library of the CR to make a reproduction of a work for its own archiving and conservation purposes
- **Online access** – based on contract with publishers or on Creative Commons licence, the entire archive is available in the library building

less than **0,4 %** of the content is available outside the library building

- **Legal deposit act** – does not cover born digital documents
- **Directive of the European Parliament and of the Council on Copyright in the Digital Single Market** – has not yet been implemented in Czech legislation

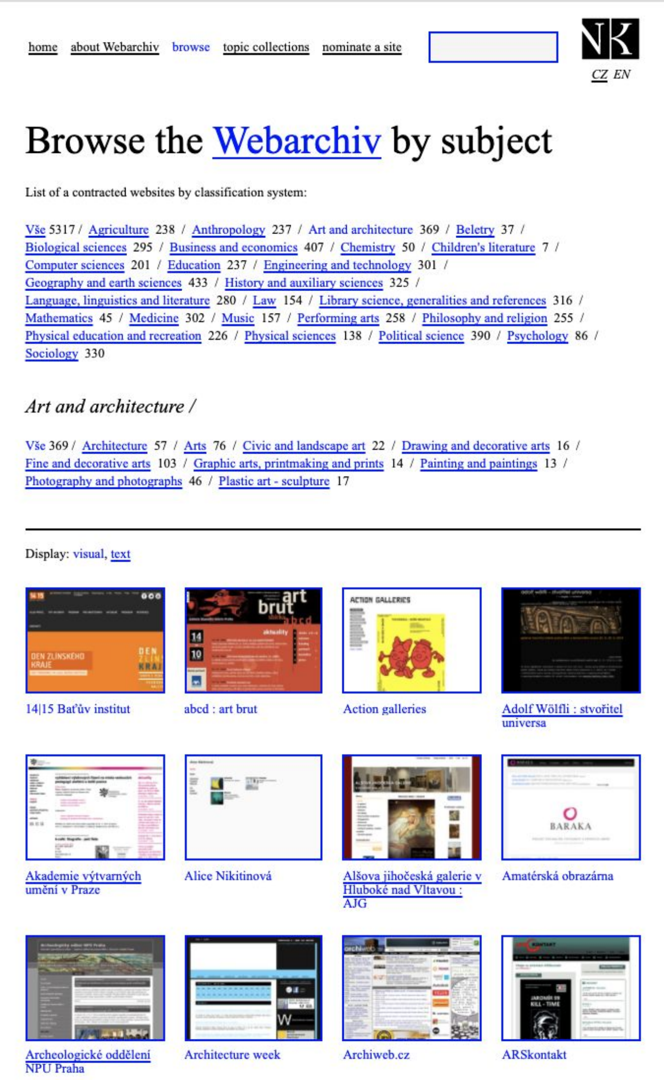
webarchiv.cz

collection policy

- **Comprehensive harvests**
  - contract with czech domain provider CZ.NIC
  - once or twice a year crawl of the whole .cz domain
  - 1,4 million of second order domains / domain.cz
- **Selective harvests**
  - selective approach, curated resources
- **Topic collections**
  - collections of resources related to certain event or topic

webarchiv.cz
selective harvest

- selective approach
- resources with historical, scientific or cultural value
- online access – contract or CC, more than 5000 archived websites with online access
- crawled periodically
- cataloging records in Czech national bibliography
- catalog - resources sorted by topics according to the conspectus method



Browse the Webarchiv by subject

List of a contracted websites by classification system:

Vše 5317 / Agriculture 238 / Anthropology 237 / Art and architecture 369 / Beletry 37 / Biological sciences 295 / Business and economics 407 / Chemistry 50 / Children's literature 7 / Computer sciences 201 / Education 237 / Engineering and technology 301 / Geography and earth sciences 433 / History and auxiliary sciences 325 / Language, linguistics and literature 280 / Law 154 / Library science, generalities and references 316 / Mathematics 45 / Medicine 302 / Music 157 / Performing arts 258 / Philosophy and religion 255 / Physical education and recreation 226 / Physical sciences 138 / Political science 390 / Psychology 86 / Sociology 330

Art and architecture /

Vše 369 / Architecture 57 / Arts 76 / Civic and landscape art 22 / Drawing and decorative arts 16 / Fine and decorative arts 103 / Graphic arts, printmaking and prints 14 / Painting and paintings 13 / Photography and photographs 46 / Plastic art - sculpture 17

Display: visual, text

14|15 Baťův institut
abcd : art brut
Action galleries
Adolf Wölfli : stvořitel universa
Akademie výtvarných umění v Praze
Alice Nikitinová
Alšova jihočeská galerie v Hluboké nad Vltavou : AJG
Amatérská obrazárna
Archeologické oddělení NPU Praha
Architecture week
Archiweb.cz
ARSkontakt

webarchiv.cz

cataloging and bibliographic metadata

- library system **Aleph**
- format for Bibliographic Data **MARC 21** (machine-readable cataloging)
- **RDA** (Resource Description and Access) - standard for descriptive cataloging providing instructions and guidelines on formulating bibliographic data, since 2015
- **WA-KAT** – cataloging tool, an application we developed for cataloging web resources, available at: https://kat.webarchiv.cz/
- **Cataloging manual -** recommendations on how to catalog a web resources https://webarchivcz.github.io/katalogizacni-manual/

webarchiv.cz

topic collections

collections of resources related to certain event or topic
more in depth capture of the topic in electronic resources

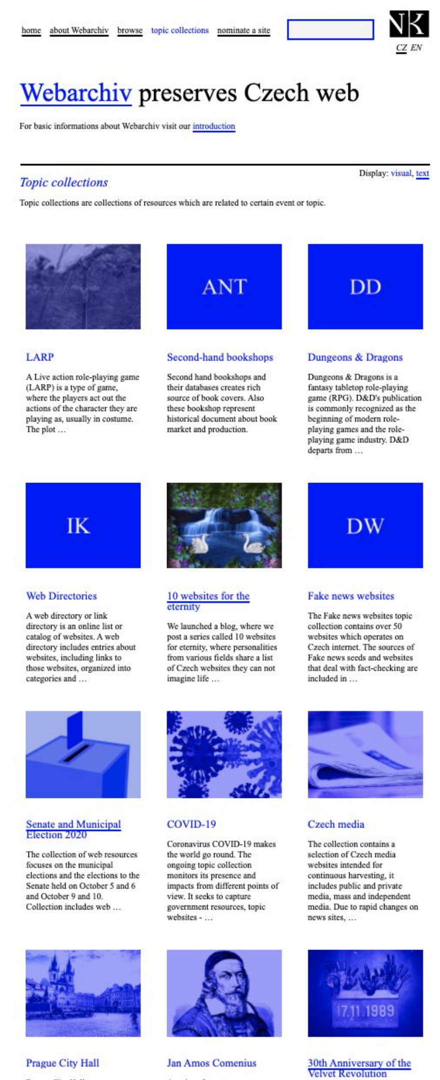
current events

- planned: elections, anniversaries
- unexpected: current political events, natural disasters

long-term collections – continuous harvesting

- Charles University, Czech media (harvested on a daily basis)

collaboration with IIPC (Olympics and Paralympics, Climate Change, Novel Coronavirus – COVID-19)



Webarchiv preserves Czech web

For basic informations about Webarchiv visit our introduction

Display: visual, text

Topic collections

Topic collections are collections of resources which are related to certain event or topic.

LARP

A Live action role-playing game (LARP) is a type of game, where the players act out the actions of the character they are playing as, usually in costume. The plot …

Second-hand bookshops

Second hand bookshops and their databases creates rich source of book covers. Also these bookshop represent historical document about book market and production.

Dungeons & Dragons

Dungeons & Dragons is a fantasy tabletop role-playing game (RPG). D&D's publication is commonly recognized as the beginning of modern role-playing games and the role-playing game industry. D&D departs from …

Web Directories

A web directory or link directory is an online list or catalog of websites. A web directory includes entries about websites, including links to those websites, organized into categories and …

10 websites for the eternity

We launched a blog, where we post a series called 10 websites for eternity, where personalities from various fields share a list of Czech websites they can not imagine life …

Fake news websites

The Fake news websites topic collection contains over 50 websites which operates on Czech internet. The sources of Fake news seeds and websites that deal with fact-checking are included in …

Senate and Municipal Election 2020

The collection of web resources focuses on the municipal elections and the elections to the Senate held on October 5 and 6 and October 9 and 10. Collection includes web …

COVID-19

Coronavirus COVID-19 makes the world go round. The ongoing topic collection monitors its presence and impacts from different points of view. It seeks to capture government resources, topic websites - …

Czech media

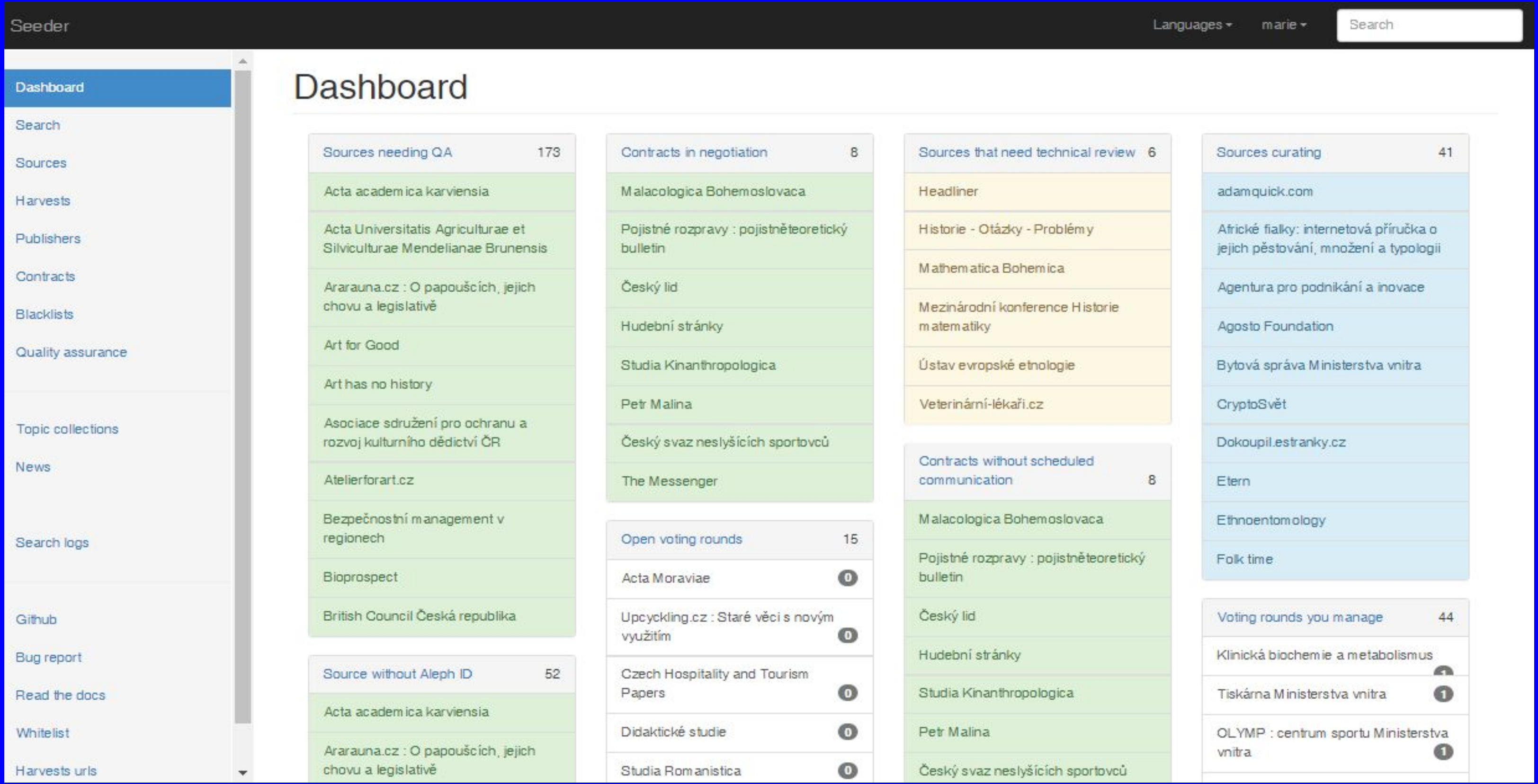
The collection contains a selection of Czech media websites intended for continuous harvesting, it includes public and private media, mass and independent media. Due to rapid changes on news sites, …

Prague City Hall

Jan Amos Comenius

30th Anniversary of the Velvet Revolution webarchiv.cz

seeder



Dashboard

Sources needing QA 173

- Acta academica karviensia
- Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis
- Ararauna.cz : O papoušcích, jejich chovu a legislativě
- Art for Good
- Art has no history
- Asociace sdružení pro ochranu a rozvoj kulturního dědictví ČR
- Atelierforart.cz
- Bezpečnostní management v regionech
- Bioprospect
- British Council Česká republika

Source without Aleph ID 52

- Acta academica karviensia
- Ararauna.cz : O papoušcích, jejich chovu a legislativě

Contracts in negotiation 8

- Malacologica Bohemoslovaca
- Pojistné rozpravy : pojistněteoretický bulletin
- Český lid
- Hudební stránky
- Studia Kinanthropologica
- Petr Malina
- Český svaz neslyšících sportovců
- The Messenger

Open voting rounds 15

Acta Moraviae	0
Upcykling.cz : Staré věci s novým využitím	0
Czech Hospitality and Tourism Papers	0
Didaktické studie	0
Studia Romanistica	0

Sources that need technical review 6

- Headliner
- Historie - Otázky - Problémy
- Mathematica Bohemica
- Mezinárodní konference Historie matematiky
- Ústav evropské etnologie
- Veterinární-lékaři.cz

Contracts without scheduled communication 8

- Malacologica Bohemoslovaca
- Pojistné rozpravy : pojistněteoretický bulletin
- Český lid
- Hudební stránky
- Studia Kinanthropologica
- Petr Malina
- Český svaz neslyšících sportovců

Sources curating 41

- adamquick.com
- Africké fialky: internetová příručka o jejich pěstování, množení a typologii
- Agentura pro podnikání a inovace
- Agosto Foundation
- Bytová správa Ministerstva vnitra
- CryptoSvět
- Dokoupil.estranky.cz
- Etern
- Ethnoentomology
- Folk time

Voting rounds you manage 44

Klinická biochemie a metabolismus	1
Tiskárna Ministerstva vnitra	1
OLYMP : centrum sportu Ministerstva vnitra	1

open source software for managing electronic resources, website and harvests, developed in-house, https://github.com/webarchivcz/

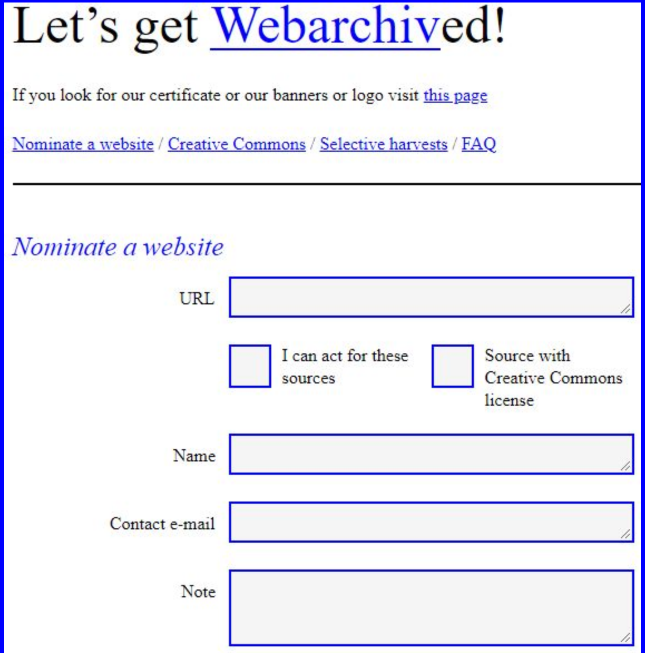
webarchiv.cz

cooperation & research

- **archiving specific resources**
  – Czech Language Institute of the Czech Academy of Sciences – Czech Literary Internet
- **methodological support for building own archives**
  – Office for supervision of economic affairs of political parties and political movements
- **IIPC, University Library in Bratislava**
  – collaborative collections
- **Development of centralized interface for extracting big data from web archives**
  – ongoing research project, making data available to the research community
  (NL CR, University of West Bohemia – Faculty of Applied Sciences, The Department of Cybernetics, Institute of Sociology of the Czech Academy of Sciences)

webarchiv.cz

user generated seeds



Let’s get Webarchived!

If you look for our certificate or our banners or logo visit this page

Nominate a website / Creative Commons / Selective harvests / FAQ

Nominate a website

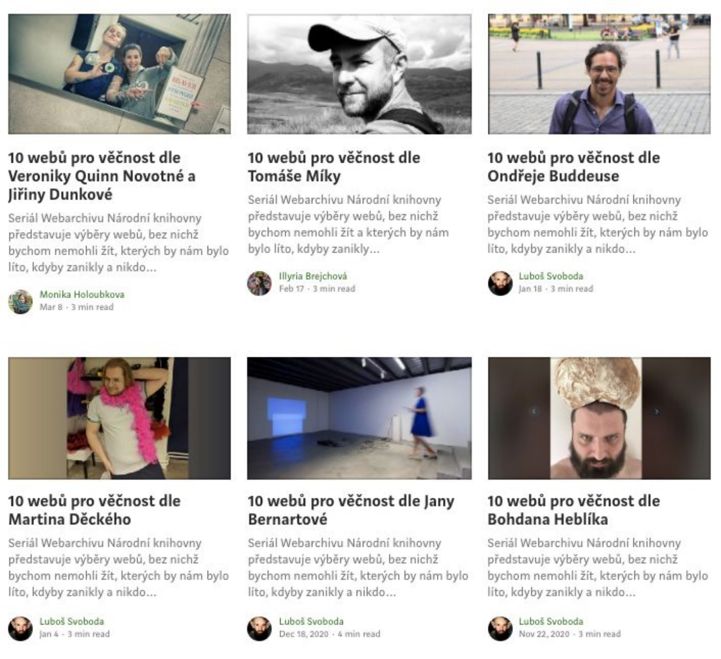
URL ▭

□ I can act for these sources □ Source with Creative Commons license

Name ▭

Contact e-mail ▭

Note ▭



10 webů pro věčnost dle Veroniky Quinn Novotné a Jiřiny Dunkové

Seriál Webarchivu Národní knihovny představuje výběry webů, bez nichž bychom nemohli žít, kterých by nám bylo líto, kdyby zanikly a nikdo…

Monika Holoubkova
Mar 8 · 3 min read

10 webů pro věčnost dle Tomáše Míky

Seriál Webarchivu Národní knihovny představuje výběry webů, bez nichž bychom nemohli žít a kterých by nám bylo líto, kdyby zanikly…

Illyria Brejchová
Feb 17 · 3 min read

10 webů pro věčnost dle Ondřeje Buddeuse

Seriál Webarchivu Národní knihovny představuje výběry webů, bez nichž bychom nemohli žít, kterých by nám bylo líto, kdyby zanikly a nikdo…

Luboš Svoboda
Jan 18 · 3 min read

10 webů pro věčnost dle Martina Děckého

Seriál Webarchivu Národní knihovny představuje výběry webů, bez nichž bychom nemohli žít, kterých by nám bylo líto, kdyby zanikly a nikdo…

Luboš Svoboda
Jan 4 · 3 min read

10 webů pro věčnost dle Jany Bernartové

Seriál Webarchivu Národní knihovny představuje výběry webů, bez nichž bychom nemohli žít, kterých by nám bylo líto, kdyby zanikly a nikdo…

Luboš Svoboda
Dec 18, 2020 · 4 min read

10 webů pro věčnost dle Bohdana Heblíka

Seriál Webarchivu Národní knihovny představuje výběry webů, bez nichž bychom nemohli žít, kterých by nám bylo líto, kdyby zanikly a nikdo…

Luboš Svoboda
Nov 22, 2020 · 3 min read

https://www.webarchiv.cz/en/add
we accept proposals for resources for archiving

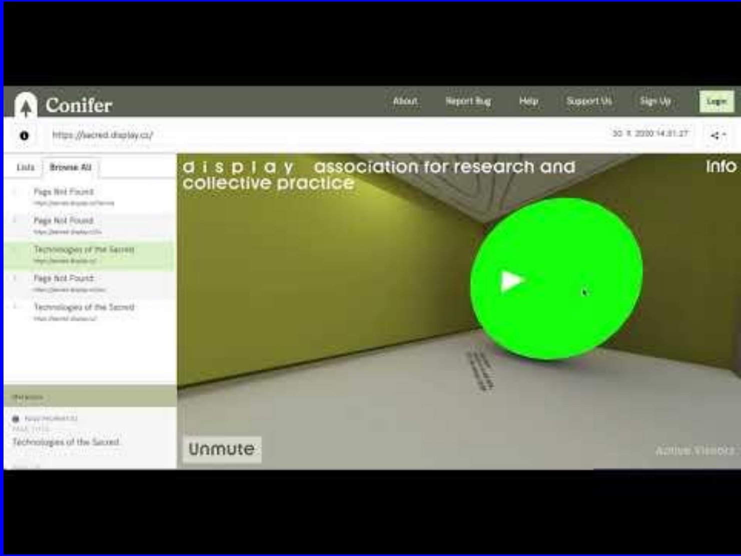
https://medium.com/webarchiv
10 websites for eternity - personalities suggest webs for archiving

webarchiv.cz

chalenges

- make the archive data and metadata as accessible to the public as possible
- social media archiving (personalized content) and dynamic content

  Webrecorder / archiveweb page / browsertrix (FB, IG, TW, dynamic websites)

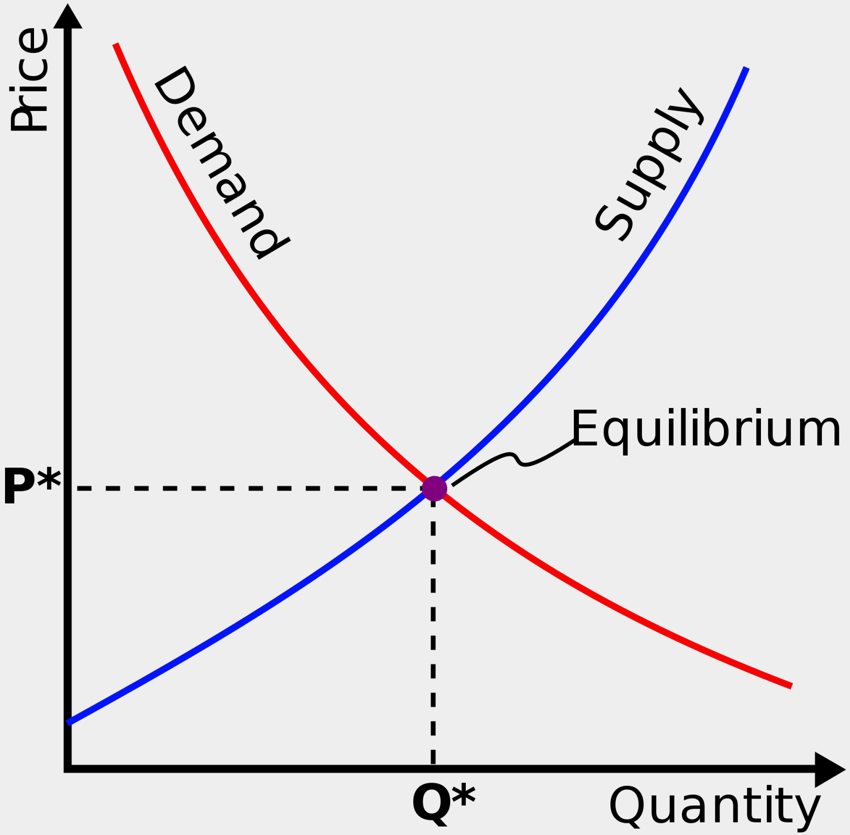

Conifer
About
Report Bug
Help
Support Us
Sign Up
Login
https://sacred.display.cz/
30. 11. 2020 14:31:27
Lists
Browse All
Page Not Found
Page Not Found
Technologies of the Sacred
Page Not Found
Technologies of the Sacred
Technologies of the Sacred
display association for research and collective practice
Info
Unmute
Active Visitors

- cooperation with research communities
- quality assurance
- automatization, data protection

IT operation

Key issues

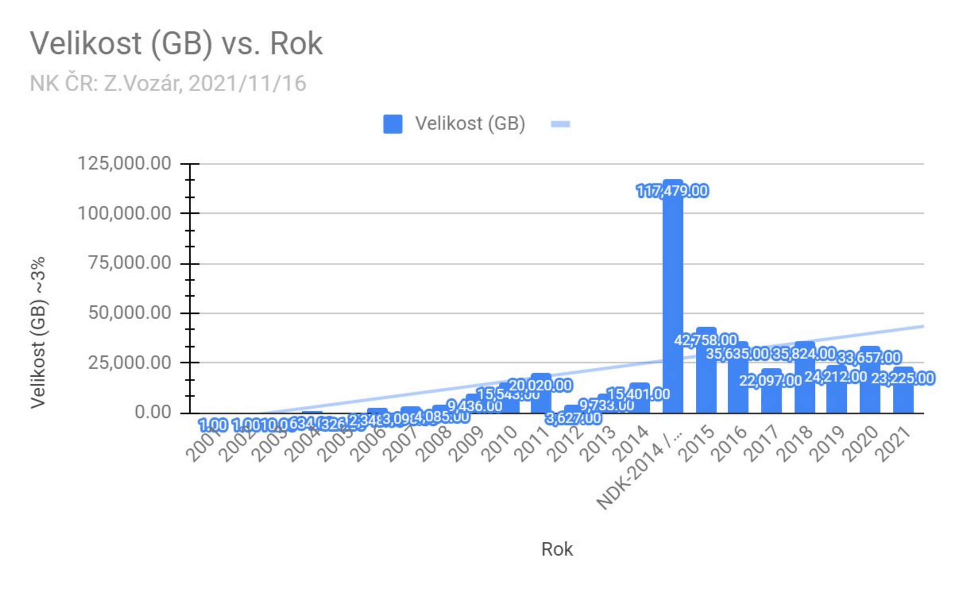
- Absorption capacity each year
- Scrapers ban
- Reliable archiving / Deduplication
- LTP logical protection
- Optimization and QA
- Automatisation
- Sufficient personal capacities
- Legal depot / questions



Czech Webarchive - Acquisition

- stored: 409 TB zipped data:
    - yearly acquisition 25-50 TB (last five years)
    - this year 23 TB so far

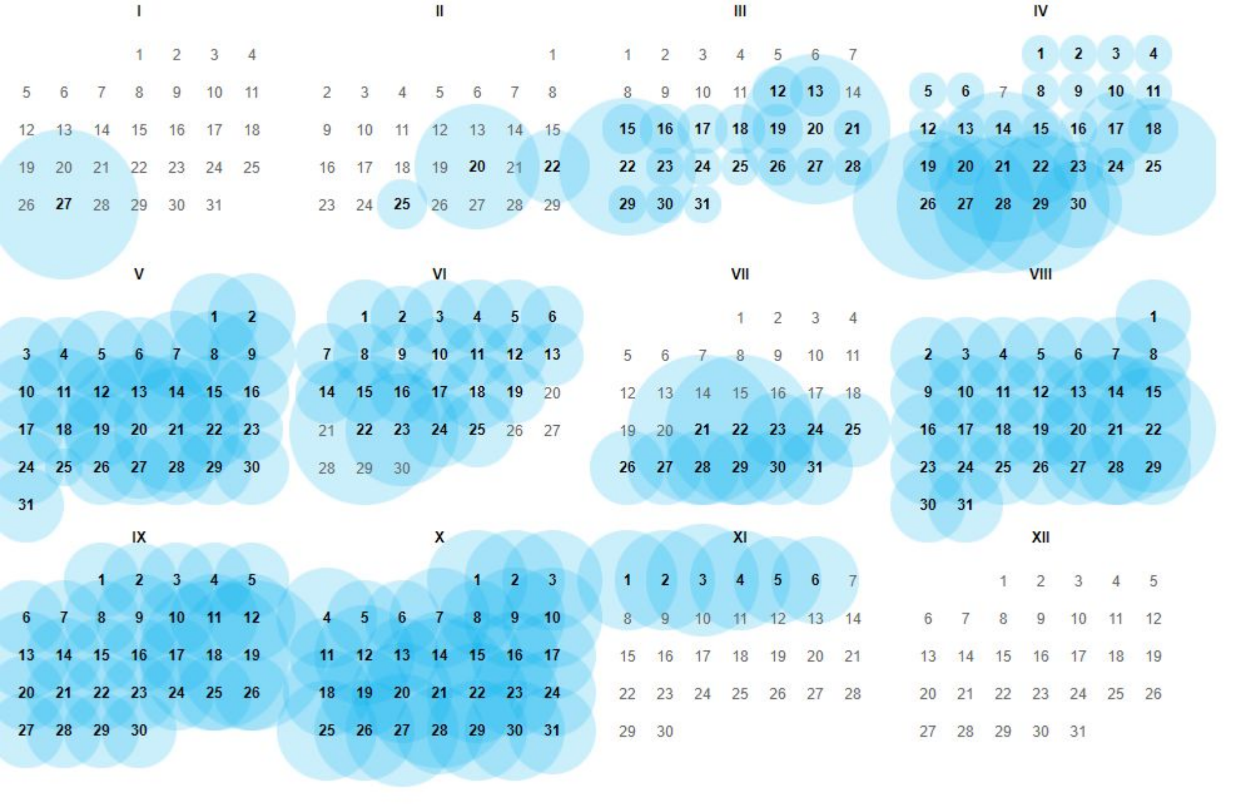
    - daily
        - Continuous - from 170 GB till 40 GB with deduplication
    - monthly
        - Serials - from 4 TB till 700 GB with deduplication
        - Topics - from 2 TB till 300 GB with deduplication
    - once/twice a year
        - Totals - from 30 TB till 15 TB - with no dedup / with dedup



Velikost (GB) vs. Rok

NK ČR: Z.Vozár, 2021/11/16

Continuous campaign Czech media
(eg. *blisty.cz*)



Key technologies and principles

- virtualisation via VMware
- hierarchical storage via Spectrum Protect + TSM
- custom networking policies / absorption capacity
- quantity of crawlers

SW Tools

PROD

Heritrix web - crawler

- version 3.4.0 (latest stable version: 3.4.0-20210923)
- diff. flavors of 3.4.0
- GH: https://github.com/internetarchive/heritrix3

Openwayback - presentation playback

- version 2.3.2
- java application used to play back archived websites
- *no longer* under active development
- customised in 2016
- GH: https://github.com/iipc/openwayback/

CDX server

- massive index of 7 bil. items
- no alone instance

Seeder - own

- Ca. 3 times a year new version

WA-KAT - own

- dockerised catalogisation SW

TEST / DEV

pywb

- HiFi Webarchive - python web archiving toolkit for replaying archive
- Implementation with OutbackCDX and pywb index cdxj
- customization of UI and integration of UKWA UI faceted search

Seeder

- Czech Webarchive curating tool
- CI / CD integration in Test env, Jenkins
- New functions for collections and crawl automatisation
- Python / Django
- open-source: https://github.com/WebarchivCZ/Seeder

Grainery + Extractor

- Revision of WA content
- Preparation of metadataset for LTP ingest
- Python / Flask

R&D - Centralized interfaces - collab. (2018 -2022)

Data mining

- WARC processing
  - AUT tools - SPARK - initial DF
  - text extraction from txt/html
  - boilerplate removal
  - sound/video text extraction
  - other formats TD
- NER, Link extraction
- Network analysis

Topic modelling

- Topic identification based on catal. metadata
- Use of deep neural networks (Kerberos + Tensorflow, Pytorch)

New storage techs

- shift or paradigm:
  - hierarchical to object storage
  - capacity is cheaper than data loss
  - I/O resiliency
- set of 6 new servers for HDFS POC
- data operations as service

R&D - Centralized interfaces - Discovery and Output

Data exploration & exportation

- datasets for scientists based on their research requirements
- analysis of topics and their automatic detection or analysis of audio files
- approaches based on deep neural networks for document classification

- 1. Discovery UI interface
  - filtration via facets (eg harvest, dates, contents, types, formats)
  - creation of collections
  - creation of very own data filtres
  - Stop words
- 2. REST API interface
  - Jupyter NB
  - JAva, Python, Scala
- 3. Export on demand
  - JSON, CSV
  - Fulltexts, Collocations, Network analysis data

R&D - Centralized interfaces - Discovery and Output

Discovery UI:

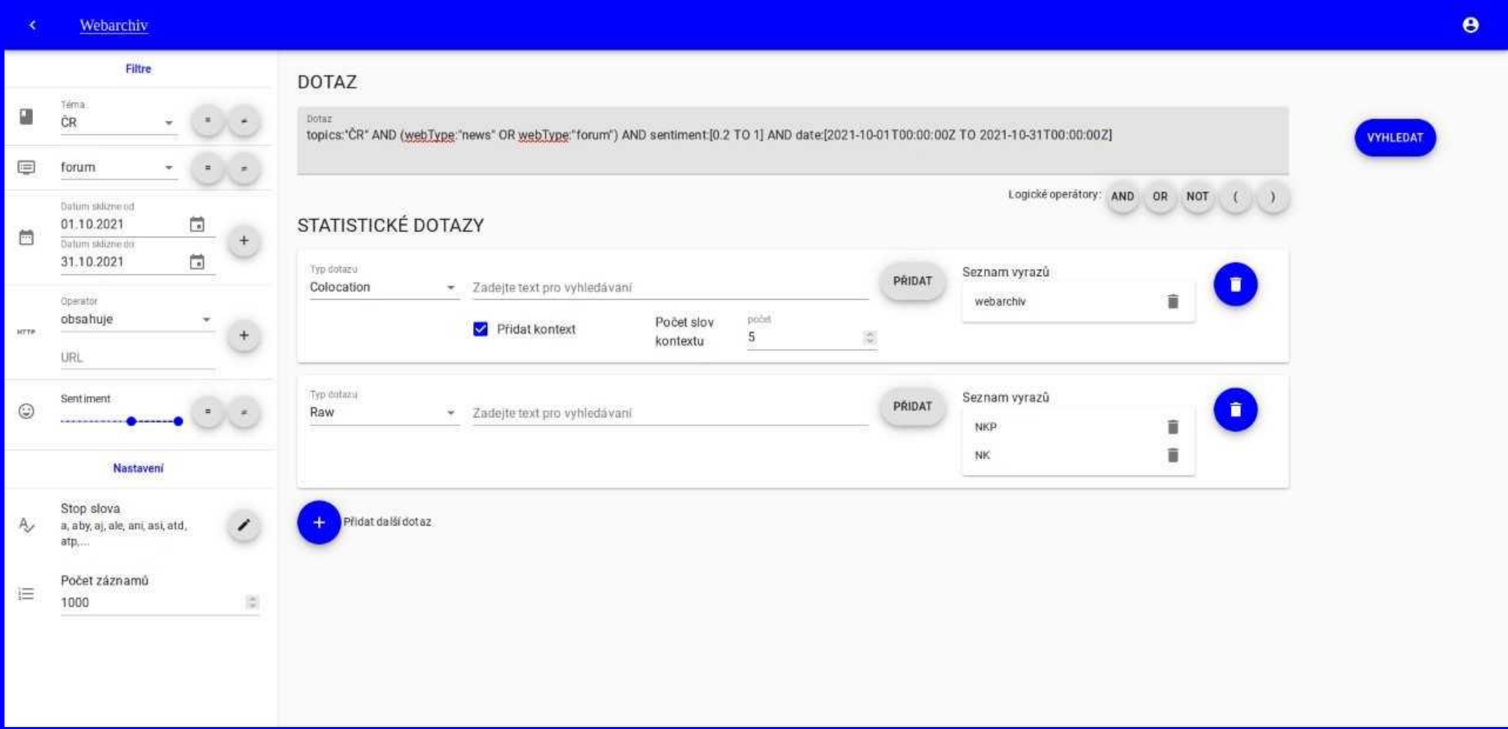
- Filtration
- Data Base

  Evocation

- Dry run

Iterative export:

- Filter It
- Extract It !



Webarchiv

Filtre

Téma: ČR

forum

Datum sklizne od: 01.10.2021

Datum sklizne do: 31.10.2021

Operator: obsahuje

URL

Sentiment

Nastavení

Stop slova: a, aby, aj, ale, ani, asi, atd, atp,...

Počet záznamů: 1000

DOTAZ

Dotaz:

```
topics:"ČR" AND (webType:"news" OR webType:"forum") AND sentiment:[0.2 TO 1] AND date:[2021-10-01T00:00:00Z TO 2021-10-31T00:00:00Z]
```

Logické operátory: AND OR NOT ( )

STATISTICKÉ DOTAZY

Typ dotazu: Colocation

Zadejte text pro vyhledávaní

☑ Přidat kontext

Počet slov kontextu: počet 5

Seznam vyrazů: webarchiv

Typ dotazu: Raw

Zadejte text pro vyhledávaní

Seznam vyrazů: NKP, NK

POC 2021/Q3

Thank you!

www.webarchiv.cz

www.facebook.com/webarchivcz, https://twitter.com/webarchiv_cz
https://www.instagram.com/mrtve_weby/

webarchiv@nkp.cz