



Centralized interface for extracting big data  
from web archives – new perspectives for web  
archive data research

*Tomáš Foltýn – Marie Haškovcová*

# Webarchiv

Czech web archive of the National Library of the Czech Republic

2000 – project of National Library of the CR, Moravian Library  
and Masaryk University

2022 – 440 TB of data, 25–40 TB per year

[home](#) [about Webarchiv](#) [browse](#) [topic collections](#) [nominate a site](#)



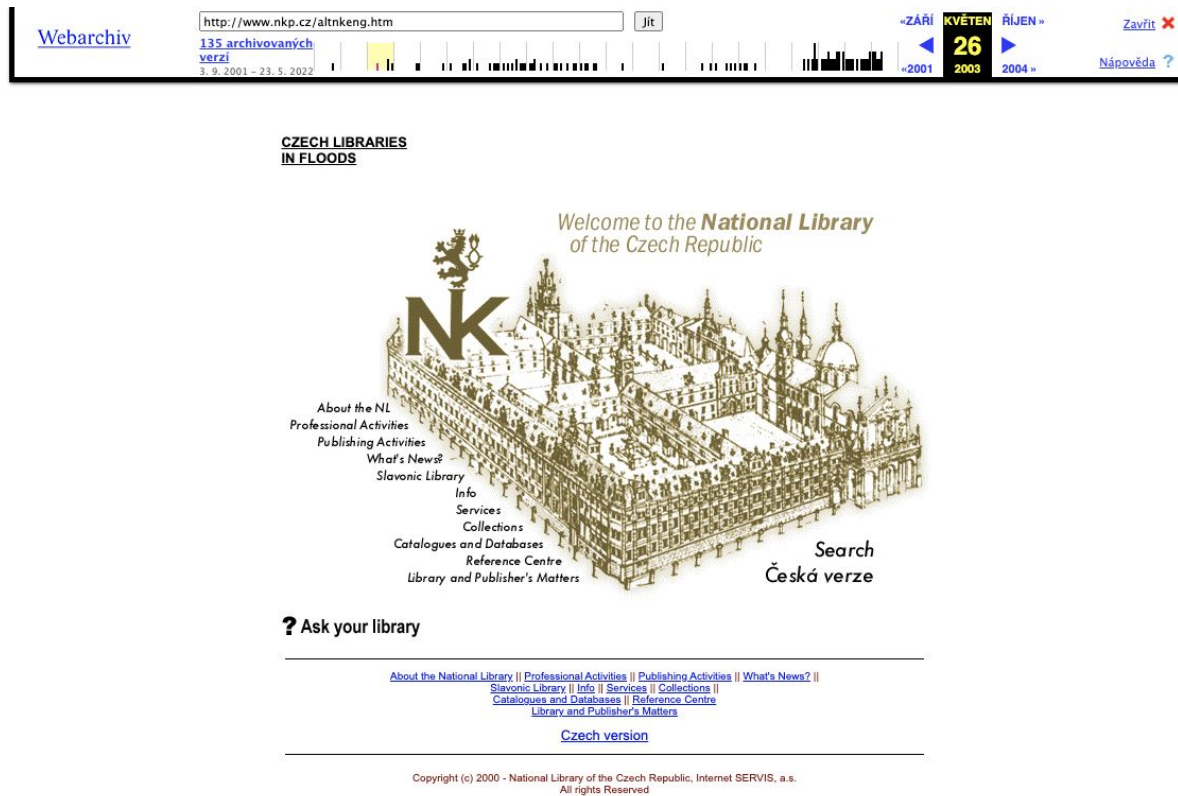
[CZ](#) [EN](#)

Webarchiv

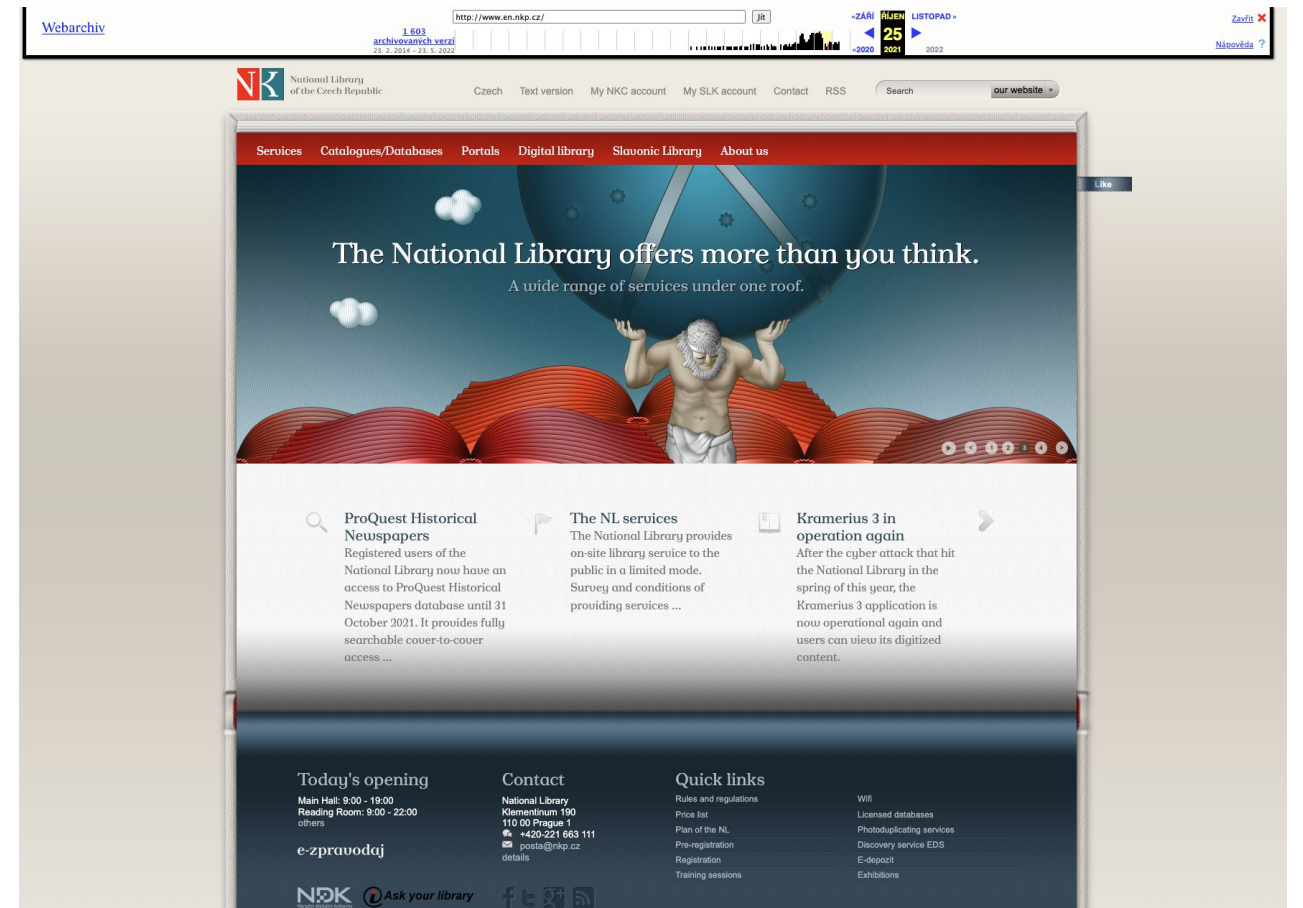
*the Museum of Czech web, [more](#)*

search „webarchiv.cz“ or „webarchiv“

<https://www.webarchiv.cz/en/>



2003



2021

*archive copy of National Library of the Czech Republic website*

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

Ministry of Culture of the Czech Republic – NAKI (Applied research and development of national and cultural identity programme), 2018 – 2022

- **National Library of the Czech Republic**  
(data, experience with web archiving, infrastructure)
- **University of West Bohemia – Faculty of Applied Sciences,  
The Department of Cybernetics**  
(machine processing of large volumes of data, SW solutions)
- **Institute of Sociology of the Czech Academy of Sciences**  
(research community in the social sciences)

## *Goals of the research project*

- create advanced data extraction interface that would allow work with large amounts of data
- solving the problem of accessing data from the **Czech web archive** and providing them to the research community
- expansion of the technological infrastructure, which should no longer serve only for the preservation of funds, but also for the possibilities of further analytical data processing
- evaluation of the legal framework for working with big data

## *Project outputs*

- faceted and full-text search engine that will allow researchers to define part of the data for their research
- graphical user interface convenient for users
- an export application that allows researchers to obtain data sets for further scientific and research use





# *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** – contract with publishers or on Creative Commons licence

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

**Directive of the European Parliament and of the Council on Copyright in the Digital Single Market**



# 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
- long-term harvesting

- **Topic collections**

- collections of resources related to certain event or topic

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

Are you a human?

I'm not a robot  reCAPTCHA  
Privacy - Terms

Přidat web

to

# Selective harvests

- 5300 resources available online
- cataloging record in Czech national bibliography

## Browse the [Webarchiv](#) by subject

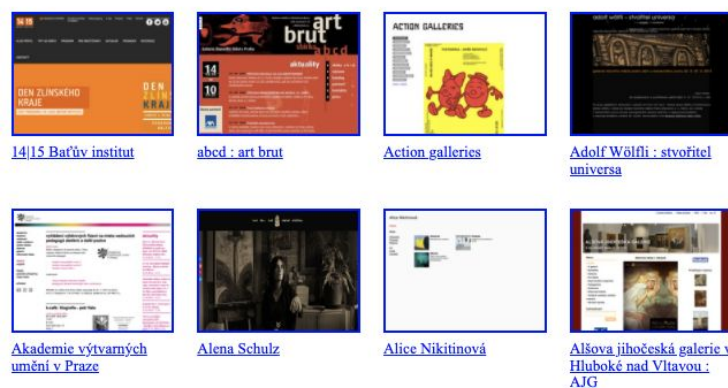
List of a contracted websites by classification system:

[Vše](#) 5347 / [Agriculture](#) 244 / [Anthropology](#) 244 / [Art and architecture](#) 378 / [Beletry](#) 40 / [Biological sciences](#) 300 / [Business and economics](#) 412 / [Chemistry](#) 51 / [Children's literature](#) 7 / [Computer sciences](#) 206 / [Education](#) 246 / [Engineering and technology](#) 309 / [Geography and earth sciences](#) 441 / [History and auxiliary sciences](#) 331 / [Language, linguistics and literature](#) 283 / [Law](#) 155 / [Library science, generalities and references](#) 329 / [Mathematics](#) 46 / [Medicine](#) 309 / [Music](#) 164 / [Performing arts](#) 261 / [Philosophy and religion](#) 257 / [Physical education and recreation](#) 231 / [Physical sciences](#) 140 / [Political science](#) 395 / [Psychology](#) 89 / [Sociology](#) 339

### Art and architecture /

[Vše](#) 378 / [Architecture](#) 58 / [Arts](#) 77 / [Civic and landscape art](#) 23 / [Drawing and decorative arts](#) 16 / [Fine and decorative arts](#) 105 / [Graphic arts, printmaking and prints](#) 14 / [Painting and paintings](#) 13 / [Photography and photographs](#) 46 / [Plastic art - sculpture](#) 17

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



# Topic collections

- current event or long-term collection
- cooperation with the IIPC

## War in Ukraine

Keywords of harvest:

[ozbrojené konflikty](#), [válečné oběti](#), [Rusko](#), [uprchlíci](#), [Ukrajina](#), [mezinárodní konflikty](#)



[ct24.ceskatelevize.cz](#)

<https://ct24.ceskatelevize.cz/specialy/3432412-rusko-ukrajinsky-konflikt> [\[current\]](#)

[www.mzv.cz](#)

[https://www.mzv.cz/jnp/cz/cestujeme/aktualni\\_doporuceni\\_a\\_varovani/ukr](https://www.mzv.cz/jnp/cz/cestujeme/aktualni_doporuceni_a_varovani/ukr) [\[current\]](#)

[www.mvcr.cz](#)

<https://www.mvcr.cz/clanek/informace-pro-obcany-ukrajiny.aspx> [\[current\]](#)

[czechia.mfa.gov.ua](#)

<https://czechia.mfa.gov.ua/> [\[current\]](#)

[brno.mfa.gov.ua](#)

<https://brno.mfa.gov.ua/> [\[current\]](#)

On February 24th Russia invaded Ukraine when it launched an invasion in full scale attack by bombarding military targets. Majority of the Western world Russian aggression towards sovereign state strictly condemned as unprovoked and unjustified. In result, these actions leads into escalation of the Russia-Ukraine conflict. As part of the continuously updated collection we focused on the reflection of the ongoing conflict in Czech media, including public service media, independent media, government sites, volunteer activities, humanitarian and

# *Metadata*

## *Cataloging and bibliographic metadata*

- library system **Aleph**
- format for Bibliographic Data **MARC 21**
- **RDA**, since 2015

## *Technical and administrative metadata*

- information about the file format, technical data obtained during the harvest, e.g. start and end date of the harvest, its type or author
- relate to harvests, the container format and the index

# *Centralized interface for extracting big data from web archives*

- analysis text document topics and their automatic detection, analysis of audiofiles, approaches based on deep neural networks for document classification, automatic analytical tools, automatic assignment of metadata to individual documents
- **infrastructure** – HADOOP cluster and Hbase solutions  
used technologies: WARC – PySpark – Python – Scikit-Learn – Apache Hbase – Hadoop – Apache Spark
- **intermediary format** = format of records during web processing archive, defines items for raw downloaded data and metadata, each web page = 1 JSON object, to which individual algorithms store the retrieved metadata  
[https://github.com/NLCR/Centralized\\_interfaces-FE](https://github.com/NLCR/Centralized_interfaces-FE)

# *Graphical user interface*

- faceted full text search engine for analyzing large quantities of web archive data
- integrated application for exporting selected datasets for scientists based on their research requirements

[Webarchiv](#)

**Sign in**

Centralized interface for extracting big data from  
web archives

**Log in**

FILTERS

**Theme**

Theme  
divadlo

**Page type**

Page type

**Date of harvest**

From

To

**URL**

Operator  
Contain

URL

**Sentiment**

Sentiment

SETTINGS OF LIMITS

**Stop words**

a, aby, aj, ale, ani, asi, atd, atp,...

**Number of entries**

Number of entries  
1000

Random records?

Random seed  
255006

QUERY

topics:"knihovna a muzeum" OR

Logical operators:

HARVESTS

<p><b>Se</b> <input type="checkbox"/></p> <p>Harvest's name: Serials-2021-02-1M-2M_OneShot-crawler00</p> <p>Size: 4.7 TB</p> <p>Number of WARC's: 1,221,667</p> <p>Date of start: 2/26/2021</p>	<p><b>Se</b> <input type="checkbox"/></p> <p>Harvest's name: Serials-2021-01-1M-4M_OneShot-crawler00</p> <p>Size: 991 MB</p> <p>Number of WARC's: 3,920,197</p> <p>Date of start: 1/22/2021</p>	<p><b>ex</b> <input type="checkbox"/></p> <p>Harvest's name: example</p> <p>Size: 322 GB</p> <p>Number of WARC's: 4</p> <p>Date of start: 6/23/2020</p>
<p><b>Te</b> <input type="checkbox"/></p> <p>Harvest's name: Test-2019-06-30-cc-naki-crawler00</p> <p>Size: 260 GB</p> <p>Number of WARC's: 3,756</p> <p>Date of start: 7/1/2019</p>	<p><b>Vo</b> <input type="checkbox"/></p> <p>Harvest's name: Volby-2013-01</p> <p>Size: 296 GB</p> <p>Number of WARC's: 686</p> <p>Date of start: 11/12/2013</p>	<p><b>Vo</b> <input type="checkbox"/></p> <p>Harvest's name: Volby-2013-00</p> <p>Size: 676 GB</p> <p>Number of WARC's: 4,322</p> <p>Date of start: 10/9/2013</p>
<p><b>Te</b> <input type="checkbox"/></p> <p>Harvest's name: Tests-2021-03-T-UDHPSH_QAtestII</p> <p>Size: 30 GB</p> <p>Number of WARC's: 31</p> <p>Date of start: 3/29/2021</p>	<p><b>Rq</b> <input type="checkbox"/></p> <p>Harvest's name: Requests-2019-12-25_special</p> <p>Size: 2.7 GB</p> <p>Number of WARC's: 3</p> <p>Date of start: 12/25/2019</p>	



# Query, Filtres, Settings od limits, Harvests

topics:"knihovna a muzeum" OR topics:"divadlo"NOT webType:"eshop"date:[2021-08-01T00:00:00Z TO 2022-05-20T00:00:00Z]url:/\*nkp.cz.\*/sentiment:[0 TO 1]

**Webarchiv** [New query](#) [My queries](#) [User](#)

**FILTERS**

**Theme** =  
Theme ≠

**Page type** =  
Page type ≠

**Date of harvest**  
From +  
To +

**URL**  
Operator: Contain +  
URL

**Sentiment** =  
Sentiment ≠

**SETTINGS OF LIMITS**

**Stop words**  
a, aby, aj, ale, ani, asi, atd, atp,...

**Number of entries**  
Number of entries: 1000  
 Random records?  
Random seed: 255006

**QUERY**  
topics:"knihovna a muzeum" OR topics:"divadlo"NOT webType:"eshop"date:[2021-08-01T00:00:00Z TO 2022-05-20T00:00:00Z]url:/\*nkp.cz.\*/sentiment:[0 TO 1]

Logical operators: **AND** **OR** **NOT** ( )

**HARVESTS**

<b>Se</b> <input checked="" type="checkbox"/> Harvest's name: Serials-2021-02-1M-2M_OneShot-crawler00 Size: 4.7 TB Number of WARC's: 1,221,667 Date of start: 2/26/2021	<b>Se</b> <input checked="" type="checkbox"/> Harvest's name: Serials-2021-01-1M-4M_OneShot-crawler00 Size: 991 MB Number of WARC's: 3,920,197 Date of start: 1/22/2021	<b>ex</b> <input type="checkbox"/> Harvest's name: example Size: 322 GB Number of WARC's: 4 Date of start: 6/23/2020
<b>Te</b> <input type="checkbox"/> Harvest's name: Test-2019-06-30-cc-naki-crawler00 Size: 260 GB Number of WARC's: 3,756 Date of start: 7/1/2019	<b>Vo</b> <input type="checkbox"/> Harvest's name: Volby-2013-01 Size: 296 GB Number of WARC's: 686 Date of start: 11/12/2013	<b>Vo</b> <input type="checkbox"/> Harvest's name: Volby-2013-00 Size: 676 GB Number of WARC's: 4,322 Date of start: 10/9/2013
<b>Te</b> <input type="checkbox"/> Harvest's name: Tests-2021-03-T-UDHPSH_QAtestII Size: 30 GB Number of WARC's: 31 Date of start: 3/29/2021	<b>Rq</b> <input type="checkbox"/> Harvest's name: Requests-2019-12-25_special Size: 2.7 GB Number of WARC's: 3 Date of start: 12/25/2019	

**Continue**



# Query, Harvests, Analytic Queries

**NK Webarchiv** [New query](#) [My queries](#) [User](#)

> **QUERY**

topics:"knihovna a muzeum" OR topics:"divadlo"NOT webType:"eshop"date:[2020-08-01T00:00:00Z TO 2022-05-20T00:00:00Z]url:/\*nkp.cz\*/sentiment:[0 TO 1]








Logical operators:    ( )

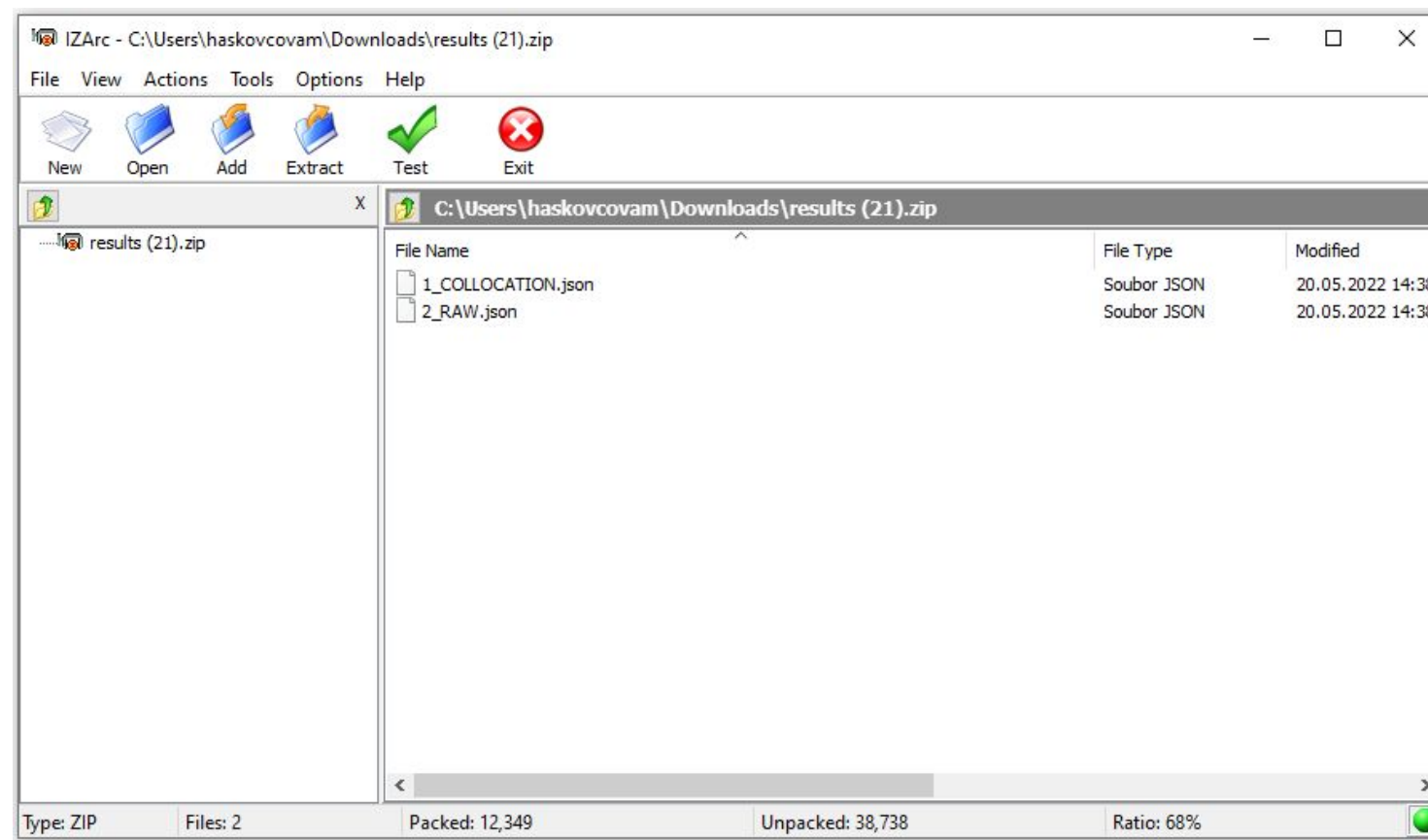
**HARVESTS**  Se  Se

**ANALYTIC QUERIES**

Query type Colocation	Insert text for search	<input type="button" value="+"/>	List of words: čtenář	<input type="button" value="Delete"/>
<input checked="" type="checkbox"/> Add context	Number of words in context: 2			
Query type Raw	Insert text for search	<input type="button" value="+"/>	List of words: knihovna	<input type="button" value="Delete"/>
	Number of entries: 4			

## My queries

QUERY	CREATED	STATE	
topics:"knihovna a muzeum" OR topics:"divadlo"NOT webType...	5/20/2022, 4:37:49 PM	Finished	 <a href="#">Download</a>
topics:"knihovna a muzeum" OR topics:"divadlo"NOT webType...	5/20/2022, 4:27:31 PM	Finished	 <a href="#">Download</a>
topics:"knihovna a muzeum" OR topics:"divadlo"NOT webType...	5/20/2022, 3:53:45 PM	Finished	 <a href="#">Download</a>
topics:"knihovna a muzeum" OR topics:"divadlo"NOT webType...	5/20/2022, 3:33:32 PM	Finished	 <a href="#">Download</a>
topics:"knihovna a muzeum" OR topics:"divadlo" AND NOT we...	5/20/2022, 3:14:12 PM	Finished	 <a href="#">Download</a>
topics:"knihovna a muzeum" OR topics:"divadlo"NOT webType...	5/20/2022, 3:07:35 PM	Finished	 <a href="#">Download</a>
topics:"knihovna a muzeum" OR topics:"divadlo" NOT webTyp...	5/20/2022, 2:57:59 PM	Finished	 <a href="#">Download</a>



## Raw

```
2_RAW (1).json
1 [ {
2   "id" : "cz,skipcr,bulletin,bf996313-ff6b-4319-b9f0-6dd901b6ab9a",
3   "url" : "bulletin.skipcr.cz/bulletin/bull04_108.htm",
4   "title" : "Bulletin SKIP, č. 1, 2004",
5   "language" : "cs",
6   "plainText" : "1. 3. 2004 vyhlašuje SKIP při zahájení BMI dlouhodu",
7   "headlines" : [ "BULLETIN", "SKIP", "1", "2004", "www.mojekniha.cz",
8   "links" : [ "https://bulletin.skipcr.cz/bulletin/Bulletin.htm", "p",
9   "topics" : [ "knihovna a muzeum", "literatura" ],
10  "sentiment" : 0.7467844144718006
11 }, {
12  "id" : "cz,blogarchiv,knihovnaskutec,4eadca7e-17d0-4a01-ac40-3f492",
13  "url" : "knihovnaskutec.blogarchiv.cz/890247-nad-klementinem-vycha",
14  "title" : "Nad Klementinem vychází slunce » Čtete knihy – jsou zd",
15  "language" : "cs",
16  "plainText" : "ředilel NK Böhm včera představil veřejnosti plány r",
17  "headlines" : [ "Leave a Reply", "Nejnovější příspěvky", "Nejnově",
18  "links" : [ "http://knihovnaskutec.blogarchiv.cz/", "http://knihovi",
19  "topics" : [ "knihovna a muzeum", "literatura" ],
20  "sentiment" : 0.3912378271722256
21 }, {
22  "id" : "cz,nkp,full,9e9e9358-cbd3-46d1-a0a1-a22ecae91317",
23  "url" : "full.nkp.cz/nkkcr/registrik2/legislativa.htm",
24  "title" : "AUTORSKÁ PRÁVA",
25  "language" : "cs",
26  "plainText" : "Služby knihoven v elektronickém prostředí a autorsk",
27  "headlines" : null,
28  "links" : [ "http://full.nkp.cz/nkkcr/search.html", "http://full.nk",
29  "topics" : [ "zločin, zákon a spravedlnost" ],
30  "sentiment" : 0.43037679125394024
31 }, {
32  "id" : "cz,konzervativnilisty,,d0584e83-9f0f-4c90-ad43-eb3f639b5ee",
33  "url" : "konzervativnilisty.cz/index.php/tema-mesice/starsi-temata",
34  "title" : "Ne Bernie, policie ani hasiči, to není socialismus",
35  "language" : "cs",
36  "plainText" : "Levičáci, při své neochabující propagandistické vá",
37  "headlines" : [ "Sledujte:", "Obsah" ],
38  "links" : [ "http://konzervativnilisty.cz/", "http://konzervativn",
39  "topics" : null,
40  "sentiment" : -0.03363580451224124
41 } ]
```

## Collocation

```
1_COLLOCATION (1).json
1 {
2   "apatykar.info/kratke-zpravy-19598" : {
3     "našich" : [ "Vážíme si našich <em>čtenářů</em>, kteří" ],
4     "preferují" : [ "<em>čtenářů</em>, kteří preferují kvalitní informace." ]
5   },
6   "cbf.cz/online-prenosy/play_352608.html" : {
7     "idnes.cz" : [ "milionů <em>čtenářů</em>. idNES.cz nabízí obrovskou" ],
8     "milionů" : [ "než 6 milionů <em>čtenářů</em>. idNES.cz" ],
9     "přináší" : [ "republice. <em>Čtenářům</em> přináší aktuální a" ],
10    "republice" : [ "v České republice. <em>Čtenářům</em> přináší" ]
11  },
12  "cbf.cz/souteze/detail_6811_soutez_3314.html" : {
13    "idnes.cz" : [ "milionů <em>čtenářů</em>. idNES.cz nabízí obrovskou" ],
14    "milionů" : [ "než 6 milionů <em>čtenářů</em>. idNES.cz" ],
15    "přináší" : [ "republice. <em>Čtenářům</em> přináší aktuální a" ],
16    "republice" : [ "v České republice. <em>Čtenářům</em> přináší" ]
17  },
18  "ceskaskola.cz/2020/06/maturantum-z-orlove-udelila-hygiena.html" : {
19    "diskusní" : [ "svým <em>čtenářům</em> diskusní prostor k" ],
20    "poskytuje" : [ "škola poskytuje svým <em>čtenářům</em> diskusní" ]
21  },
22  "ceskaskola.cz/search/label/unicef" : {
23    "diskusní" : [ "svým <em>čtenářům</em> diskusní prostor k" ],
24    "poskytuje" : [ "škola poskytuje svým <em>čtenářům</em> diskusní" ]
25  },
26  "hoax.cz/hoax/point-focus-llc-is-now-expanding/diskuse" : {
27    "návštěvníků" : [ "příspěvky návštěvníků a <em>čtenářů</em> serveru" ],
28    "serveru" : [ "a <em>čtenářů</em> serveru HOAX.cz a" ]
29  },
30  "klubknihomolu.cz/date/2021" : {
31    "dodala" : [ "a dodala svým <em>čtenářům</em> možnost" ],
32    "historií" : [ "<em>čtenáře</em> s historií včelařství a" ],
33    "možnost" : [ "svým <em>čtenářům</em> možnost přijít na" ],
34    "seznámí" : [ "pět komor seznámí <em>čtenáře</em> s" ]
35  },
}
```

datasets – new perspective at archive data and new possibilities for making it available to researchers



*Thank you for your attention*

*Tomáš Foltýn*

[tomas.foltyn@nkp.cz](mailto:tomas.foltyn@nkp.cz)

*Marie Haškovcová*

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

W

W W

W W W