

Depozit slovenského webu

Ing. Peter HAUSLEITNER

Ing. Alojz ANDROVIČ, PhD.

Bc. Andrej BIZÍK

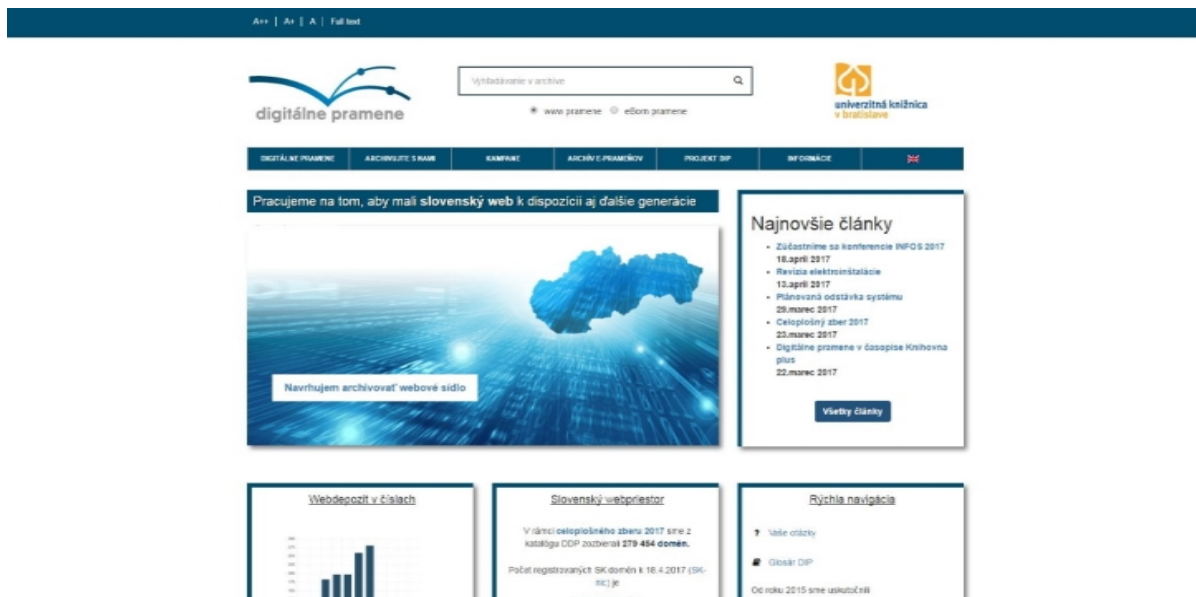
PhDr. Jana MATÚŠKOVÁ

Sympóziu INFOS 2017, 25. 4. 2017

www.webdepozit.sk

Národný projekt Digitálne pramene – webharvesting a
archivácia E-Born obsahu

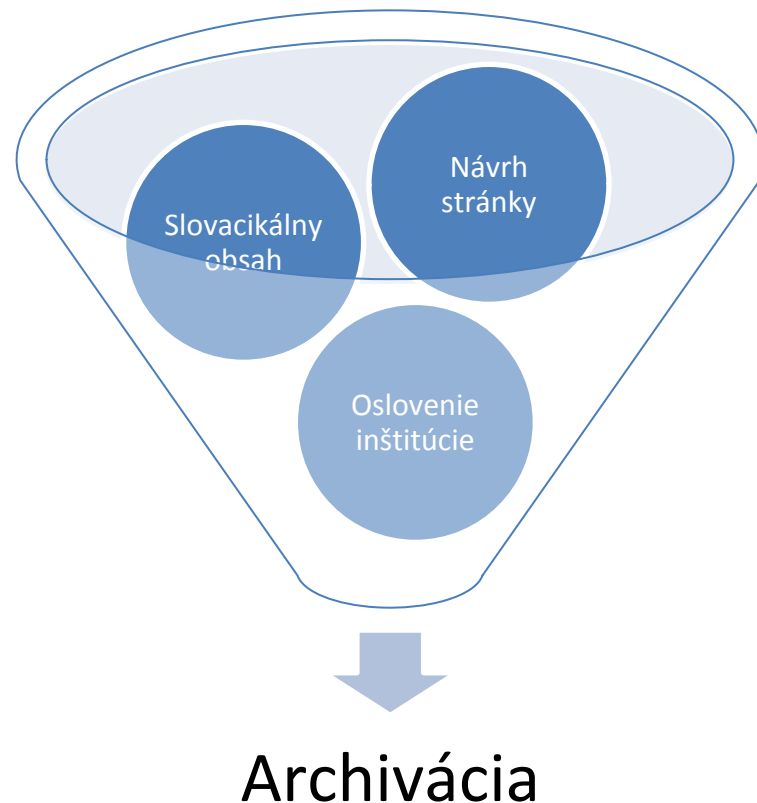
Archivujeme webové stránky a E-Born pramene



The screenshot shows the homepage of the webdepozit.sk website. At the top, there is a navigation bar with the text "Pracujeme na tom, aby mal slovenský web k dispozícii aj ďalšie generácie". Below this, there is a search bar and a navigation menu with categories like "DIGITÁLNE PRAMENE", "ARCHIVÁCIE S MAIB", "KAMPAŇE", "ARCHÍV E-PRAMENŮV", "PROJEKT OP", "INFORMÁCIE", and "O NÁS". The main content area features a large blue graphic with the text "Navrhujem archivovať webové sídlo". To the right, there is a section titled "Najnovšie články" (Latest articles) with a list of recent publications and dates. At the bottom, there are three smaller sections: "Webdepozit v číslach" (Webdeposit in numbers) with a bar chart, "Slovenský webpriestor" (Slovak web space) with statistics, and "Rýchla navigácia" (Quick navigation) with links to "Všetky články" and "Oblasť DP".

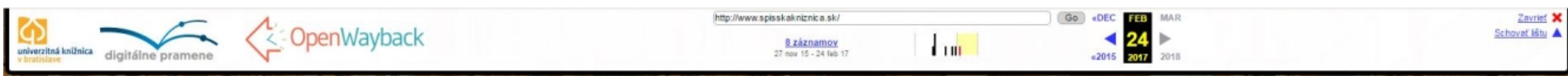
Archivácia webu

- zachytiť, uchovať
a v budúcom horizonte
spätne sprístupniť
uložené webové stránky



Technológie

- Heritrix, OpenWayback, SOLR, LAMP, JAVA, PostgreSQL



Obr.: lišta OW s dátumom archivácie zobrazenej stránky, interný portál DIP, výsledky vyhľadávania v katalógu

Názov predlohy	Typ predlohy	Plánované spustenie	Plánované ukončenie	Plánovaný počet	Najbližšie spustenie	Počet spustení	Aktívna predloha
Celoplošný zber 2016	Celoplošný zber	7.10.2016	7.10.2016	-	-	2	✓
Celoplošný zber 2017	Celoplošný zber	24.2.2017	24.2.2017	-	-	2	✓
E-šors WWW	Výberový zber	28.3.2017	28.3.2017	-	-	3	✓
Kultúrny profil Slovenska	Výberový zber	27.12.2016	27.12.2016	-	-	3	✓
Kultúrny profil Slovenska	Výberový zber	26.7.2016	26.7.2016	-	-	4	✓
Kultúrny profil Slovenska	Výberový zber	28.12.2016	28.12.2016	-	-	2	✓

Výsledky vyhľadávania - web archiv (nájdenej 92 záznamov)

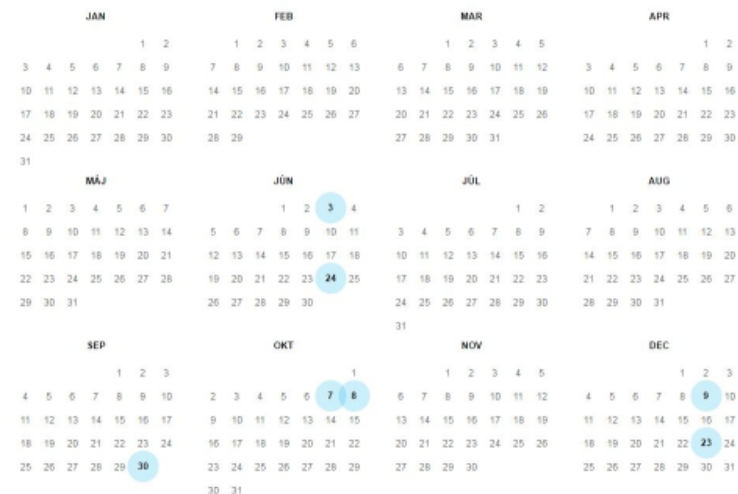
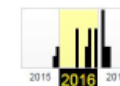
- Prílohy vedy, politiky, práva, verejné sprisah, spravosť**
- Prílohy vedy, politiky, práva, verejné sprisah, spravosť**
- Prílohy vedy, politiky, práva, verejné sprisah, spravosť**
- Prílohy vedy, politiky, práva, verejné sprisah, spravosť**

Technológie



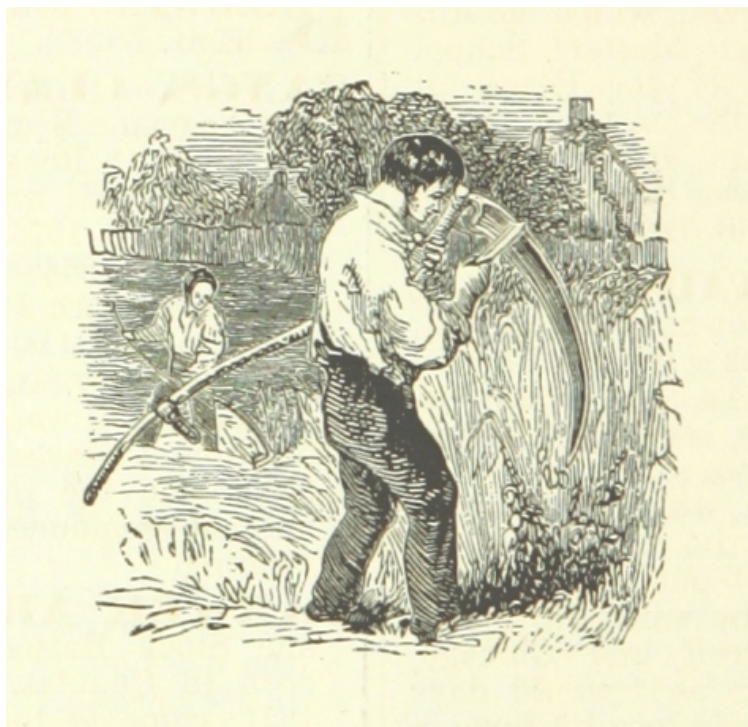
<http://www.archiv.sav.sk/> Hľadaj

<http://www.archiv.sav.sk/> bolo preskúmané 27 krát, ič späť na 23. novembra 2015. Priskúm môže byť duplikátom predchádzajúceho. Stáva sa to asi v 25% prípadoch z 420,000,000 stránok.



Obr.: zobrazenie archivovanej stránky a kalendár so záznamami dátumov archivácie

Zber webu (web harvest)



- proces automatického stiahnutia obsahu webových stránok na účely ich dlhodobej ochrany a archivácie

Kampane:

- celoplošné
- tematické
- výberové

Obr.: Scythe, The history of Winnebago County, Ill., its past and present. ... Illustrated, Europeana Collections, Public Domain

Prístup k archivovanému obsahu

- implicitne bez verejného prístupu
- Open Access, Creative Commons
- je možné definovať na základe Zmluvy o poskytovaní elektronických online prameňov (verejný, v priestoroch knižnice, bez prístupu)

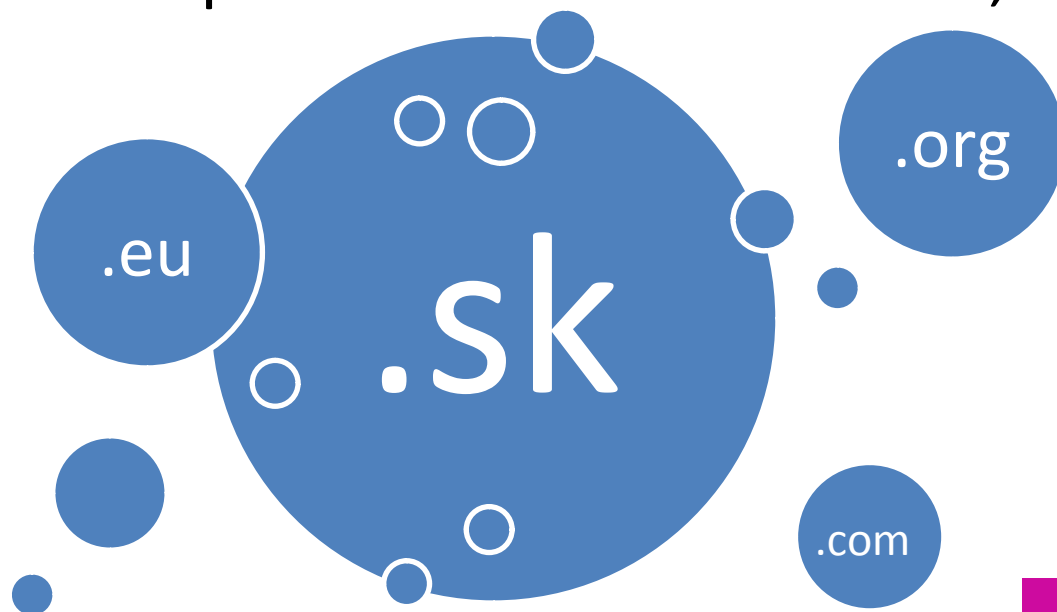
110 zmlúv,
takmer 150 URL



Obr.: Gate in Puebla Mexico,
Rebecca Hardgrave, flickr, CC
Attribution

Celoplošné zbery

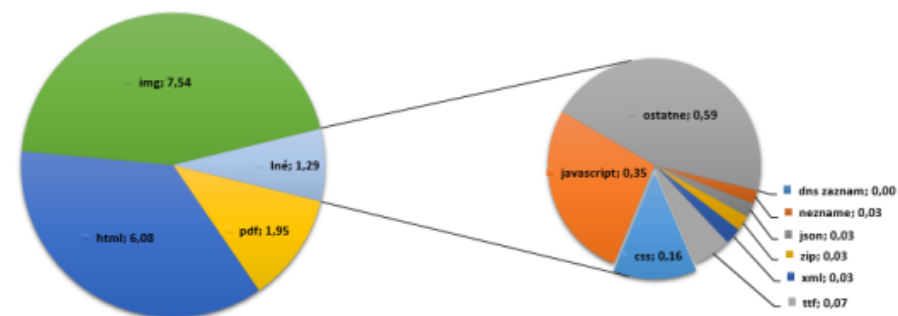
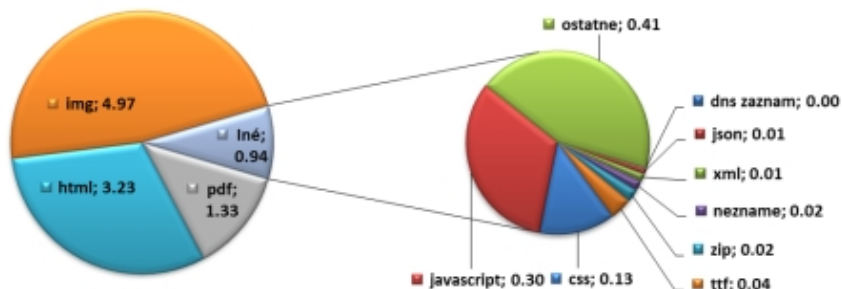
- získanie obrazu o slovenskom webe
- pilotný zber v decembri 2015
- rutinná prevádzka – október 2016, február 2017



2016 vs 2017

- 200 MB, 5000 objektov, 1 hod./ doména
- 10,46 TB nekomprimovaných dát

- 400 MB, 10 000 objektov, 2 hod./ doména
- 14,19 TB nekomprimovaných dát



Grafy: štruktúra objektov pri celoplošných zberoch 2016 a 2017, veľkosť v TB

2016 vs 2017

- 360 628 domén v katalógu
 - 278 610 zozbieraných úspešne
 - 16 253 zozbieraných neúspešne
 - 65 765 vynechaných domén
- 386 228 domén v katalógu
 - 279 454 zozbieraných úspešne
 - 24 913 zozbieraných neúspešne
 - 81 861 vynechaných domén

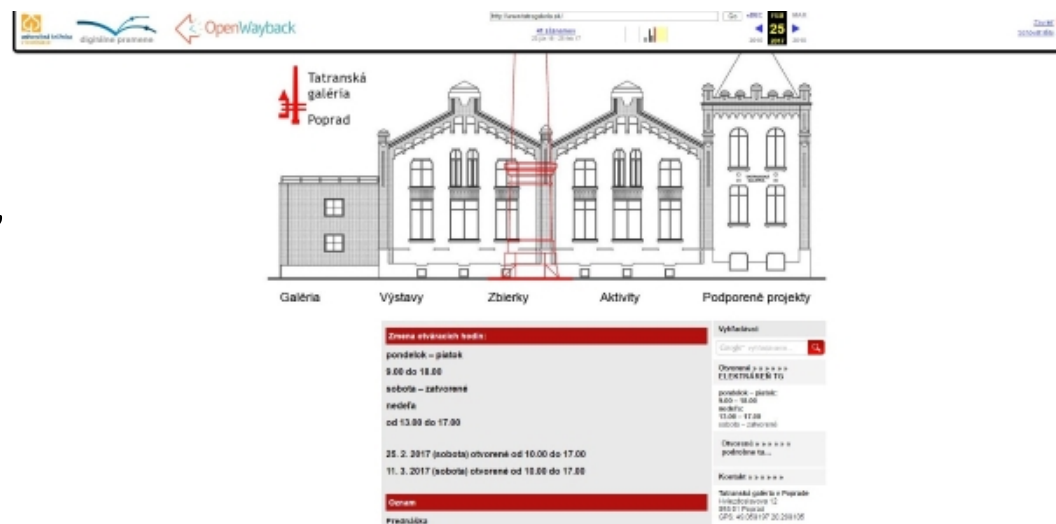
	Priemerná veľkosť pozbiaraného obsahu	URL s dosiahnutými limitmi objemu	URL s dosiahnutými limitmi objektov	Počet WARC súborov
2016	39,35 MB/URL	20 562 (200 MB)	8 872 (5 000 obj.)	278 663
2017	63,26 MB/URL	16 461 (400 MB)	5 227 (10 000 obj.)	232 747

Tab.: Porovnanie ukazovateľov celoplošných zberov

Tematické zbery

- webové stránky sa vyberajú podľa témy a aktuálneho diania
- štandardné nastavenia: 99 999 999 objektov, 1 GB, dni/ doména

Obr.: Archivovaná stránka
Tatranskej galérie v Poprade



Realizované tematické zbery WWW

2015

Pilotný tematický zber:

1. Kultúra
2. Vzdelávanie
3. Veda a spoločnosť

2016

Parlamentné voľby 2016

Kultúrny profil Slovenska:

- knižnice
- galérie
- múzeá

Tour de France 2016

Olympijské hry Rio 2016

Predsedníctvo SR v Rade EÚ

2017

Kultúrny profil Slovenska –
divadlá

Pravidelné mesačné zbery

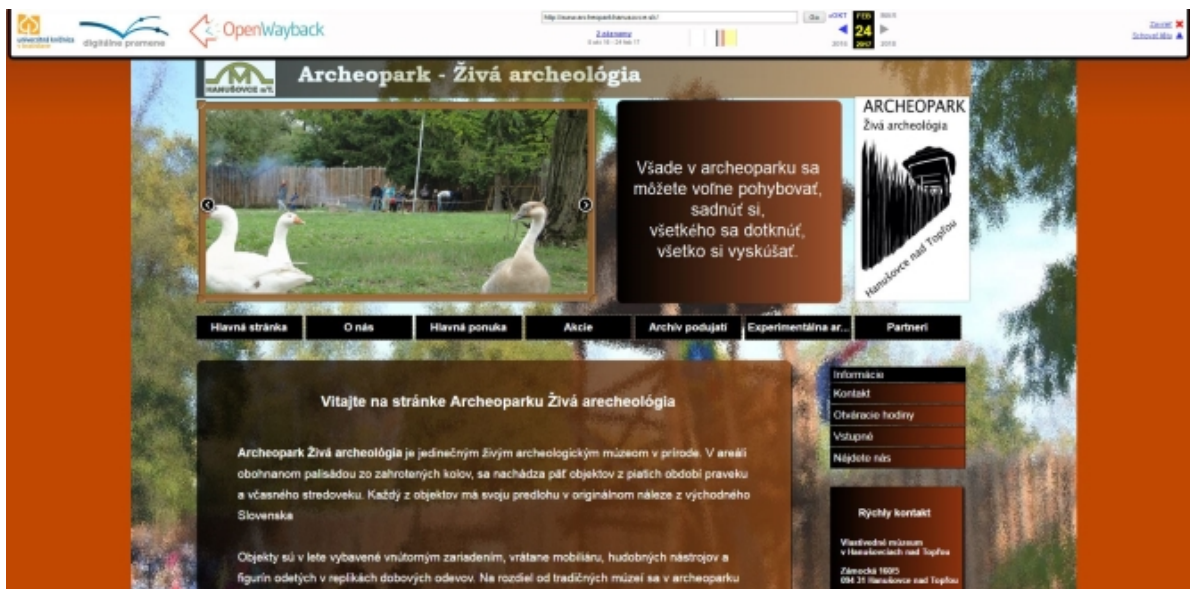
www.ulib.sk a

www.webdepozit.sk



Výberové zbery

- zbery domén zazmluvnených inštitúcií
- prebiehajú spravidla 4 x ročne
- štandardné nastavenia:
1 500 000 objektov, 2 GB, 3 dni/ URL



Obr.: snímka obrazovky archivovanej stránky
www.archeoparkhanusovce.sk k 24.2.2017

Bádateľňa webdepozitu



- špecializovaný priestor v rámci Univerzitnej knižnice v Bratislave slúži na prístupňovanie archivovaného obsahu - webových prameňov a pôvodných (E-Born) dokumentov



Ďakujem za pozornosť 😊



Obr.: Knihovníčka Lucinda deLorimier pri oslave 40. výročia Truckee Library.

Použité zdroje

Literatúra

1. CELBOVÁ, Ludmila, Lukáš GRUBER, Tomáš SÍBEK a Libor COUFAL. Archivace webu. Praha: Národní knihovna České republiky, 2008. 45s.
2. ANDROVIČ, Alojz, Andrej BIZÍK, Peter HAUSLEITNER, Beáta KATRINCOVÁ, Iveta LACKOVÁ a Jana MATÚŠKOVÁ. Digitálne pramene – národný projekt zberu a archivácie v roku 1. *Knihovna Plus* [online]. Národní knihovna ČR. 2017, č. 1. ISSN 1801-5948. Dostupné na internete: <http://knihovnarevue.nkp.cz/kplus-web/archiv/2017-01/historie-a-soucasnost/digitalne-pramene-2013-narodny-projekt-zberu-a-archivacie-v-roku-1#lib>
3. ANDROVIČ, Alojz, Ivan CIGLAN a Jana MATÚŠKOVÁ. Digitálne pramene – webharvesting a archivácia e-Born obsahu. *ITlib* [online]. Centrum vedecko – technických informácií SR. 2016, č. 2. ISSN 1336-0779. Dostupné na internete: http://itlib.cvtisr.sk/buxus/docs/05_digitalne%20pramene.pdf

Použité zdroje

Grafické prvky

1. Snímky obrazoviek, grafy a fotografie - www.webdepozit.sk
2. HAUSLEITNER, Peter. Horí. Public Domain. Dostupné online: https://www.webdepozit.sk/obrazky/hori_publ.dom.svg
3. HARDGRAVE, Rebecca. Gate in Puebla Mexico. Flickr. CC Attribution. Dostupné online: <https://www.flickr.com/photos/hardgravephoto/28978309371>
4. Kosa. The history of Winnebago County, Ill., its past and present. ... Illustrated. Europeana Collections. Public Domain. Dostupné online: http://www.europeana.eu/portal/en/record/9200387/BibliographicResource_3000117256703.html
5. Ikony Font Awesome. Dostupné online: <http://fontawesome.io/icons/>
6. Knižovnička Lucinda deLorimier pri oslave 40. výročia Truckee Library. Dostupné online: <http://www.sierrasun.com/news/local/video-officials-past-and-present-celebrate-truckee-librarys-40th-b-day/>