

Webarchiv

Český webový archiv, jeho historie, fungování a collection policy

Marie Haškovcová

Webarchiv a archivace webu

Webarchiv

digitální knihovna českých webových zdrojů, uchovává významnou část národního kulturní dědictví

výběr webového obsahu, jeho sběr, dlouhodobé uchování sbírek v archivním formátu a zpřístupnění

404 Not Found

nginx/1.10.3

Historie

2000 – Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet - projekt NK ČR, MZK a MU

2001 – první kopie

2005 – pravidelné sklizení

2007 – IIPC



2020 – více než 440 TB dat

Univerzita Karlova

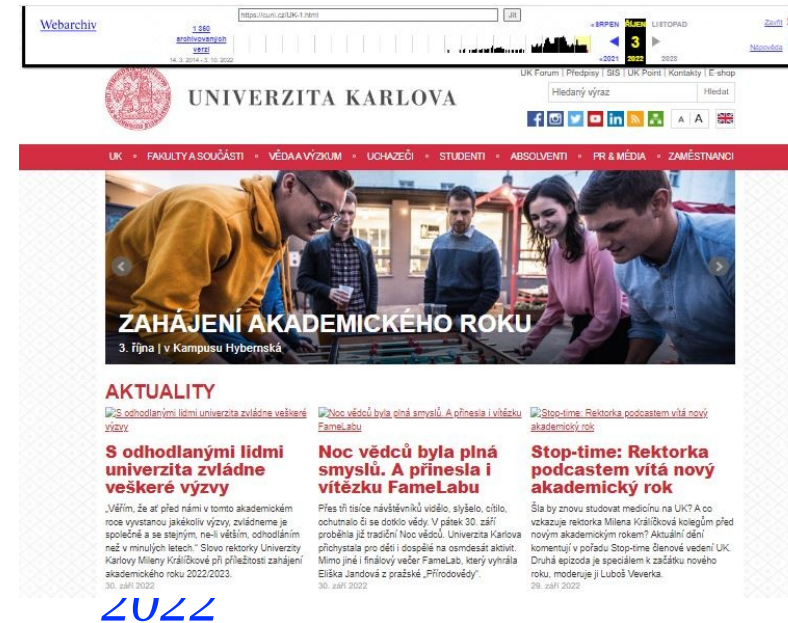


2001

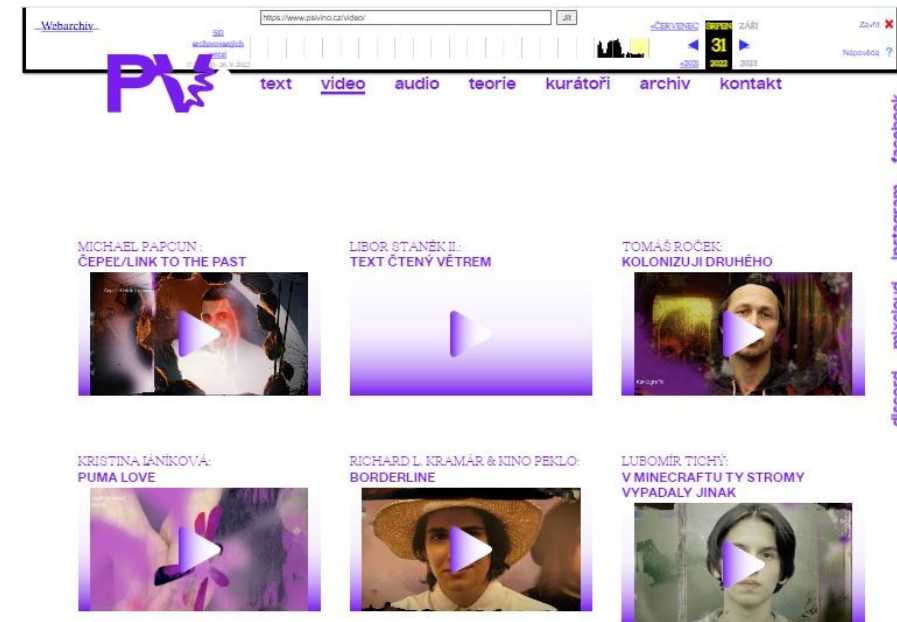
Psí víno



2010



2022



Legislativa

- Autorský zákon 121/2000 Sb., Knihovní licence § 37 - rozmnoženina díla, archivní a konzervační potřeby
- volný přístup – smlouva nebo Creative Commons
0,4 % archivu
- trojnovela – právní úprava web-harvestingu
- implementace směrnice o autorském právu na jednotném digitálním trhu do české legislativy

Jak archivujeme? Collection policy

- Celoplošné sklizně
.cz, CZ.NIC, cca 1,4 milionů domén
- Výběrové sklizně
kurátorský přístup
- Tematické kolekce
událost, téma

Nechte se [Webarchivovat!](#)

Pokud hledáte certifikát či jiné naše grafické materiály pro umístění na vašem webu pokračujte [zde](#)

[Přidejte web](#) / [Creative Commons](#) / [Výběr zdrojů](#) / [Pomozte nám uchovat český web](#) / [FAQ](#)

Přidejte web


URL

Mohu jednat za uvedené zdroje Zdroj s licencí Creative Commons

Jméno

Kontaktní e-mail

Poznámka

Nejsem robot  reCAPTCHA
Ochrana soukromí · Smluvní podmínky

Přidat web

Strategie budování sbírky Webarchivu

<https://webarchiv.cz/static/www/download/collection-policy.pdf>
pravidla pro zpřístupňování obsahu, technická nastavení crawlerů

Výběrové sklizně

- online přístup (smlouva, Creative Commons), 5300 zdrojů
- pravidelné a dlouhodobé sklizení
- katalogizace

[Webarchiv](#) podle oborů

Přehled nasmlouvaných webů dle oborového třídění:

Vše 5380 / [Antropologie, etnografie](#) 248 / [Beletrie](#) 40 / [Biologické vědy](#) 304 / [Chemie, Krystalografie, Mineralogické vědy](#) 51 / [Divadlo, film, tanec](#) 264 / [Ekonomické vědy, obchod](#) 415 / [Filozofie a náboženství](#) 260 / [Fyzika a příbuzné vědy](#) 140 / [Geografie, Geologie, Vědy o Zemi](#) 444 / [Historie a pomocné historické vědy, Biografické studie](#) 332 / [Hudba](#) 164 / [Jazyk, lingvistika, literární věda](#) 283 / [Knihovnictví, informatika, všeobecné, referenční literatura](#) 335 / [Lékařství](#) 319 / [Literatura pro děti a mládež](#) 7 / [Matematika](#) 46 / [Politické vědy \(Politologie, politika, veřejná správa, vojenství\)](#) 396 / [Právo](#) 155 / [Psychologie](#) 89 / [Sociologie](#) 342 / [Technika, technologie, inženýrství](#) 311 / [Tělesná výchova a sport, Rekreace](#) 231 / [Umění, architektura](#) 385 / [Výchova a vzdělávání](#) 253 / [Výpočetní technika](#) 213 / [Zemědělství](#) 245

Umění, architektura /

Vše 385 / [Architektura](#) 59 / [Fotografie](#) 50 / [Grafické umění, Grafika](#) 16 / [Kresba, Umělecká řemesla](#) 16 / [Malířství](#) 15 / [Sochařství, keramika, porcelán, umělecké zpracování kovů](#) 17 / [Umění](#) 79 / [Územní plánování, Urbanismus, Památková péče](#) 23 / [Výtvarné umění](#) 106

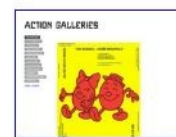
Zobrazení: vizuální, [textové](#)



[14|15 Bařův institut](#)



[abod : art brut](#)



[Action galleries](#)



[Adolf Wolfli : stvořitel universa](#)



[Akademie výtvarných umění v Praze](#)



[Alena Schulz](#)



[Alice Nikitinová](#)



[Alšova jihočeská galerie v Hluboké nad Vltavou : AIG](#)

Tematické sbírky

- aktuální události (volby, povodně)
- dlouhodobě budované sbírky (česká média)
- spolupráce s IIPC

[úvod](#) / [o Webarchivu](#) / [katalog stránek](#) / [tematické sbírky](#) / [přidat web](#)



CZ [EN](#)

Klimatická změna

Klíčová slova sklizně:

[ekologie](#), [ochrana klimatu](#), [ekologické instituce a organizace](#), [klimatologie](#), [klimatické změny](#)



[Česká republika 2030: společně - udržitelně](#)

[www.cr2030.cz](#) [[archivovaná verze](#)]

[Fakta o změně klimatu](#)

[faktaoklimatu.cz](#) [[archivovaná verze](#)]

[Klimatická koalice](#)

[klimatickakoalice.cz](#) [[archivovaná verze](#)]

[Umění pro klima](#)

[umeniproklima.cz](#) [[archivovaná verze](#)]

Změna klimatu je jednou z největších výzev současnosti. Hledání řešení na vědecké i politické úrovni doprovází intenzivní celospolečenská debata. Kolekce, která vznikla ve spolupráci s mezinárodním konsorciem webových archivů IIPC, zahrnuje weby vědeckých ústavů a projektů, státních institucí, neziskových organizací i ohlasy v médiích.

Data a metadata

České webové zdroje - bohemikální charakter (území, autorství, jazyk, obsah), nejen na české doméně

sklizení: Heritrix 3.4, Webrecorder, archiveweb page / browsertrix (sociální média, dynamický obsah)

zpřístupnění: OpenWayback 3.0

ukládání: WARC

kurátorský nástroj: Seeder, <https://github.com/webarchivcz/>

25–40 TB ročně

metadata:

- popisná / bibliografická
- technická a administrativní

Katalogizace a bibliografická metadata

- knihovní systém **Aleph**
- **MARC 21**
- **RDA**, od roku 2015

WA-KAT

<https://kat.webarchiv.cz/>

Katalogizační manuál

<https://webarchivcz.github.io/katalogizacni-manual/>

WebArchiv - Úplné zobrazení záznamu

Zvolte formát: Standardní -- Katalogizační záznam -- Stručný záznam -- MARC -- [Citace](#)

Záznam 1 z 2

```
LDR      -----nai-a22-----i-4500
FMT      SE
BAS      |a 01
BAS      |a WAR
001      web20172886912
003      CZ-PrNK
005      20170320175238.0
006      m-----o-----
007      cr-cn-
008      170320c20069999xr--x-w-o-----0--b2cze--
015      |a cnb002886912
040      |a ABA001 |b cze |e rda
0410     |a cze |a eng
043      |a e-xr-- |b e-xr-pl |2 czenas
072 7    |a 785 |x Instrumentální hudba. Symfonická hudba. Hudba pro více nástrojů |2 Konspekt |9 9
072 9    |a 784-788 |x Instruments and instrumental ensembles and their music |2 Conspectus |9 9
080      |a 785.11.071 |2 MRF
080      |a 785.11 |2 MRF
080      |a 78 |2 MRF
080      |a 78-027.22 |2 MRF
080      |a 78.082.4 |2 MRF
080      |a (0.034.2)004.738.12 |2 MRF
1102     |a Plzeňská filharmonie |7 ko2004209449 |4 pbl
24510    |a Plzeňská filharmonie
264 1    |a Plzeň : |b Plzeňská filharmonie, |c [2006]-
300      |a 1 online zdroj
310      |a Aktualizováno nepravdělně
336      |a text |b txt |2 rdacontent
337      |a počítač |b c |2 rdamedia
338      |a online zdroj |b cr |2 rdacarrier
520      |a Stránky zahrnují základní informace o Plzeňské filharmonii, o orchestru, údaje koncertech a vystoupeních.
588      |a Název z titulní obrazovky (náhled z 20.03.2017) |5 CZ-PrNK
61027    |a Plzeňská filharmonie |7 ko2004209449 |2 czenas
65007    |a symfonické orchestry |7 ph135216 |2 czenas
65007    |a orchestrální hudba |7 ph123783 |2 czenas
65007    |a hudba |7 ph114719 |2 czenas
65007    |a hudební život |7 ph135177 |2 czenas
65007    |a koncerty |7 ph121790 |2 czenas
65009    |a symphonic orchestras |2 eczenas
65009    |a orchestral music |2 eczenas
65009    |a music |2 eczenas
65009    |a musical life |2 eczenas
65009    |a concerts |2 eczenas
651 7    |a Plzeň (Česko) |7 ge130439 |2 czenas
651 9    |a Plzeň (Czechia) |2 eczenas
655 7    |a www dokumenty |7 fd186892 |2 czenas
655 9    |a www documents |2 eczenas
85640    |u http://www.plzenskafilharmonie.cz |q text/html |4 N
85640    |u http://wayback.webarchiv.cz/wayback/http://www.plzenskafilharmonie.cz |q text/html |z archivní verze stránek |4 N
929      |a Text
SYS      002886912
```

[http://aleph.nkp.cz/F/?func=direct&doc_number=002886912&local_base=Web]

[Citace](#)

© 2014 Ex Libris, NK ČR

Technická a administrativní metadata

technické údaje získané během sklizně:
datum zahájení a ukončení sklizně, typ,
autor ad.

3 typy:

- sklizeň
- kontejner
- index

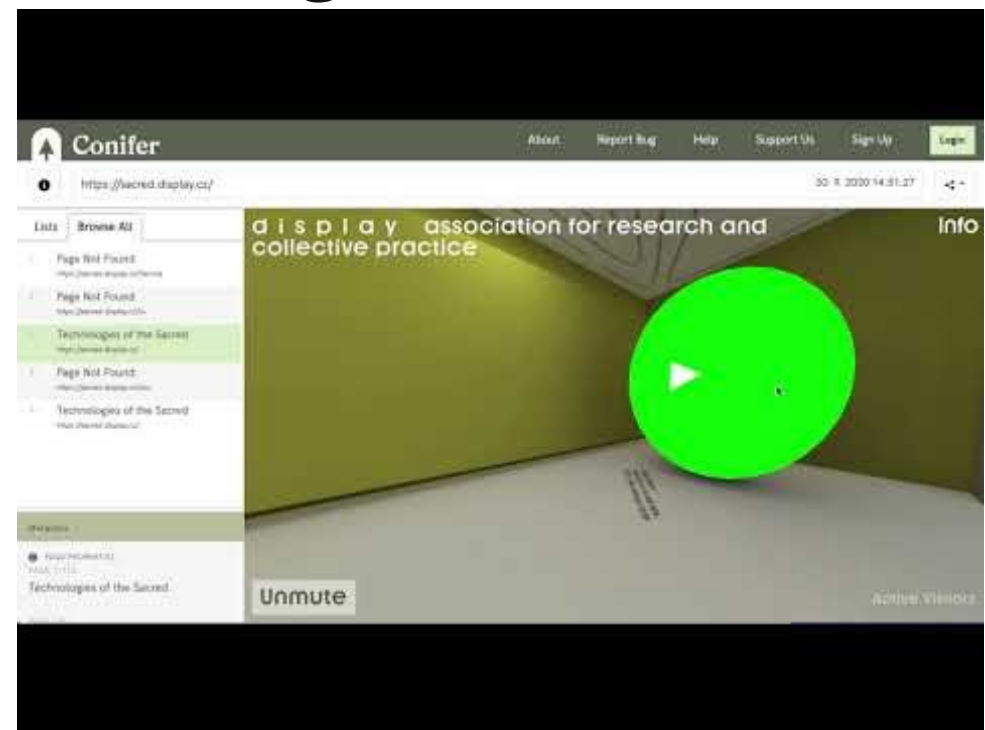
```
1 {
2   "_id": "5dcd98a695bcf5a1a194f0be",
3   "recType": "harvest",
4   "author": "NKR",
5   "date": "2019-11-14T19:10:46.983Z",
6   "standard": "Grainery 0.4",
7   "harvest": {
8     "harvestPrefix": {
9       "harvestNameStand": "V6M_2017-10-05",
10      "harvestFromWarinfo": "V6M_2017-10-05",
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-dcf272874d9f"],
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 }
```

Metodika pro tvorbu, uložení a zpřístupnění technických a administrativních metadat z webového archivu

<https://invenio.nusl.cz/record/432325?ln=cs>, <https://github.com/WebarchivCZ/grainery>

Výzvy

- collection policy – profilování sbírek
- dynamický obsah, sociální média - tw, fb, ig



- kontrola kvality, ochrana dat
- zpřístupnění dat / metadat veřejnosti, nástroje
- spolupráce s vědci a institucemi

Vývoj centralizovaného rozhraní pro vytěžování velkých dat z webových archivů

- **Národní knihovna ČR**
(data, zkušenosti s archivací webu, infrastruktura)
- **Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Katedra kybernetiky**
(strojové zpracování dat, přístupy založené na hlubokých neuronových sítích pro klasifikaci dokumentů)
- **Sociologický ústav AV ČR, v.v.i.**
(výzkumníci v oblasti sociálních věd)

MK ČR – Program na podporu aplikovaného výzkumu a experimentálního vývoje národní a kulturní identity, 2018 – 2022 (NAKI II)

Cíle a výstupy výzkumného projektu

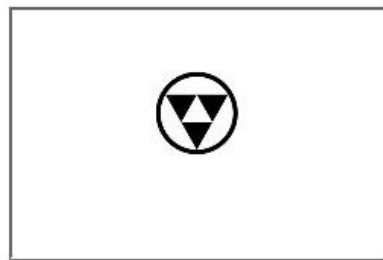
Cíle

- umožnit přístup k datům Wearchivu badatelské komunitě
- rozšíření technologické infrastruktury, která umožní další analytické zpracování velkého objemu dat
- přehodnocení právního rámce

Výstupy

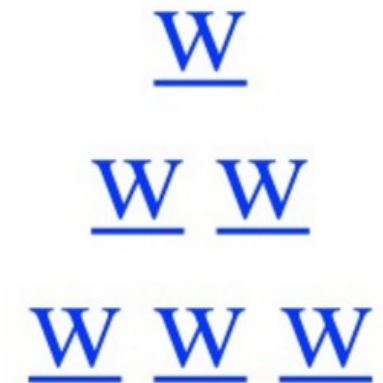
- uživatelsky přívětivé grafické rozhraní
- fasetový a fulltextový vyhledávač, který umožní výzkumníkům definovat část dat pro jejich výzkum
- exportní aplikace, která umožňuje výzkumníkům získat datové sety

Archivní kopie X Dataset



```
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   } ]
}
```

Děkuji za pozornost



www.webarchiv.cz

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

marie.haskovcova@nkp.cz