

# Webarchiv

*Czech web archive of National Library of the Czech Republic*

*Curatorial approaches, topic collections and cooperation with the research communities*

# We are Webarchiv

the digital library, which preserves web sites for future generations.

Without continuous web content preserving, the significant part of national heritage would be lost.

WWW



# *History*

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



# Univerzita Karlova v Praze

Vyhledávání

## Základní informace

Fakulty a další součásti

Studium

Věda a granty

Zahraníční styky

Informační služby

Co je nového

Staré stránky naleznete zde

## Základní informace

- **Kontakt**
  - Adresa : Ovocný trh 5, Praha 1, 116 36, Czech Republic
  - Telefon : +420 2 24491111
  - Fax : +420 2 24210695
  - E-mail : UK@cuni.cz
- **Základní dokumenty**
  - Základní listina
  - Historie
  - Statut a další vnitřní předpisy
- **Orgány UK**
  - Rektor
  - Akademický senát
  - Vědecká rada
  - Správní rada
  - Kvestor
  - **Poradní orgány rektora**
    - Kolegium rektora, prorektorů
    - Grantová rada
    - Ediční rada
    - Historická komise
- **Rektorát**
- **Struktura a seznam osob**
- **Úřední deska**
- **Carolinum - Spolek absolventů a přátel Univerzity Karlovy**



**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](#)

# *Today*

- 385 TB archived data
- 9,5 billion digital objects (text, images, audio and video objects, software, scripts, etc.)
- 3,5 people in the department + 1 IT guy

# How do we archive?

We perform a complete archivation  
of the „entire“ Czech web.

The selected websites  
are harvested contemporaneously.

W

W

W

W

W

W

W

W



# Unfortunately,

not entire data are available online.

This is caused by current legal state of copyright.  
The entire Webarchiv data are accessible only  
in the library building.

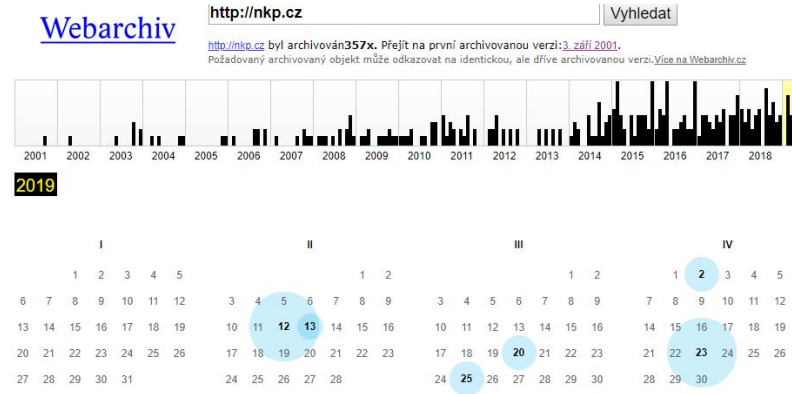


# *Legal Issues*

- **Copyright act** – Library Licence allows library to make a reproduction of a work for its own archiving and conservation purposes
- **Legal deposit act** – does not cover born digital documents
- **Online access** – based on contract with publishers or on Creative Commons licence

# Software

- crawler: Heritrix 3.4
- access: Open Wayback 2.3.1



- curators: Seeder (developed in-house, available on github <https://github.com/webarchivcz/>)

## Dashboard

Search

Sources

Harvests

Publishers

Contracts

Blacklists

Quality assurance

Topic collections

News

Search logs

Github

Bug report

Read the docs

Whitelist

Harvests urls

## Dashboard

|  |     |   |    |   |   |   |    |
|--|-----|---|----|---|---|---|----|
| Sources needing QA   | 179 | Contracts in negotiation                        | 8  | Sources that need technical review              | 6 | Sources curating  | 41 |
| Acta academica karviensia  |     | Malacologica Bohemoslovaca                      |    | Headliner                                       |   | adamquick.com   |    |
| Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis |     | Pojistné rozpravy : pojistněteoretický bulletin |    | Historie - Otázky - Problémy                    |   | Atrické fičky: internetová příručka o jejich pěstování, množení a typologii |    |
| Ararauna.cz : O papoušcích, jejich chovu a legislativě                 |     | Český lid                                       |    | Mathematica Bohemica                            |   | Agentura pro podnikání a inovace  |    |
| Art for Good   |     | Hudební stránky                                 |    | Mezinárodní konference Historie matematiky      |   | Agosto Foundation   |    |
| Art has no history   |     | Studia Kinanthropologica                        |    | Ústav evropské etnologie                        |   | Bytová správa Ministerstva vnitra   |    |
| Asociace sdružení pro ochranu a rozvoj kulturního dědictví ČR          |     | Petr Malina                                     |    | Veterinární-lékaři.cz                           |   | CryptoSvět  |    |
| Atelierforart.cz   |     | Český svaz neslyšících sportovců                |    |   |   | Dokoupil.estranky.cz  |    |
| Bezpečnostní management v regionech                                    |     | The Messenger                                   |    | Contracts without scheduled communication       | 8 | Etern   |    |
| Bioprospect  |     |   |    | Malacologica Bohemoslovaca                      |   | Ethnoentomology   |    |
| British Council Česká republika  |     | Open voting rounds                              | 15 | Pojistné rozpravy : pojistněteoretický bulletin |   | Folk time   |    |
|  |     | Acta Moraviae                                   | 0  | Český lid                                       |   |   |    |
| Source without Aleph ID  | 52  | Upcycling.cz : Staré věci s novým využitím      | 0  | Hudební stránky                                 |   | Voting rounds you manage  | 44 |
| Acta academica karviensia  |     | Czech Hospitality and Tourism Papers            | 0  | Studia Kinanthropologica                        |   | Klinická biochemie a metabolismus   | 1  |
| Ararauna.cz : O papoušcích, jejich chovu a legislativě                 |     | Didaktické studie                               | 0  | Petr Malina                                     |   | Tiskárna Ministerstva vnitra  | 1  |
|  |     | Studia Romanistica                              | 0  | Český svaz neslyšících sportovců                |   | OLYMP : centrum sportu Ministerstva vnitra                                  | 1  |

*Seeder – software for managing electronic resources, websites and harvests*

# Collection policy

- Comprehensive harvests
- Selective harvests
- Topic collections

```
653         </td><td>
654             <div class="date">
655                 <div class="position">
656                     <div class="hidden">2</div>
657                     <div class="measure opacity20" id="5Fe-25-2019"><img width="100%" height="100%" />
658                 </div>
659             </div>
660         </td><td>
661             <div class="date"></div>
662         </td><td>
663             <div class="date"></div>
664         </td><td>
665             <div class="date"></div>
666         </td><td>
667             <div class="date"></div>
668         </td><td>
669             <div class="date"></div>
670         </td></tr><tr><td>
671             <div class="date"></div>
672         </td><td></td><td></td></td></td></td></td></td></td>
673     </tr>
674 </tbody>
675 </table>
676 </div>
677
678
679 <div class="month" id="2019-3">
680     <table>
681         <thead>
682             <tr>
683                 <th colspan="7"><span class="label"></span></th>
684             </tr>
685         </thead>
686         <tbody>
687             <tr>
688                 <td><div class="date"></div></td><td>
689                     <div class="date"></div>
690                 </td><td>
691                     <div class="date">
692                         <div class="position">
693                             <div class="hidden">2</div>
694                             <div class="measure opacity20" id="Dub-2-2019"><img width="100%" height="100%" />
695                         </div>
696                     </div>
697                 </td><td>
698                     </div></td>
```

# *Comprehensive harvests*

- contract with czech domain provider CZ.NIC
- once or twice a year crawl of the whole .cz domain
- accessible only in the library
- 1,4 millions of second order domains / domain.cz
- maximum of 5000 harvested files per site

Vložte webovou adresu

# Tuto stránku Webarchiv nemůže zobrazit

Z důvodu autorského zákona nemůžeme tuto stránku zpřístupnit online. Archivované verze této stránky jsou dostupné pouze z [Referenčního centra NK ČR](#). Pro více informací o zpřístupnění navštivte naše [často kladené dotazy](#).

*“Archived versions of this page are only available from the Reference Centre of the National Library of the Czech Republic.”*

# *Selective harvests*

- selective approach
  - bohemical character (territory, language, authorship, topic/content), not only on czech domain
  - resources with historical, scientific or cultural value
- curated resources
- online access – contract or Creative Commons
  - more than 5000 archived websites with online access
- crawled periodically
- maximum of 15 000 harvested files per site



# Browse the [Webarchiv](#) by subject

List of a contracted websites by classification system:

[Vše](#) 5172 / [Agriculture](#) 176 / [Anthropology](#) 164 / [Art and architecture](#) 349 / [Beletry](#) 29 / [Biological sciences](#) 224 / [Business and economics](#) 314 / [Computer sciences](#) 147 / [Education](#) 222 / [Engineering and technology](#) 280 / [Geography and earth sciences](#) 408 / [History and auxiliary sciences](#) 293 / [Chemistry](#) 49 / [Children's literature](#) 1 / [Language, linguistics and literature](#) 213 / [Law](#) 147 / [Library science, generalities and references](#) 294 / [Mathematics](#) 45 / [Medicine](#) 284 / [Music](#) 156 / [Performing arts](#) 167 / [Philosophy and religion](#) 238 / [Physical education and recreation](#) 213 / [Physical sciences](#) 88 / [Political science](#) 338 / [Psychology](#) 83 / [Sociology](#) 315

## *Art and architecture / Arts*

[Vše](#) 349 / [Architecture](#) 55 / [Arts](#) 72 / [Civic and landscape art](#) 21 / [Drawing and decorative arts](#) 15 / [Fine and decorative arts](#) 98 / [Graphic arts, printmaking and prints](#) 11 / [Painting and paintings](#) 13 / [Photography and photographs](#) 45 / [Plastic art - sculpture](#) 17

---

Display: [visual](#), [text](#)



[abcd : art brut](#)



[Adolf Wolfli : stvoritel universa](#)



[Arte-fakt : sdružení pro ochranu památek](#)



[Art for Good](#)



[Artforum : osobnosti v síti](#)



[Artiki](#)



[Arts Lexikon : on-line](#)



[Asociace sdružení pro](#)

# Curators – workflow

- selecting and evaluating resources
- contracting with publishers
- cataloging (RDA rules, conspectus method)
- access and quality assurance

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

# *Topic collections*

collections of resources related to certain event or topic

deeper capture of the topic in electronic resources

***current events*** — harvesting usually in several stages: before, during and after the event

- planned: elections, anniversaries
- unexpected: floods, terrorist attacks

***long-term collections*** — continuous harvesting

- Creative Commons, Periodical publications, Charles University

***collaboration with IIPC*** (Olympics and Paralympics, Climate Change, Artificial Intelligence)



## Webarchiv preserves Czech web

For basic informations about Webarchiv visit our [introduction](#)

### *Topic collections*

Display: [visual](#), [text](#)

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



### [Floods 2013](#)

In 2013 a large part of the Czech Republic hit a flood. In thematic harvest, we tried to capture feedback in space of Czech internet, specialized websites, responses of authorities, ...



### [Olympic Prague](#)

On the basis of the resolution of 22 March 2007, the Prague City Hall decided to apply for the Olympic Games in 2016. The information about this decision, its response ...



### [Charles IV. - 700th anniversary.](#)

On the occasion of the 700th anniversary of the birth of Charles IV., we prepared a thematic collection of web sources, that currently relate from various perspectives to jubilee of ...



### [The National Archives - Public Authorities](#)

The National Archives, by law, supervises the removal of documents and the scrutiny of the records service of a number of organizations. The topic collection includes websites of public authorities ...



### [Climate Change](#)

Climate change is one of the greatest challenges of today. Finding solutions at scientific and political levels is accompanied by intense societal debate. The collection, which was created in collaboration ...



### [European Parliament Election 2019](#)

On 24 and 25 May, the Czech Republic elects its representatives to the European Parliament. The topic collection includes websites of political parties, their leaders and responses to the political ...



### [New Building of the National technical library.](#)

In 2006 construction works have been started in order to build up new building of the National technical library located on Fleming square (Dejvice) next to the Czech Technical University ...



### [Václav Havel](#)

The topic harvest dedicated to the death of the first Czechoslovak and Czech president Václav Havel, who died on December 18, 2011, took place in several rounds of the turn ...



### [European Parliament Election 2014](#)

The European Parliament Election in May 2014 showed record low turnout of Czech voters. The thematic collection maps the election campaigns of political parties and their leaders, results and the ...



Explore >> International Internet Preservation Consortium >> 2018 Winter Olympics and Paralympics



## 2018 Winter Olympics and Paralympics

Collected by: [International Internet Preservation Consortium](#)

Archived since: [Led, 2018](#)

Description: A collection of websites related to the 2018 Winter Olympic and Paralympic Games, held in Pyeongchang, South Korea. International Internet Preservation Consortium member institutions contributed suggested websites for inclusion in the collection. Public seed nominations were also solicited and received.

Subject: [Society & Culture](#), [Olympics](#), [Paralympic Games](#), [Olympic Winter Games \(23rd - 2018 - Pyeongch'ang-gun, Korea\)](#)

Creator: [International Internet Preservation Consortium](#)

Collector: [International Internet Preservation Consortium](#)

### Narrow Your Results

Subject Sort By: Count | (A-Z)

- Athletes/teams (53)
- General news/Commentary (36)
- Other (8)
- Social and Economic Issues (2)
- Fandom (1)

Language Sort By: Count | (A-Z)

Czech (102)

Olympic Sport Sort By: Count | (A-Z)

- Multiple sports (35)
- Biathlon (5)
- Figure Skating (4)
- Ice Hockey (4)
- Snowboard (4)

[More ▾](#)

Event Sort By: Count | (A-Z)

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.

[Search](#) [Clear](#)

The following results were found for the term(s): **czech**

- 102 Sites were found.
- Additional results for **czech** may be found by searching within the page text.

[Sites](#) [Search Page Text](#)

Page 1 of 2 (102 Total Results) [Next Page ▶](#)

Sort By: [Best Match](#) | [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Title: **Czech team**

URL: <http://czechteam.info/>

Captured 5 times between [Uno 9, 2018](#) and [Uno 27, 2018](#)

Videos: [377 Videos Captured](#)

Subject: [Athletes/Teams](#)

Language: [Czech](#)

Olympic Sport: [Multiple sports](#)

Event: [Winter Olympics](#)

Website type: [Athletic teams](#)

Country: [Czech Republic](#)

# *Challenges*

- make the archive as accessible to the public as possible (legislative restrictions)
- collection policy – collection profiling
- full-text search
- development of tools for working with archived data, big data and metadata
- cooperation with the research communities

## *Current cooperation with the researchers / institutions*

- methodological support for building own archives

The National Archives, Czech Academy of Sciences, Office for supervision of economic affairs of political parties and political movements

- topic collections

The National Archives – Public Authorities

- archiving specific resources

Czech Language Institute of the Czech Academy of Sciences (periodical publications)

# The National Archives - Public Authorities

Keywords of harvest:

[státní správa](#), [Národní archiv](#), [veřejná správa](#), [ministerstva](#)

---



The National Archives, by law, supervises the removal of documents and the scrutiny of the records service of a number of organizations. The topic collection includes websites of public authorities including ministries, central government offices, legal entities established by law, and organizational units of the state, including contributory organizations established by them, as well as top judicial institutions.

## [Agentura ochrany přírody a krajiny ČR](#)

[www.ochranaprirody.cz](http://www.ochranaprirody.cz) [[current](#)]

## [Antidopingový výbor ČR](#)

[antidoping.cz](http://antidoping.cz) [[current](#)]

## [Asociace neprofesionálních komorních a symfonických těles : ANKST](#)

[ankst.cz](http://ankst.cz) [[current](#)]

## [Celní správa České republiky](#)

[www.cs.mfcr.cz/cmsgrc/](http://www.cs.mfcr.cz/cmsgrc/) [[current](#)]

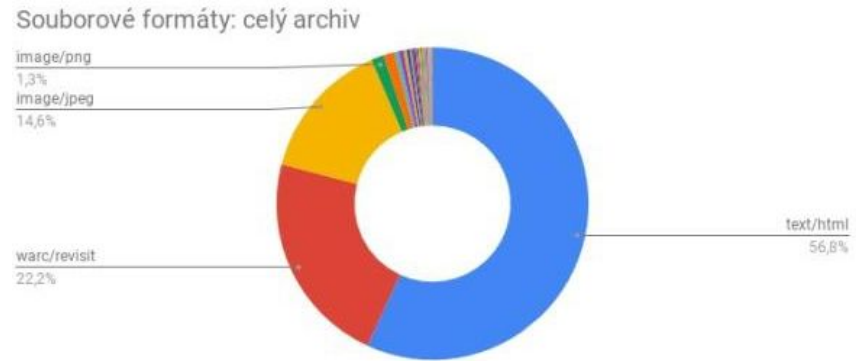


# *Development of centralized interface for extracting big data from web archives – research project*

- Webarchiv – National Library of the Czech Republic
- University of West Bohemia – Faculty of Applied Sciences, The Department of Cybernetics
- Institute of Sociology of the Czech Academy of Sciences

## *First steps:*

- legislative analysis
- index analysis
- analysis of provenance, authenticity and technical parameters of archive data
- workshop for researchers



*analysis of file formats in Webarchiv*

project accepted into the program of the Ministry of Culture which helps to support applied research and experimental development of national and cultural identity (NAKI)

# Future

The present day we work on a creating the full-text search. We recognize the importance of understanding of the digital objects and their historical meaning.

Webarchiv expects opening our data to an analytic data exploration and connection with other web archiving initiatives.

The logo consists of three stylized 'W' characters in a blue serif font. The first 'W' is smaller and positioned to the left of the other two, which are larger and more prominent.

*Thank you for your attention*

Marie Haškovcová

[www.webarchiv.cz](http://www.webarchiv.cz)

[www.facebook.com/webarchivcz](https://www.facebook.com/webarchivcz)

[marie.haskovcova@nkp.cz](mailto:marie.haskovcova@nkp.cz)

