23-24 November 2021

404 Not Found Who will preserve the Internet?



Online video conference and workshop

6-YEAR JOURNEY OF SLOVAK WEB ARCHIVE

Peter Hausleitner & Jana Matúšková, University Library in Bratislava

DIGITAL RESOURCES NATIONAL PROJECT



- web archiving in Slovakia first attempts 2005/2006 ULIB participated in project Web Cultural Heritage
- there were no necessary legislative, organizational and technical conditions for systematic harvesting and archiving of web pages and original electronic resources in Slovakia
- Operational Programme Informatization of the Society national project Digital Resources Web Harvesting and e-Born Content Archiving
- implementation: 01. 04. 2015 − 31. 12. 2015 (application development, ICT installation, pilot harvests)
- ▶ 2016 2021: 6 years of routine operation
- November 2021: the end of sustainability of the Digital Resources Project

DIGITAL RESOURCES - ORGANISATIONAL SUPPORT

Digital Resources Project - Mission & Goal:

- to build the technological and organisational infrastructure for a controlled web harvesting and born digital documents (e-Born) archiving
- to provide systematic collecting and long-time archiving of Slovak websites and born digital content (electronic monographs and electronic serials) as an integral part of the Slovak Cultural Heritage
- 2 parts of the Digital Resources archive: the archive of Slovak web and the archive of electronic publications

Organisational support:

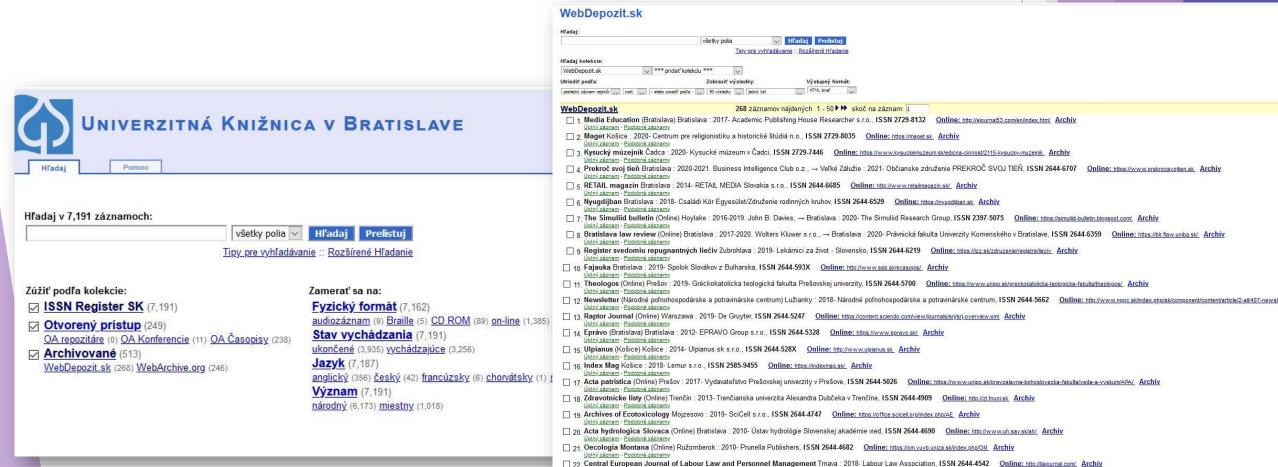
- Deposit of Digital Resources department of ULIB created in 2015
- Head of the Department (manager)
- Curator I Administrator of the Information System for Digital Resources
- Curator II Web Archive Curator
- Curator III Born Digital Archive Curator
- part-time person and manager for Born Digital Resources
- external capacities: support Service Level Agreement

ARCHIVING OF ELECTRONIC PUBLICATIONS

- electronic serials and monographs (e-Born, born digital) are archived in the special part of Digital Resources Archive
- ISSN/ISBN is a requirement for archiving
- cooperation with the National ISSN Centre, The Slovak National ISSN Database is the source of e-serials titles, the database is connected with the Digital Resources Archive (links to the archive)
- ▶ titles in PDF form archiving via file insertion (publisher or curator)
- titles in form of updated websites archiving via harvesting (curator)
- access and search via <u>www.webdepozit.sk</u> public portal (fulltext search)
- ▶ 31.10.2021: 394 titles in the archive 67 monographs and 327 serials (260 in PDF form a 67 in form of updated websites)
- e-Born titles with agreement: 130
- CC and OA titles: 167
- ▶ 53 titles of archived serials are no longer published, 25 of them are on the web unavailable

ARCHIVING OF ELECTRONIC PUBLICATIONS

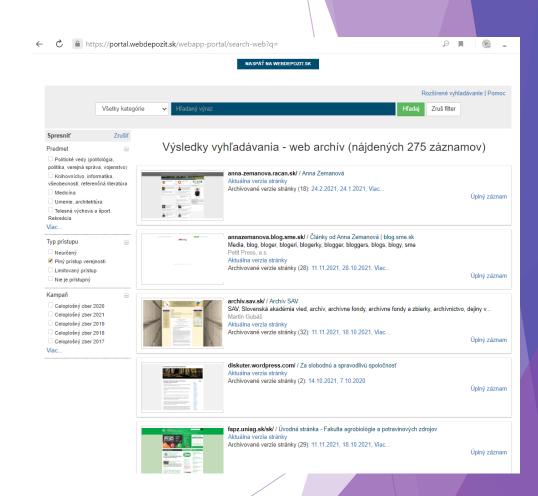
► The Slovak National ISSN Database and its connection with the Digital Resources Archive



Úplný záznam - Podobné záznamy

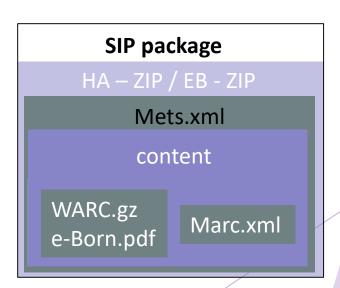
ACCESS TO THE ARCHIVE

- www.webdepozit.sk
- according to the actual Copyright Act
- default: without public access
- OpenAccess, CreativeCommons no agreement necessary
- in other cases it is necessary to sign an agreement with provider/publisher
- ▶ 3 types of access: public, local and forbidden
- local access to the archived content in the ULIB research room (26 PCs)



LONG-TERM PRESERVATION

- the content of the archive is sent into the Central Data Archive (University Library in Bratislava), which serves as a repository for long-term preservation
- ▶ the LTP storage is designed and operates in accordance with the requirements and standards for trusted long-time storages (ISO 16363, ISO 14721)
- ▶ 29 170 archival (SIP) packages (4,5860 TB) stored in the Central Data Archive (31.10.2021)
- SIPs are compressed
- ► SIP types: HA harvest & EB e-Born
- ► HA mets.xml, WARC marc.xml + warc.gz
- ► EB mets.xml, PDF marc.xml + name.pdf



DOMAINS IN IS DIR CATALOGUE (04TH NOVEMBER 2021)

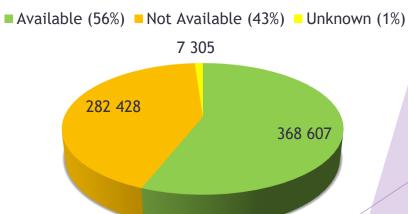
- nic.sk 442 701 registered .sk domains
- ▶ DIR web catalogue:

multiple-leveled domains

domains other than .sk

- born-digital titles websites
- domains extinct
- ▶ 658 340 domains in DIR catalogue
- harvested at least once 654 428 domains
- not available today 281 961 dom-s
- more than 5 MB WARCs 71 512 dom-s





DOMAIN SHARE* IN THE CATALOGUE (15.11.2021)

* top-level domains

TLD Count Percent 99,87 sk 658 844 0,04 283 eu 0,03 217 com 145 0,02 org 0,01 CZ 61 info 49 0,01 0,01 35 net online de other 0,01 40

Automatically - sk-nic.sk + manually inserted (blog.domain.sk, www.site.sk/sth, ...)

manually inserted <

WHOLE-DOMAIN CRAWLS

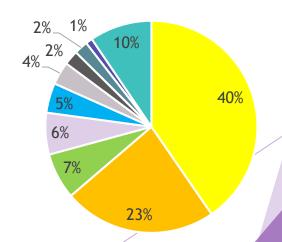
Basic settings / per domain	2016	2017	2018	2019	2020	2021
Max. time limit	1 hour	2 hours	2 hours	3 hours	4 hours	4 hours
Max. size	200 MB	400 MB	400 MB	400 MB	400 MB	300 MB
Max. objects	5 000	10 000	10 000	10 000	20 000	7 000
Domains in harvest	360 628	386 228	459 832	533 883	594 434	653 471

2021

316 233 successfuly harvested domains 16 TB more than 255 mil. objects

W-DC 2021 - Object types

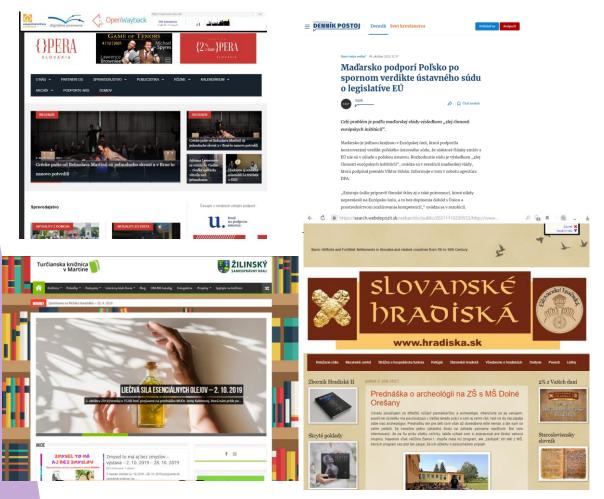




SELECTIVE HARVESTS 82 CAMPAIGNS UNTIL NOV. 16TH '21

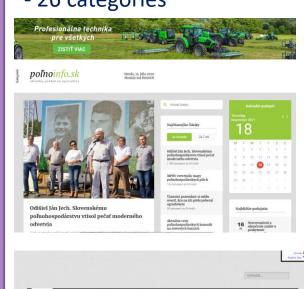
Cooperating subjects:

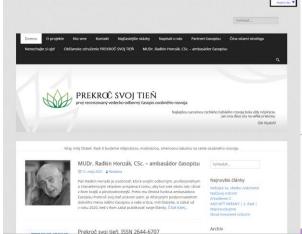
- 229 URLs, 223 active agreements

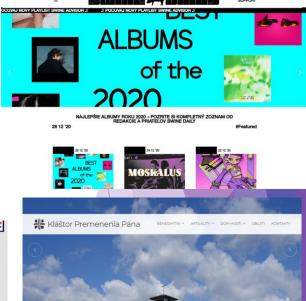


Conspectus categories:

- 26 categories







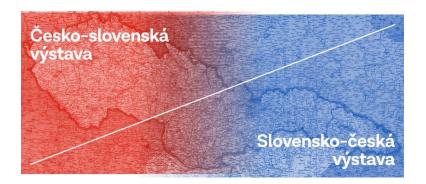


Benediktíní sú pokračovateľmi starobylého mníšskeho hnutia Katolíckej církvi, ktorého korene sishahjú do prvých storočí kresťanstva. Regula svátého Benedikta, napisaná na konci 6. storočka odoreh shawým programovým dokumentom kažého benediktínskeho kklástra. Kložičov Premenenia Pána v Sampore je od roku 2010 domovom nového mníšskeho spoločenstva, snažiaceho sa obnoviť mníšsky spôsob života na Slovensku. Od roku 2012 je klášťor nezávislým priorátom a plnohodonotným dlemom Benediktínsky kongregácie Zvestvania, Predstaveným klášťora je otec prior Vlodímir Koson OSB. V súčasnosti prebieha v Sampore druhá etapa výstanby klášťora, ktorej cielom je poskyntú dostatok priestoru pre rozrastajúce sa spoločenstvo, ale a ponúkut nové možnost pre mávšetenikov, ktori maj už dnes možnosť



EVENT-BASED HARVESTS 57 CAMPAIGNS UNTIL NOV. 16TH '21



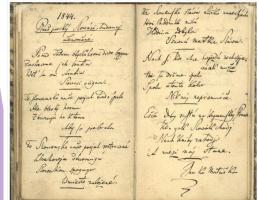














Američan sa nakazil koronavírusom dvakrát, druhýkrát to bolo vážnejšie

V Spojených štátoch sa 25-ročný muž nakazil koronavírusom dvakrát, pričom druhýkrát malo u neho ochorenie covid-19 vážnejší priebeh, musel vyhľadať pomoc v nemocnici. S odvolaním sa na štúdiu vo vedeckom časopise The Lancet o tom dnes informoval spravodajský veb BBC.





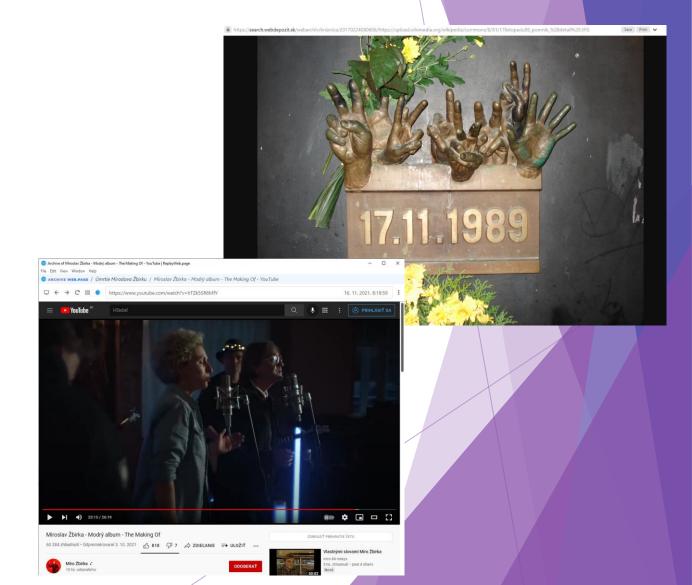
EVENT-BASED HARVESTS COOPERATION WITH WEBARCHIV.CZ

- events with impact on both countries
- seed exchange
- two collaborative harvests:

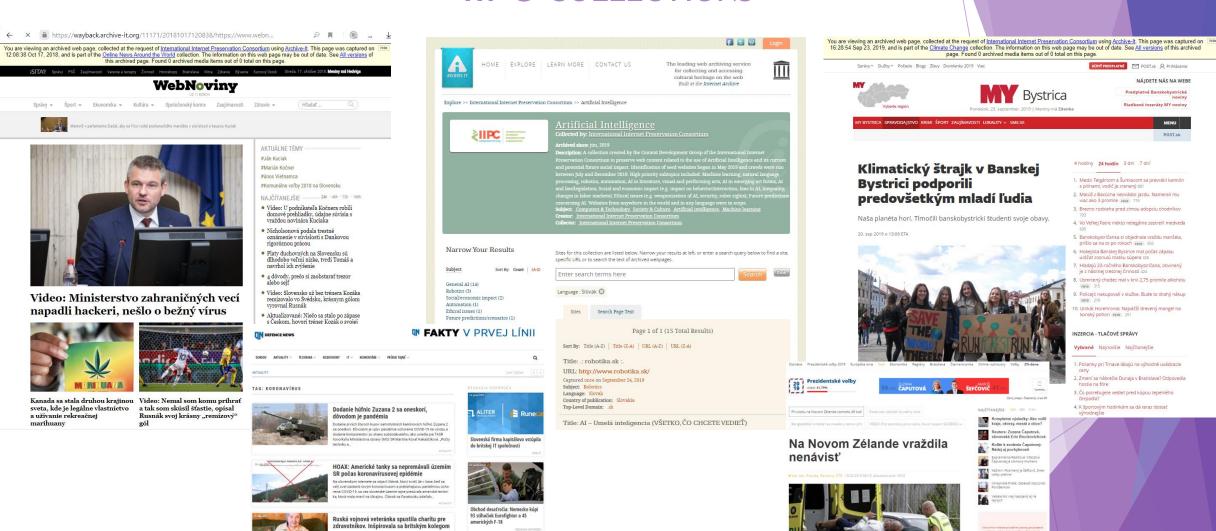
30th Anniversary of Velvet revolution, Nov. 2019

Death of Miroslav Žbirka, Nov. 2021





IIPC COLLECTIONS



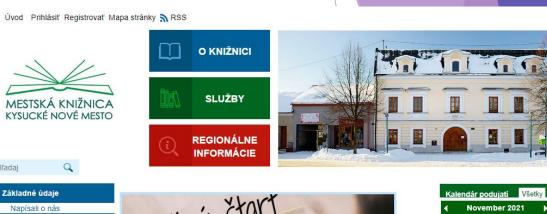
HOAX: Mobilná technológia 5G nemá nič spoločné s novým koronavírusom

NAKA tvrdí, že šéf vojenského spravodajstva získal miliónové majetky legálne

pravicového teroristu na moslimskú komunitu. Jeho obeťami vo dvoch mešítách v meste Christchurch počas piatkových modlitieb sa stalo 49 ľudi.

Nenávisť a zvrátenosť. Novým Zélandom otriasol hrutálny útok kraine





Hľadaj

Kontakt

Projekty

Fotogaléria

Užitočné linky Vypožičný čas

Online katalóg

Služby knižnice

Spýtajte sa knižnice

Elektronické informačné

Podujatia pre školy

Požiadavka na nákup kníh

Knižný štart

Ostatné projekty do roku



MOBILNÁ

aplikácia



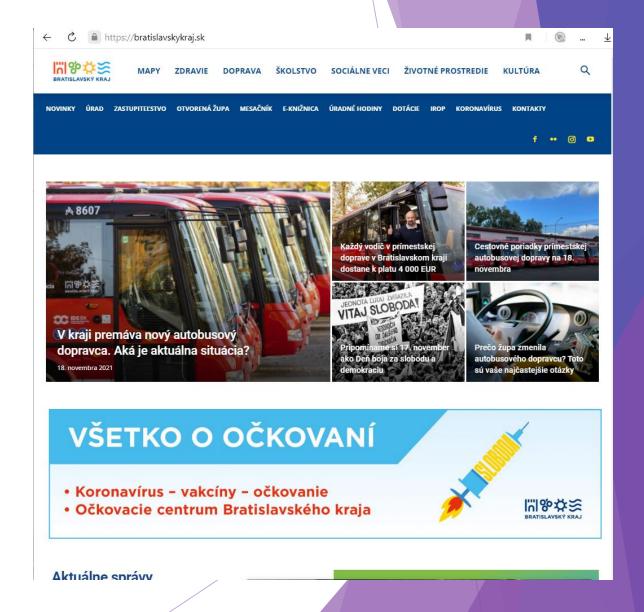


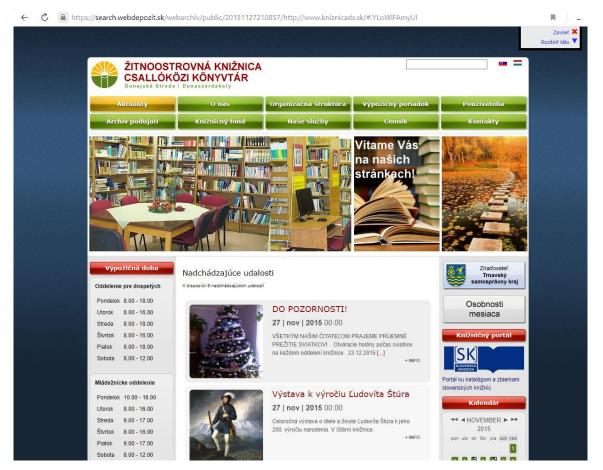
Library of Kysuce in Čadca - 2015 vs. Nov. 2021

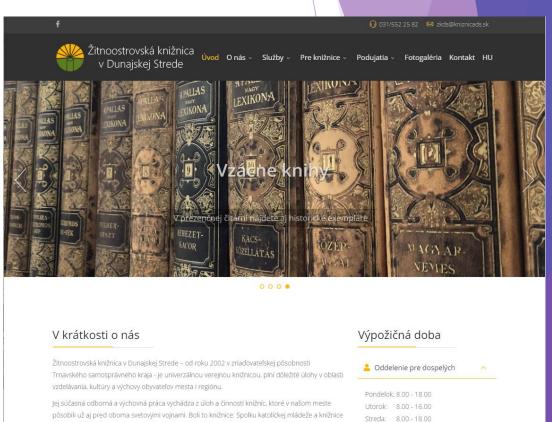
SITE CHANGES

(WEBDEPOZIT ARCHIVE / LIVE SITE)









Štvrtok: 8.00 - 16.00

Piatok: 8.00 - 18.00

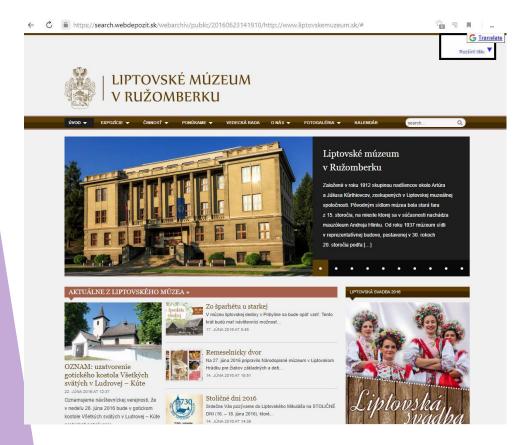
Sobota: 8.00 - 12.00

škôl mesta bohaté na krásnu aj populárno-náučnú literatúru. Počas vojny sa veľká časť týchto kníh

zničila a stratila, ale po roku 1946 obecná rada obnovila zničenú knižnicu s cieľom pomôcť v šírení

oddychovej kultúry medzi obyvateľstvom. Táto knižnica mala len jedného odborného pracovníka,

Rye Island Library Dunajská Streda - 2015 vs. Nov. 2021



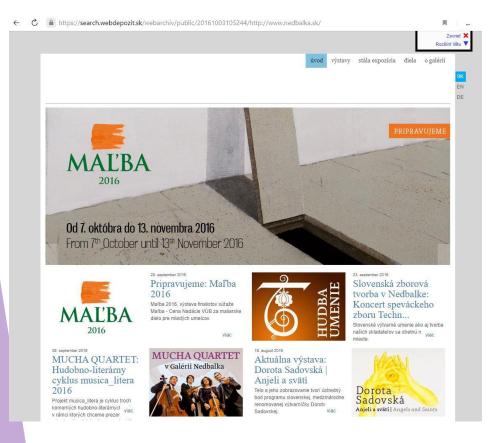


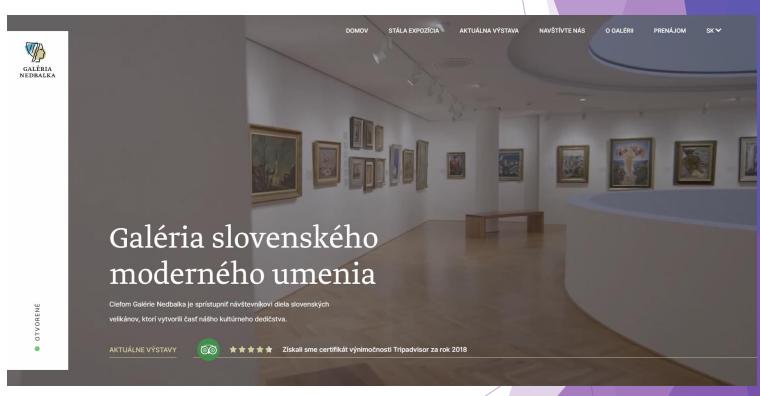
Ludrová - Kút

Liptovský Mikuláš

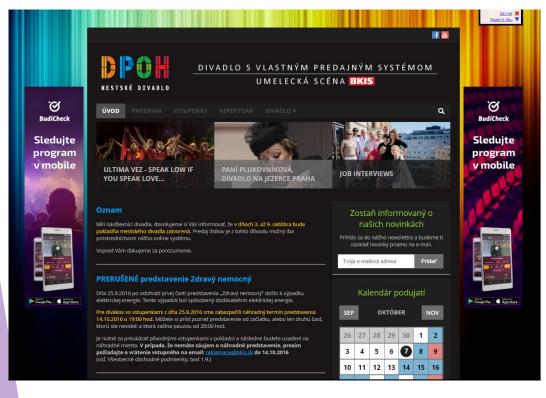
Liptovský Hrádok

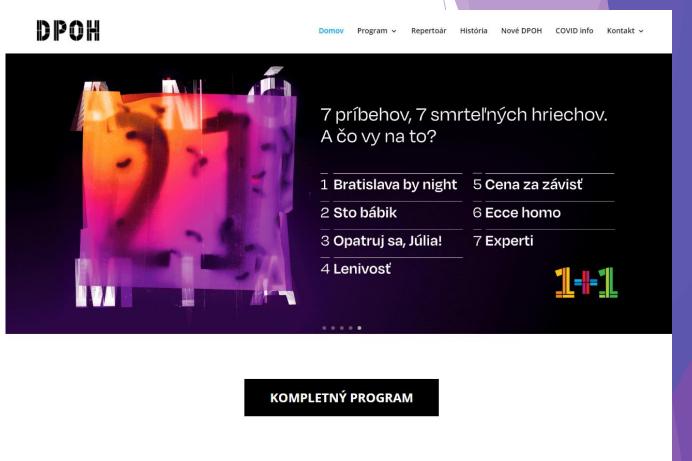
Museum of Liptov in Ružomberok - 2016 vs. Nov. 2021





Nedbalka Gallery Bratislava - 2016 vs. Nov. 2021





P. O. Hviezdoslav Theater, Bratislava - 2016 vs. Nov. 2021

SITES EXTINCT

YOU WON'T FIND THESE ANYMORE (404, FORBIDDEN, ...)

Novinky

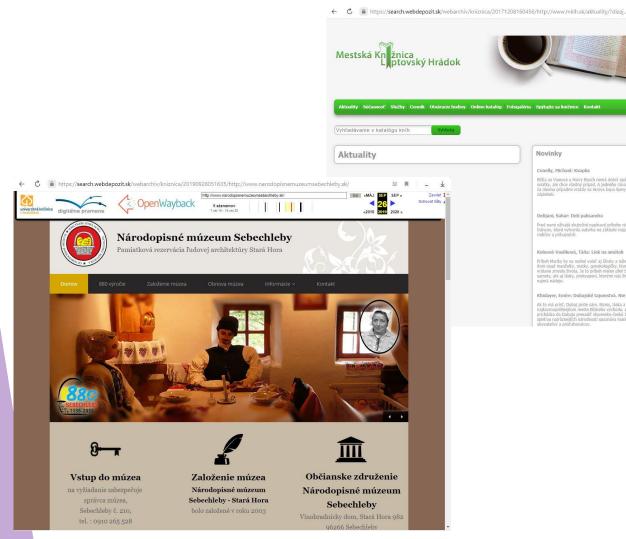
Conelly, Michael: Kvapka

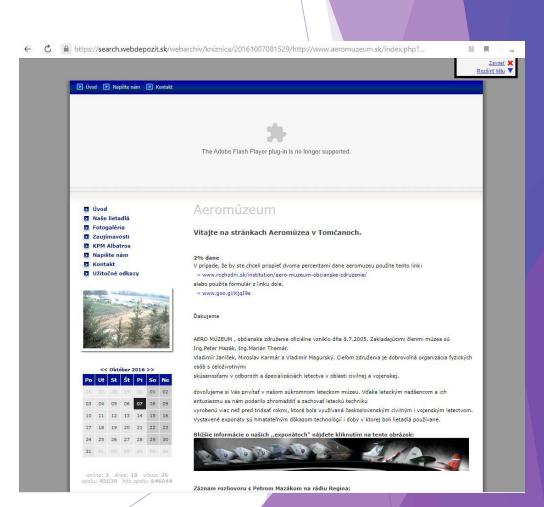
Delijani, Sahar: Deti palisandra

Keleová-Vasilková, Táňa: Liek na smútok

Príbeh Maríky by sa mohol volať aj Straty a nálezy. Rozvíja sa v ňom osud manželky, matky, gynekologičky, ktorá stráca všetko vrátane zmyslu života. Je to príbeh nielen plný bolesti, smútku a

Blížia sa Vianoce a Harry Bosch nemá dobrý spánok – nie pre sviatky, ale chce vlastný prípad. A jedného rána dostane naraz dva, za oboma prípadmi vraždy sa skrýva kopa špíny a nečakaných zápletek.





LEGISLATION

- Legal Deposit Act 212/1997
- there is no covering of web archiving and electronic legal deposit in the legislation now
- October 2021 draft on new Electronic Publications Act (about publication issues and media and audiovisual registers (MoC of SR)
 - defining web portals as daily updated communicate, of news nature, as a sort of periodical publication
 - ▶ § 14: Web portal provider is due to allow the University Library in Bratislava to create a copy and save the content of the portal into the long-term web cultural heritage digital deposit fund

PROBLEMS AND CHALLENGES

Technical problems:

- ► Heritrix restrictions searching for new options of collecting multimedia resources and social networks, Webrecorder (ArchiveWeb.page)
- robots.txt (respecting, uncollected content, styling)
- ▶ formats, non-mainstream CMS (Buxus, ...), plugins, themes, dynamic content

Legislative restrictions:

- ▶ the necessity to amend the Copyright Act and the Legal Deposit Act to extend the scope on archiving and making available electronic online publications and websites published in Slovakia
- Legal Deposit for online publications would make our job easier, there would be no need to sign agreements with providers/publishers
- ▶ Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and its implementation into Slovak legislation (exceptions in the field of data mining for institutions dealing with the preservation of cultural heritage)

WHAT NEXT?

- document Strategy for the Development of Digital Resources after the End of Sustainability (10/2021)
 intentions for the future
- archiving of new types of sources and formats (multimedia content, videos, social networks)
- connection of Digital Resources catalogues with the University Library in Bratislava and the Slovak National Library (SNL) catalogues
- cooperation in the selection of resources suitable for archiving (SNL, scientific libraries, government institutions, scientific institutes of the Slovak Academy of Sciences)
- solution of dissemination of SIP packages from CDA
- ▶ SW and HW for data re-export and data extraction
- cooperation within V4 countries new tools, data exchange, data mining...
- cooperation with IIPC creation of international collections
- survey about the needs of researchers

RESOURCES

- Archived content www.webdepozit.sk
- Webarchiv logo https://webarchiv.cz/cs/
- IIPC collections (www.archive-it.org/)
- ► Free images from www.pexels.com

THANK YOU FOR ATTENTION

Ing. Peter Hausleitner

peter.hausleitner@ulib.sk

University Library in Bratislava

Deposit of Digital Resources (ddp@ulib.sk)

www.webdepozit.sk











