Web archive of National Library of the Czech Republic

Marie Haškovcová, Zdenko Vozár
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
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
• **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
• **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
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
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/](https://kat.webarchiv.cz/)
- **Cataloging manual** - recommendations on how to catalog a web resources [https://webarchivcz.github.io/katalogizacni-manual/](https://webarchivcz.github.io/katalogizacni-manual/)
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.cz

Seeder

Dashboard

Sources needing QA 173
- Acta academia karvinensia
- Acta Universitatis Agriculturae et Silviculturae Mendelae Brunensis
- Araneta.cz: O populacích, jejich chovu a legislativě
- Arl for Good
- Art has no history
- Asiace sbírání pro ochranu a rozvoj kulturního dědictví ČR
- Atelier.cz
- Bezpečnostní management v regionech
- Bioprospect
- British Council Česká republika
- Source without Aleph O 52
- Acta academia karvinensia
- Araneta.cz: O populacích, jejich chovu a legislativě

Contracts in negotiation 0
- Malacologia Bohemoslovačka
- Pojistné rozpravy : pojistnětheoretický bulletin
- Český lid
- Hudební stránky
- Studia Kinematografica
- Petr Malina
- Český svaz neslyšících sportovců
- The Messenger

Sources that need technical review 0
- Headliner
- Historie - Dějiny - Problémy
- Mathematica Bohemica
- Mezinárodní konference Historie matematiky
- Ústav evropské etnologie
- Veterinární lékaři.cz

Contracts without scheduled communication 0
- Malacologia Bohemoslovačka
- Pojistné rozpravy : pojistnětheoretický bulletin
- Český lid
- Hudební stránky
- Studia Kinematografica
- Petr Malina
- Český svaz neslyšících sportovců

Sources curating 41
- adam beard
- Atrocíci českého internetu: příběh o jejich pěstování, množení a typologii
- Agentura pro podnikání a inovace
- Agosto Foundation
- Bytová správa Ministerstva unitr
- Cryptotev
- Dokioupladávečky.cz
- Etern
- Ethnomontology
- Folk lines

Open voting rounds 15
- Acta Moraviae
- Upeckydlence.cz: Staré věci a nový vydátkem
- Czech Hospitality and Tourism Papers
- Datadokumentace
- Studia Kinematografica
- Petr Malina
- Český svaz neslyšících sportovců

Voting rounds you manage 44
- Klinická biochemie a metabolismus
- Tiskárna Ministerstva unitr
- OLYMP : centrum sportu Ministerstva unitr

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

<table>
<thead>
<tr>
<th>URL</th>
<th>I can act for these sources</th>
<th>Source with Creative Commons license</th>
<th>Name</th>
<th>Contact e-mail</th>
<th>Note</th>
</tr>
</thead>
</table>

**https://www.webarchiv.cz/en/add**

We accept proposals for resources for archiving.
• 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))
• 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
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**

**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](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/](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

**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](https://github.com/WebarchivCZ/Seeder)

**Grainery + Extractor**
- Revision of WA content
- Preparation of metadataset for LTP ingest
- Python / Flask

**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 filters
  - 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!

POC 2021/Q3
Thank you!

www.webarchiv.cz

https://www.instagram.com/mrtve_weby/

webarchiv@nkp.cz