



Development of the Centralized Interface for the Web Content and Social Networks Data Mining

Springboard to open Czech web archive

Tomáš Foltýn – Marie Haškovcová

Webarchiv

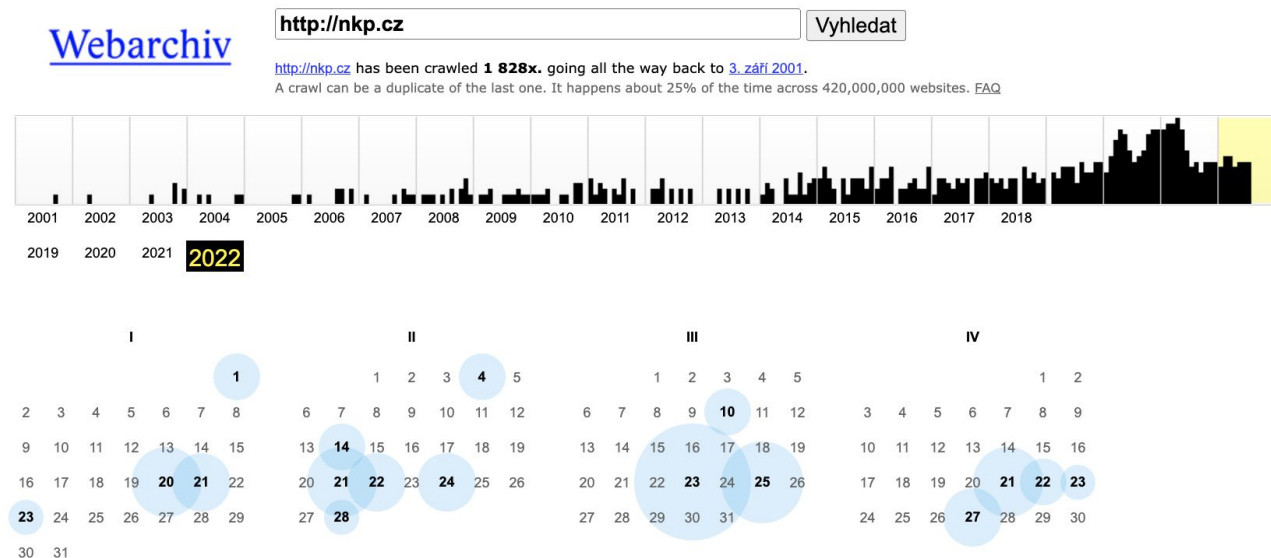
Czech web archive of the National Library of the Czech Republic

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

2001 – first archived website

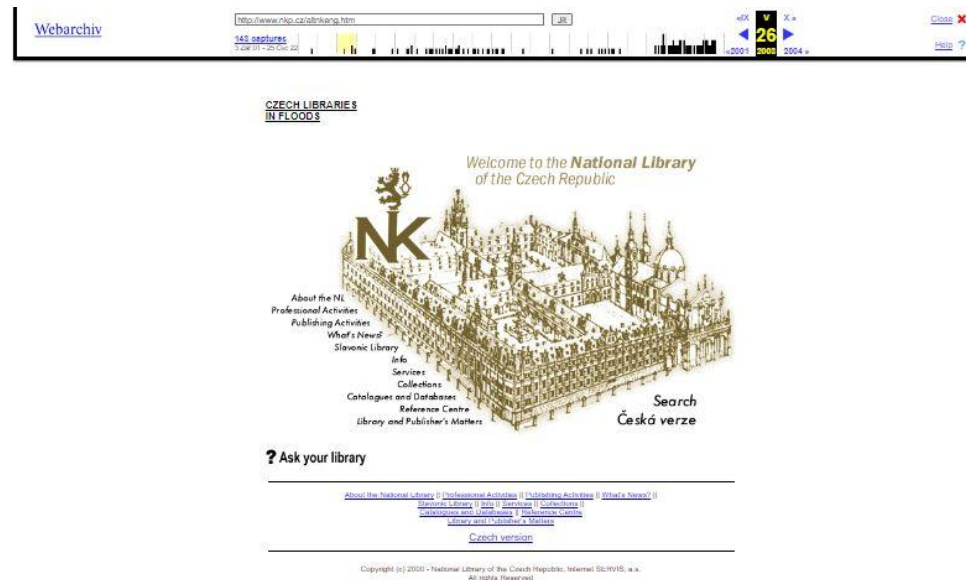
2005 – regular harvesting of content

2022 – 440 TB of data

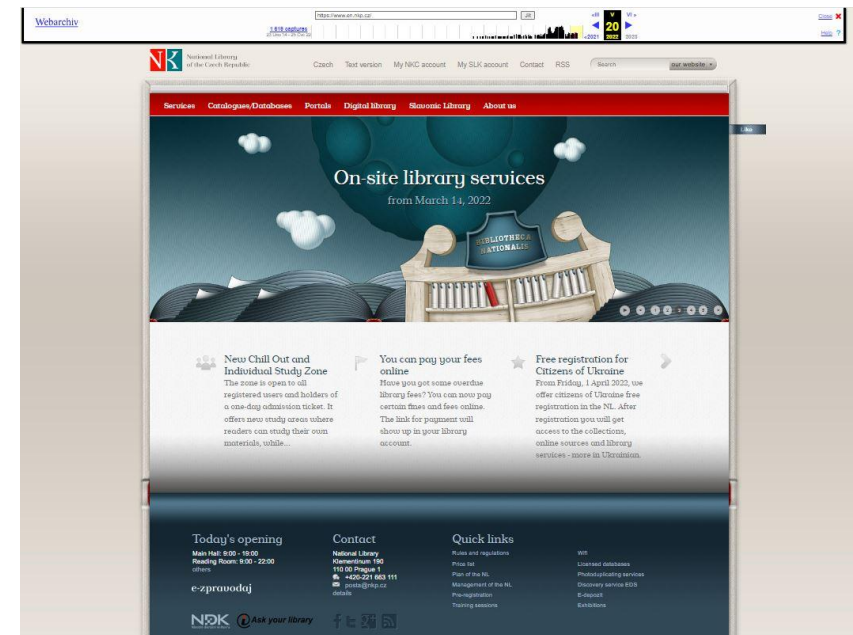


<https://www.webarchiv.cz/en/>

archive copy of National Library of the Czech Republic website

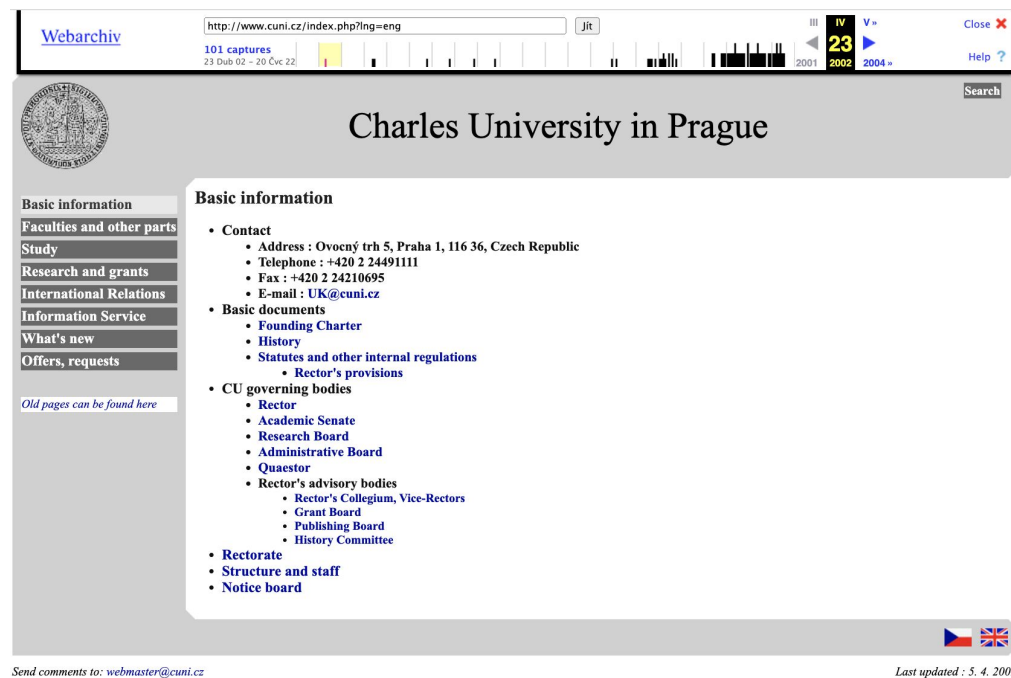


2003

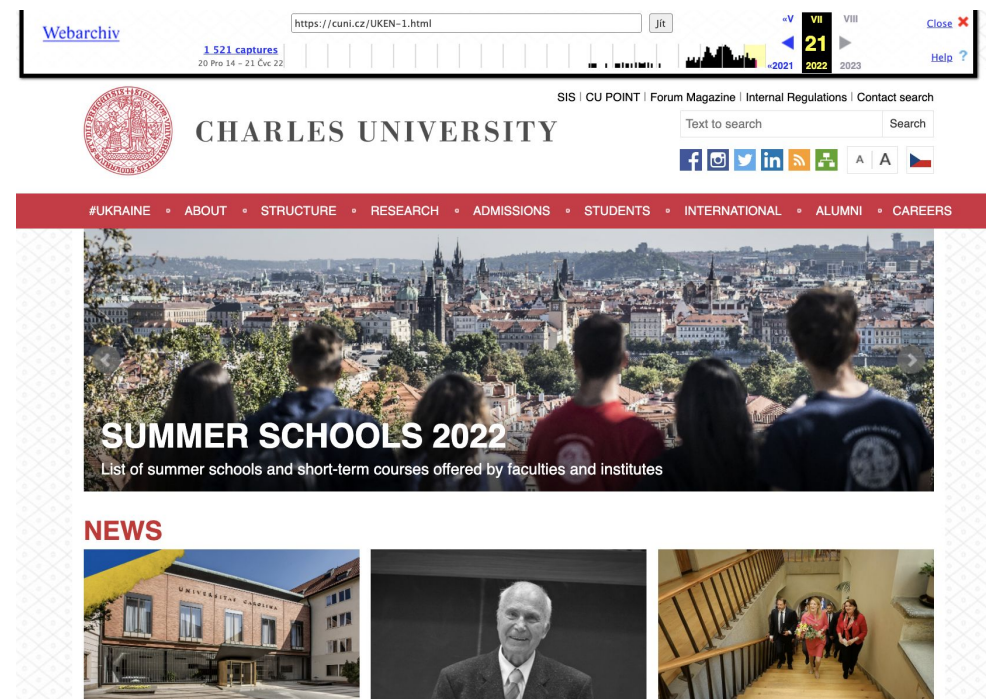


2022

archive copy of Charles University in Prague website



2001



2022

Development of centralized interface for extracting big data from web archives

- **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)

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

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

Data a metadata

Webarchiv - digital library, preserves websites for future generations

Czech web resources (territory, language, authorship or topic/content)
not only within the Czech domain

crawler: Heritrix 3.4, Webrecorder

access: OpenWayback 3.0

storing: WARC, container file standard for storing
web content in its original context

managing resources, websites, harvests: Seeder

<https://github.com/webarchivcz/>

25–40 TB per year

metadata:

- descriptive / bibliographic metadata
- technical and administrative metadata

```
1 {
2   "_id": "5dcd98a695bcf5a1a194f0be",
3   "recType": "harvest",
4   "author": "NKCR",
5   "date": "2019-11-14T19:10:46.983Z",
6   "standard": "Grainery 0.4",
7   "harvest": {
8     "harvestPrefix": {
9       "harvestNameStand": "V6M_2017-10-05",
10      "harvestFromWarinfo": "V6M_2017-10-05",
11      "harvestNameFNtrunc": "V6M_2017-10-05",
12      "harvestDirName": "V6M_2017-10-05",
13      "harvestType": "?výběrová",
14      "harvestSuffix": ["V6M", "2017-10-05"]
15    },
16    "date": "2017-10-05T11:26:00.000Z",
17    "harvestID": "105f23c9-b037-4c1d-901c-dcf272877d9f",
18    "size": 618029,
19    "warcsNumber": 12092
20  },
21  "harvestCrawl": {
22    "logs": true,
23    "path": "logs/crawl",
24    "fileName": ["crawler00.tar.gz", "crawler01.tar.gz", "crawler03.tar.gz"]
25  },
26  "paths": {
27    "cdxsID": ["105f23c9-b037-4c1d-901c-dcf272877d9f"],
28    "warcsID": ["105f23c9-b037-4c1d-901c-dcf272877d9f"],
29    "warcsFileNames": ["V6M_2017-10-05-crawler00.webarchiv.cz-warcs.gz"]
30  },
31  "revision": {
32    "dateOfValidation": "2019-12-04T19:10:46.983Z",
33    "statusOfValidation": "NA",
34    "nextLastDateOfValidation": "2021-12-03T19:10:46.983Z",
35    "hashOrig": "NA",
36    "hashLast": "NA",
37    "commentaries": { "exists": false, "text": "NA" }
38  }
39 }
```

metadata record in JSON

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

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 5373 / [Agriculture](#) 244 / [Anthropology](#) 248 / [Art and architecture](#) 384 / [Beletry](#) 40 / [Biological sciences](#) 304 / [Business and economics](#) 415 / [Chemistry](#) 51 / [Children's literature](#) 7 / [Computer sciences](#) 211 / [Education](#) 251 / [Engineering and technology](#) 311 / [Geography and earth sciences](#) 444 / [History and auxiliary sciences](#) 332 / [Language, linguistics and literature](#) 283 / [Law](#) 155 / [Library science, generalities and references](#) 335 / [Mathematics](#) 46 / [Medicine](#) 319 / [Music](#) 164 / [Performing arts](#) 263 / [Philosophy and religion](#) 260 / [Physical education and recreation](#) 230 / [Physical sciences](#) 140 / [Political science](#) 397 / [Psychology](#) 89 / [Sociology](#) 341

Education /

Vše 251 / [Adult education](#) 6 / [Curricula](#) 4 / [Education](#) 108 / [Elementary education](#) 18 / [Higher education](#) 46 / [Professional education](#) 10 / [Public policy issues in education](#) 13 / [Schools and their activities](#) 20 / [Secondary education](#) 12 / [Special education](#) 14

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



[19. výroční konference České asociace pedagogického výzkumu](#)



[Adam.cz : zpravodajský a informační servis sdružení dětí a mládeže](#)



[Akademické centrum osobnostního rozvoje](#)



[Akademie múzických umění v Praze](#)



[Alternativní školy](#)



[Apple ve školství](#)



[Archiv vítězných prací SOČ](#)



[Asociace institucí vzdělávání dospělých ČR : AIVD](#)

Topic collections

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

COVID-19

Keywords of harvest:

[krizový management](#), [infekční nemoci](#), [epidemie](#), [nemoci](#), [krizové situace](#), [globální problémy](#), [lékařská péče](#), [přenos infekčních nemocí](#)



SZU.CZ

<http://szu.cz/tema/krizove-situace/2019-ncov-novy-koronavirus-wu-chan> [\[current\]](#)

www.mzcr.cz

http://www.mzcr.cz/dokumenty/informace-pro-obcany-v-souvislosti-s-aktualnim-vyskytem-cinskeho-koronaviru-2019_18415_1.html [\[current\]](#)

www.vlada.cz

<https://www.vlada.cz/cz/media-centrum/aktualne/aktualni-informace-ke-koronaviru-2019-ncov-179250/> [\[current\]](#)

www.vlada.cz

<https://www.vlada.cz/cz/epidemie-koronaviru/> [\[current\]](#)

www.infekce.cz

<https://www.infekce.cz/> [\[current\]](#)

koronavirus.mzcr.cz

<https://koronavirus.mzcr.cz/> [\[current\]](#)

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 - professional and health portals, reactions in various types of media, volunteer activities and civic initiatives, resources exploring the economic, legal and social impacts of government measures, conspiracy theories or artistic reflections. Czech web resources are also part of the large international collection Novel Coronavirus (COVID-19), which is being prepared by IIPC (<https://archive-it.org/collections/13529>).

Metadata

Cataloging and bibliographic metadata

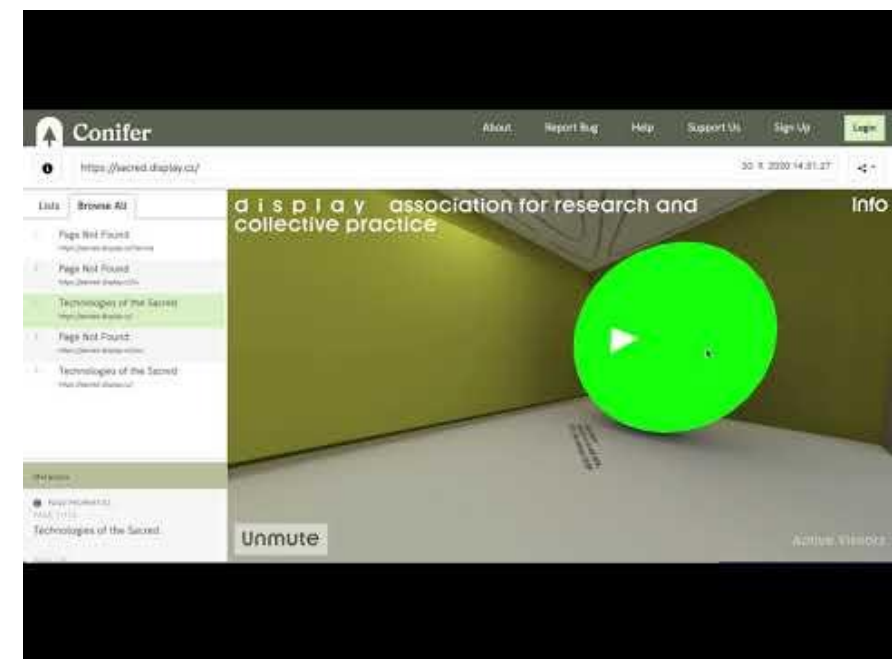
- library system **Aleph**
- format for Bibliographic Data **MARC 21**
- **RDA**, since 2015

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

Challenges

- 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)



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

Centralized interface for extracting big data from web archives - WACloud

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

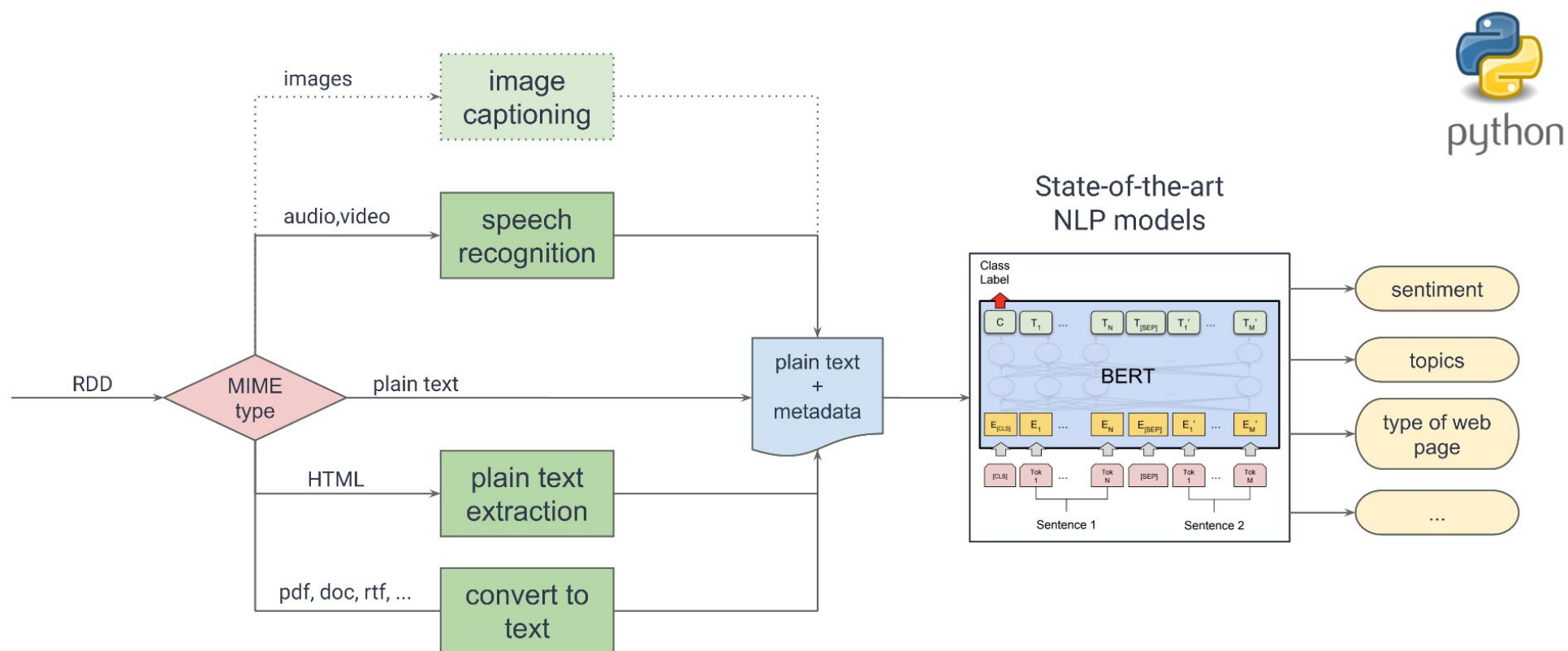
- analysis text document topics and their automatic detection, analysis of audiofiles, approaches based on deep neural networks for document classification, automatic analytical tools, automatic assignment of metadata to individual documents
- **intermediary format** = format of records during web processing 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

Infrastructure, technologies, AI

- **infrastructure:** HADOOP cluster and Hbase solutions
- **used technologies:** WARC – PySpark – Python – Scikit-Learn – Apache Hbase – Hadoop – Apache Spark – TensorFlow – PyTorch


AI Algorithms Pipeline



dataset, media types, processes, plain text + metadata, language processing model, deep neural network, trained to predict topic, sentiment, page type

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](#) [FAQ](#) [Czech](#)

[Webarchiv](#)

Sign in

Centralized interface for extracting big data
from web archives

Webarchiv

New query

Favorite

My queries

FAQ

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?

QUERY

Logical operators:

AND

OR

NOT

(

)

HARVESTS

Se

Harvest's name

Serials-2021-02-1M-2M_OneShot-crawler00

Size

4.7 TB

Number of WARC's

4,735,803

Date of start

2/26/2021

ex

Harvest's name

example

Size

991 MB

Number of WARC's

0

Date of start

6/23/2020

To

Harvest's name

Topics-2020-10-01-VolbyKrajeSenat2020

Size

322 GB

Number of WARC's

305

Date of start

6/8/2021

To

Harvest's name

Topics_2021-07_T-CUNI_MGHMP_Klima_Tornado2_Olympiada3

Size

260 GB

Number of WARC's

236

Date of start

7/30/2021

Se

Harvest's name

Serials_2021-08_1M-2M_OneShot

Size

296 GB

Number of WARC's

280

Date of start

8/30/2021

Se

Harvest's name

Serials_2021-09_1M-6M_OneShot

Size

676 GB

Number of WARC's

628

Date of start

9/29/2021

Te

Harvest's name

Tests-2021-03-T-UDHPSH_QAtestII

Size

30 GB

Number of WARC's

31

Date of start

3/29/2021

Rq

Harvest's name

Requests-2019-12-25_special

Size

2.7 GB

Number of WARC's

3

Date of start

12/25/2019

Continue

Analytic Queries - Collocation, Raw, Frequency, Network

Webarchiv

New queryFavoriteMy queriesFAQUser

> QUERY

topics:"knihovna a muzeum" OR topics:"literatura" NOT webType:"eshop" AND date:[2021-01-01T00:00:00Z TO 2022-09-03T00:00:00Z] AND url:/.*nkp.cz.*/ AND sentiment:[0 TO 1]

Logical operators:

AND

OR

NOT

(

)

HARVESTS

Edit

ANALYTIC QUERIES

+ Add new query

Query type
Colocation

▼

Insert text for search

+

List of words:

žtenářⓧ

Delete

☒ Add context

Number of words in context:
2

Query type
Raw

▼

Insert text for search

+

List of words:

knihovnaⓧ

Delete

Number of entries
3

Query type
Frequency

▼

Insert text for search

+

List of words:

knihaⓧ

Delete

Number of entries
5

Query type
Network

▼

Source nodes

+

Source nodes

nkp.czⓧ

Delete

☒ Group by domains

Target nodes

+

Target nodes

nkp.czⓧ

☒ Group by domains

STATE OF PROCESS

☒ Indexing
100%

☒ Statistical queries
100%

☒ Finished

Download

Add to favorites

QUERY	CREATED	STATE	
★ topics:"knihovna a muzeum" OR topics:"literatura" NOT web...	9/3/2022, 7:32:15 PM	Finished	<div><div>👁 Show detail</div><div>★ Add to favorites</div><div>🔄 Repeat query</div><div>📄 Download</div><div>...</div></div>
★ topics:"knihovna a muzeum" OR topics:"literatura" NOT web...	9/3/2022, 7:06:25 PM	Finished	
★ topics:"knihovna a muzeum" OR topics:"divadlo" AND NOT we...	9/3/2022, 6:48:23 PM	Finished	
★ topics:"knihovna a muzeum" NOT webType:"eshop" AND date:[...	9/3/2022, 6:44:12 PM	Finished	...

<div> <div> <div><</div> <div>></div> <div>results (4)</div> </div> <div> <div>☰</div> <div>☒</div> </div> <div> <div>☰☰☰</div> <div>☑</div> </div> <div> <div>📁</div> <div>🏷️</div> <div>⋮</div> </div> <div>🔍</div> </div>				
<div> <div></div> <div>Upgradovat...</div> <div>Další informace...</div> </div>				
Název	☒	Datum změny	Velikost	Druh
📄 1_COLLOCATION.json	☁️	Včera 17:33	6 kB	JSON file
📄 2_RAW.json	☁️	Včera 17:33	50 kB	JSON file
📄 3_FREQUENCY.json	☁️	Včera 17:33	57 kB	JSON file
📄 4_NETWORK.json	☁️	Včera 17:33	14 kB	JSON file

Raw

```
2_RAW.json
16  , {
17    "id" : "af121c3d-03cf-4b61-8b8f-39269a5a2825",
18    "url" : "text.nkp.cz/sluzby/dulezite-odkazy/knihovni-rad-a-dalsi-pravidla/kr-pr",
19    "urlDomain" : "text.nkp.cz",
20    "urlDomainTopLevel" : "nkp.cz",
21    "title" : "Přehled služeb poskytovaných NK – Národní knihovna České republiky",
22    "language" : "cs",
23    "plainText" : "Půjčování knihovních jednotek z obsluhovaných příručních knihoven
    • pobyt na území České republiky\nRegistrovanému uživateli staršímu 18 let s plat
    • vedené v evidenci podle knihovního zákona Knihovně v zahraničí\nMeziknihovní vý
    • elektronicky;\nKonzultace na objednávku o katalozích, databázích a sbírkách kni
24    "year" : 2021,
25    "headlines" : [ "Přehled služeb poskytovaných NK", "Přehled služeb poskytovaných
26    "links" : [ "https://aleph.nkp.cz/F/?func=bor-info&local_base=NKC", "https://al
    • "http://text.nkp.cz/sluzby-behem-revitalizace", "http://text.nkp.cz/o-knihovne"
    • knihovni-rad-a-dalsi-pravidla/kr-1", "http://text.nkp.cz/sluzby/vypsluzby-preze
    • reserse", "http://text.nkp.cz/sluzby/reserse", "http://text.nkp.cz/sluzby/repr"
    • text.nkp.cz/sluzby/dulezite-odkazy/knihovni-rad-a-dalsi-pravidla/sluzby/sluzby-
    • a-dalsi-pravidla/sluzby/dulezite-odkazy/cenik", "http://eds.nkp.cz", "http://te
    • zarazovani-zahranicnich-dokumentu_riv", "http://www2.nkp.cz/", "http://text.nkp
    • ", "http://www.ptejteseknihovny.cz/", "http://www.facebook.com/narodni.knihovna
    • "http://text.nkp.cz/sitemap", "http://text.nkp.cz/accessibility-info", "http://
27    "linksDomain" : [ "aleph.nkp.cz", "aleph.nkp.cz", "www.nkp.cz", "text.nkp.cz",
    • "text.nkp.cz", "text.nkp.cz", "text.nkp.cz", "text.nkp.cz", "text.nkp.cz", "tex
    • "www.602.cz", "www.cro.cz", "text.nkp.cz", "text.nkp.cz", "text.nkp.cz" ],
28    "linksDomainTopLevel" : [ "nkp.cz", "nkp.cz", "nkp.cz", "nkp.cz", "nkp.cz", "nk
    • "nkp.cz", "nkp.cz", "nkp.cz", "nkp.cz", "nkp.cz", "nkp.cz", "nkp.cz", "nkp.cz",
29    "topics" : [ "knihovna a muzeum" ],
30    "sentiment" : 0.5810921904179829
31  }, {
32    "id" : "f0219aaf-ffe6-45c8-aafc-1c8c93aa21bd",
33    "url" : "text.nkp.cz/sluzby/dulezite-odkazy/prehled-sluzeb",
34    "urlDomain" : "text.nkp.cz",
```

Frequency

```
3_FREQUENCY.json
1  {
2  "nkp.cz/sluzby/dulezite-odkazy/jak-to-tady-chodi" : {
3    "nk" : 9,
4    "cr" : 7,
5    "katalogu" : 7,
6    "byt" : 5,
7    "jednotlivych" : 5
8  },
9  "text.nkp.cz/o-knihovne/zakladni-informace/kontakty/pr
10    "2018" : 6,
11    "tiskova" : 6,
12    "zprava" : 6,
13    "3" : 4,
14    "4" : 3
15  },
16  "text.nkp.cz/o-knihovne/zakladni-informace/kontakty/pr
17    "2016" : 8,
18    "tiskova" : 8,
19    "zprava" : 8,
20    "10" : 7,
21    "cr" : 5
22  },
23  "skip.nkp.cz/bulletin/bull04_110.htm" : {
24    "ungar" : 19,
25    "knihovny" : 18,
26    "roce" : 12,
27    "byly" : 9,
28    "ceske" : 8
29  },
```


Network

```
4_NETWORK.json
1 [ {
2   "year" : 2021,
3   "nodes" : [ {
4     "name" : "authority.nkp.cz",
5     "links" : [ {
6       "name" : "aleph.nkp.cz",
7       "count" : 1
8     }, {
9       "name" : "authority.nkp.cz",
10      "count" : 4
11     }, {
12      "name" : "www.nkp.cz",
13      "count" : 4
14     } ]
15   }, {
16     "name" : "dnnt.nkp.cz",
17     "links" : [ {
18       "name" : "aleph.nkp.cz",
19       "count" : 1
20     }, {
21       "name" : "dnnt.nkp.cz",
22       "count" : 2
23     } ]
24   }, {
25     "name" : "edeposit.nkp.cz",
26     "links" : [ {
27       "name" : "edeposit.nkp.cz",
28       "count" : 2
29     } ]
30   } ]
}
```

Collocation

```
1_COLLOCATION.json
1 {
2   "nkp.cz/aktuality/novinky-titulni-strana/otevreni" : {
3     "budou" : [ "<em>čtenáři</em> si budou moci vyzvednout", "možnost vyzvednutí" ],
4     "hale" : [ "pokud v Hale služeb již" ],
5     "kontě" : [ "evidovaných na kontě <em>čtenáře</em> je" ],
6     "maximálně" : [ "<em>čtenáře</em> je maximálně 20 zadávání" ],
7     "musí" : [ "počet <em>čtenářů</em>, musí další <em>čtenáři</em>", "<em>čtenáři</em>" ],
8     "podpisu" : [ "<em>čtenář</em> k podpisu potvrzení o" ],
9     "povinni" : [ "<em>Čtenáři</em> jsou povinni: vstupovat a" ],
10    "počet" : [ "maximální povolený počet <em>čtenářů</em>, musí" ],
11    "služeb" : [ "z meziknihovních služeb <em>čtenáři</em> si" ],
12    "upozornění" : [ "budou <em>čtenáři</em> upozornění e-mailem souběžně" ],
13    "uvítáme" : [ "uvítáme, když si <em>čtenář</em> k" ],
14    "vyčkat" : [ "další <em>čtenáři</em> vyčkat a řídit" ],
15    "čr" : [ "pracovníků NK ČR <em>Čtenáři</em> jsou" ],
16    "čtyři" : [ "služeb pouze čtyři <em>čtenáři</em>; pokud" ]
17  },
18  "nkp.cz/sluzby/chci-sluzbu/stat-se-ctenarem/copy_of_chcisluzbu-ctenar" : {
19    "doloží" : [ "(pokud <em>čtenář</em> doloží trvalý nebo" ],
20    "fondu" : [ "knihovního fondu (pokud <em>čtenář</em> doloží" ],
21    "přístup" : [ "všichni <em>čtenáři</em> přístup ke službám" ],
22    "všichni" : [ "budovu mají všichni <em>čtenáři</em> přístup" ]
23  },
24  "text.nkp.cz/slovanska-knihovna/sluzby-slovanske-knihovny/vypujcni-sluzby/regis" : {
25    "aleph" : [ "v systému Aleph. <em>Čtenáři</em> mají", "v systému Aleph. <em>Čtenáři</em>" ],
26    "mají" : [ "Aleph. <em>Čtenáři</em> mají možnost zadávat", "Aleph. <em>Čtenáři</em>" ],
27    "mohou" : [ "<em>Čtenáři</em> si mohou knihy prodlužovat", "<em>Čtenáři</em>" ],
28    "nežádá" : [ "knihu nežádá další <em>čtenář</em>. <em>Čtenáři</em>", "knihu" ],
29    "registraci" : [ "bezplatně. Za registraci <em>čtenářů</em>, reprografické", "registraci" ],
30    "reprografické" : [ "registraci <em>čtenářů</em>, reprografické a kopírovací", "reprografické" ],
31    "systém" : [ "automatizovaný výpůjční systém. <em>Čtenáři</em> a", "automatizovaný" ],
32    "výpůjčky" : [ "<em>Čtenáři</em> a výpůjčky (prozatím pouze", "<em>Čtenáři</em>" ]
25  }
}
```

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

Marie Haškovcová

marie.haskovcova@nkp.cz

W

W W

W W W