Vyhodnotenie celoplošného zberu 2020

Slovenská doména .sk vyhradená pre slovenský internet vznikla v roku 1993 a jej správou je poverená spoločnosť SK-NIC. SK-NIC evidovala ku koncu roka 2020 viac ako 420 000 zaregistrovaných SK domén. V slovenskom webovom archíve Digitálnych prameňov sa pred celoplošným zberom 2020 nachádzalo 594 434 domén rôznych úrovní. Popri aktuálne zaregistrovaných doménach sa tam nachádzajú zaniknuté domény, domény vyšších úrovní, domény pre e-Born zbery a domény slovacikálneho charakteru s inou doménou ako .sk.

V októbri 2020 sa v rámci Digitálnych prameňov uskutočnil celoplošný zber. V poradí piaty celoplošný zber všetkých dostupných slovacikálnych webových sídiel katalógu DIP prebiehal od 1.10.2020 do 21.10.2020. Do zberu bolo zaradených 594 434 domén z katalógu DIP.

Na základe aktuálne platnej Politiky zberu webových prameňov boli zo zberu vylúčené multimediálne formáty (video a audio súbory), sociálne siete, súbory zakázané na základe uzatvorených licenčných zmlúv, kompresné formáty ako aj ďalšie súbory a formáty nespadajúce do politiky zberu.

Pred spustením zberu sme oslovili najväčších registrátorov domén na Slovensku s požiadavkou vzájomnej spolupráce, aby sme predišli problémom počas priebehu zberu. Išlo najmä o notifikácie o prekročení limitov alebo prípadnom zablokovaní IP adries. Odoslali sme im informácie o dátume spustenia zberu spolu s parametrami zberu a IP adresami workerov.

Pre celoplošný zber 2020 boli nastavené maximálne limity na doménu 20 000 objektov (maxDocumentsDownload), 4 hodiny (maxTimeSeconds) a 400 MB (maxBytesDownload).

Ďalšie nastavenia, ktoré sme prispôsobili, boli nasledovné:

- **delayFactor** – Tento parameter spôsobí oneskorenie medzi načítaním URI od toho istého hostiteľa. Napríklad, ak by bolo potrebné načítať posledný identifikátor URI od hostiteľa a delayFactor je 5 (veľmi vysoká hodnota), potom Frontier počká 4000 milisekúnd (4 sekundy) predtým, než povolí spracovanie iného URI od tohto hostiteľa.
- **maxDelayMs** - Tento parameter stanovuje maximálnu hornú hranicu čakacej doby vytvorenej funkciou delayFactor. Ak je nastavená na 30000 milisekúnd, maximálne oneskorenie medzi údajmi URI od toho istého hostiteľa nikdy neprekročí túto hodnotu.
- **minDelayMs** - Tento parameter stanovuje minimálnu hranicu čakacej doby. Má prednosť pred hodnotou vypočítanou oneskorením Factor. Napríklad hodnota minDelayMs môže byť nastavená na 3000 milisekúnd. Ak delayFactor vygeneruje iba 20 milisekundové čakanie, hodnota minDelayMs ho prepíše a načítanie URI sa oneskorí o 100 milisekúnd.
- **maxPerHostBandwidthUsageKbSec** – Parameter maximálnej šírky použitého pásma v KB/s .
- **maxRetries** - Tento parameter obmedzuje počet opakovaných pokusov o obnovu URI v dôsledku chýb.
- **retryDelaySeconds** – Parameter čakacej doby medzi opakovaniami parametra maxRetries.
- **fetchHttp. timeoutSeconds –** Parameter čakacej doby na požadovaný zdroj URI. Ak načítanie URI nie je dokončené v určenom počte sekúnd, ukončí sa čakanie na požadovaný zdroj, ktorý skončí s chybou a zber pokračuje v ďalších URI.
- **maxBytesDownload –** Maximálna veľkosť na zbieranú doménu. Hodnota je približná, ak sa prekročí nastavená hodnota, vždy sa archivuje posledný objekt vcelku.

- **maxDocumentsDownload –** Maximálny počet objektov na 1 doménu. Pod objektom sa rozumie txt, css, js, html, pdf, word, mp4 a všetky povolené. Maximálne ich spolu môže byť iba tento určený počet.
- **maxTimeSeconds –** Maximálna doba archivácie jednej URI.

Súčasťou nastavení je aj tzv. deduplikácia, ktorá identifikuje rovnaké objekty. Pri zistení rovnakých objektov sa tento objekt opätovne neukladá, ale vytvára sa odkaz na objekt, ktorý bol v minulosti už uložený. Takto vznikajú naprieč archívom prepojenia na prvý pôvodný objekt. Pre zabezpečenie správnej integrity archívu je po určitom čase potrebné deduplikáciu obnoviť. Po tomto kroku sa pri nasledujúcom zbieraní obsahu všetky objekty uložia nanovo. Takéto obnovenie tzv. deduplikačnej databázy sme urobili v roku 2018. V r. 2020 sme túto deduplikáciu neobnovovali a preto sa pri celoplošnom zbere vytvorili WARC súbory s odkazmi na rovnaké objekty z minulých zberov, ak také už v archíve existovali.

Pred spustením celoplošného zberu sa na testovacom prostredí overila konfigurácia v obmedzenom rozsahu domén s upraveným objemom pre doménu. Pred spustením sme ešte skontrolovali modifikovaný konfiguračný súbor a súbor s výnimkami nastavenými podľa politiky zberu.

Pri spúšťaní zberu sme v roku 2020 ponechali zapnutých 80 workerov (robotov) z celkového počtu 240. Každý worker obsahuje 15 heritrixov (zberačov), čo predstavuje maximálne 1200 zberov URL naraz. Tento počet bol odhadnutý na základe skúseností z predchádzajúcich rokov, kedy sme workery spúšťali postupne. Našim cieľom bolo zrealizovať zber do 30 dní.

Spustenie a priebeh zberu sme postupne kontrolovali a zaznamenávali. Počas priebehu zberu sme sledovali počty pribúdajúcich zozbieraných domén a vyhodnocovali ich, aby sme na ich základe vedeli určiť približný koniec zberu. Neevidovali sme žiadne problémy od registrátorov, poskytovateľov ani držiteľov domén. Zber prebehol za 21 dní a nespôsobil veľkú sieťovú záťaž.

Zber bolo potrebné nakoniec ukončiť manuálne, nakoľko sa vyskytli rôzne softvérové chyby a zber niektorých domén sa neukončil v požadovanom limite.

Po úspešnom ukončení zberu sme vyhodnotili domény prekračujúce nastavené limity. Zo všetkých úspešne zozbieraných URL bola priemerná veľkosť zozbieraného nekomprimovaného obsahu na jednu doménu 51 MB (16,12 TB / 330881 URL). 29 stránok skončilo na limite počtu objektov, tento zadaný limit ovplyvnil zber iba minimálne. Naopak, nastavenie časového limitu na štyri hodiny ovplyvnilo objem, keďže zber skončil skôr, ako mohol pozberať väčšiu časť stránky a nenaplnil limit objemu 400 MB. (Tabuľka 1).

Časový limit	4 hodiny
Dátový limit	400 MB
Limit objektov	20000 objektov
Úspešne ukončené rozdelenie	**Počet URL z katalógu DIP**
Úspešne ukončené –Finished	256321
Ukončené na veľkosť - Maximum amount of data limit hit	8047
Ukončené na objekty - Maximum number of documents limit hit	29
Ukončené na čas - Timelimit hit	66484
Celkový súčet	**330881**

Tabuľka 1: Tabuľka nastavení limitov a počty ich dosiahnutia

Z celkového počtu 594 434 domén v katalógu DIP sa úspešne zozbieralo 330 881 domén slovacikálneho charakteru v objeme približne 16,12 TB (Tabuľka 2). Celoplošný zber v danom nastavení trval približne 21 dní, čo predstavuje priemerný čas zberu na jednu doménu 3,05 sekundy (594 434).

Celoplošný zber 2020	Počet	GB	TB	
Nekomprimovaná veľkosť		16502,53	16,12	
Komprimovaná veľkosť		6047,26	5,91	
Celkový počet URL z katalógu DIP	594 434			
Úspešné URL z katalógu DIP	330 874	+7 manuálne ukončené		
Neúspešné URL z katalógu DIP	31 325			
Zrušené/vynechané zo zberu	232 235		**Spustenie**	1.10.2020
				08:45
			Ukončenie	21.10.2020
Počet WARC	330 956			02:14

Tabuľka 2: Základné údaje o celoplošnom zbere 2020

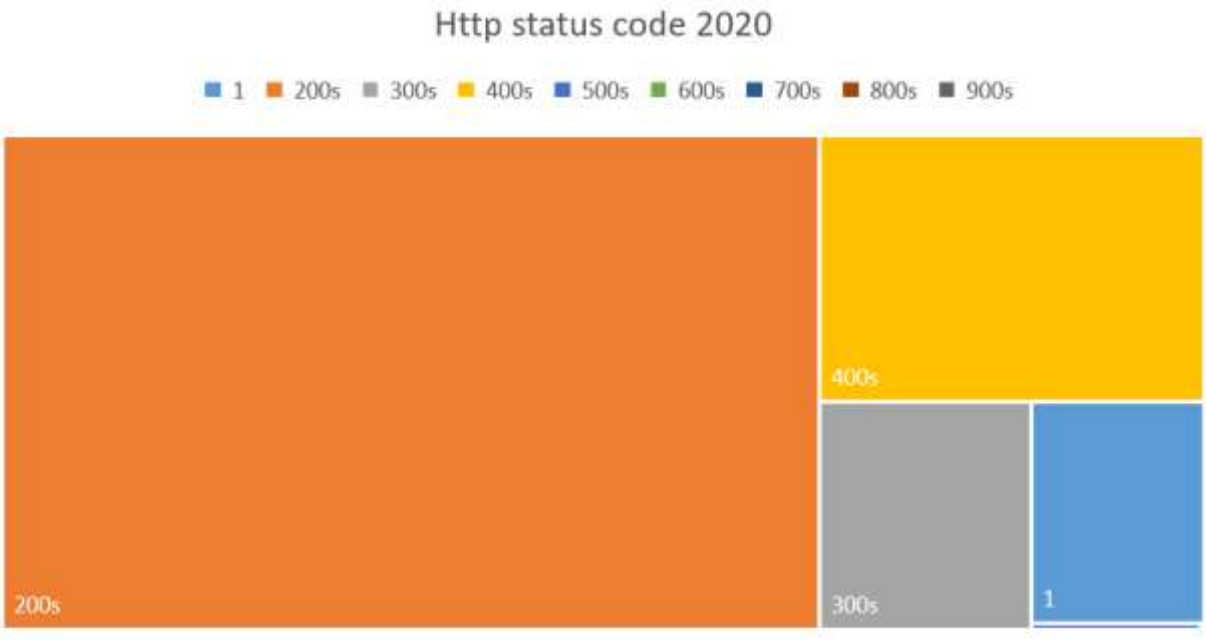
Najväčšou veľkosťou disponujú obrázkové a textové formáty. Obrázkové formáty (image) dosahujú najväčšie veľkosti, hneď za nimi sú textové súbory. Ďalšie miesta patria väčším súborom na stránkach typu pdf, word a podobne. Zvyšnú veľkosť zaberajú hlavne textové typy. V prípade, že by v zbere boli povolené multimediálne alebo zip formáty, do popredia by sa zaradili aj niektoré ďalšie, ako napríklad mp4, avi (Tabuľka 3).

text/html	111404250	4,6290
image/jpeg	57698178	5,9314
application/json	21900186	0,0949
text/dns	17506245	0,0011
image/png	15294280	1,1055
text/plain	14702722	0,0318
text/css	5240010	0,2663
application/javascript	5133076	0,2804
ostatné (6352 ďalších typov)	3620852	0,9269
text/xml	2718092	0,0091
image/svg+xml	2333936	0,1735
unknown	2128553	0,0205
image/gif	2127065	0,0660
text/javascript	1720769	0,1442
font/ttf	1640173	0,0925
application/pdf	1520602	1,5778

application/x-javascript	888199	0,0476
image/x-icon	843027	0,0074
application/vnd.ms-fontobject	785121	0,0539
application/x-font-woff	605871	0,2412
application/font-woff2	589879	0,1966
image/jpg	559708	0,0218
application/x-font-ttf	533104	0,1380
application/xml	453001	0,0026

Tabuľka 3 Počet formátov a ich zozbieraný objem v TeraBite

Ďalší pohľad na slovenský webový priestor poskytujú stavové kódy HTTP. Stavový kód HTTP, ktorý je súčasťou hlavičky odpovede servera na klientskú požiadavku upresňuje, ako bola odpoveď serverom spracovaná – či bola požiadavka vybavená kladne, záporne, alebo došlo k chybe. Najväčšia časť bola úspešná (200vky), alebo presmerovaná (300vky). Neúspešné odpovede (400vky) a serverové chyby (500vky) ukazujú na problémy so serverom alebo samostatným webom. Zaniknuté alebo chybné odkazy na internete častokrát vyjadrujú preklepy v názvoch samostatných súborov alebo adries URL. Weboví vývojári a vlastníci domén by mali používať nástroje na identifikáciu nefunkčných adries URL. Ostatných odpovedí bolo iba zanedbateľné množstvo (viď. diagram).



1	200s	300s	400s	500s	600s	700s	800s	900s
17 506 245	184 986 523	22 005 584	46 989 536	450 405	1	19	1	8 585

Diagram HTML status kódy