UNIVERZITNÁ KNIŽNICA V BRATISLAVE

CDA 2016

Formátové výzvy LTP

Zborník príspevkov z 1. medzinárodnej konferencie
o dlhodobej archivácii

univerzitná knižnica
v bratislave

Bratislava, 2016

UNIVERZITNÁ KNIŽNICA V BRATISLAVE

CDA 2016

Formátové výzvy LTP

Zborník príspevkov z 1. medzinárodnej konferencie
o dlhodobej archivácii

univerzitná knižnica
v bratislave

Bratislava, 2016

Zostavovateľka
Mgr. Lucia Kelemenová

Autori príspevkov
Ing. Milan Rakús
Mgr. Bibiána Žigová
Mgr. Jan Hutař, Ph.D.
PhDr. Ladislav Cubr
Bakk. techn. Peter Bubestinger
Bc. Andrej Bizík
Mgr. Jaroslav Kvasnica

Obálka a grafický návrh
DOLIS, s.r.o., Bratislava

CIP SR

CDA 2016 [online] : Formátové výzvy LTP : zborník príspevkov z 1. medzinárodnej konferencie o dlhodobej archivácii: Bratislava, 10. 11. 2016 / zost. Lucia Kelemenová ; obálka a graf. návrh DOLIS, s.r.o. – 1. vyd. – Bratislava : Univerzitná knižnica v Bratislave, 2016

LTP archívy. Centrálny dátový archív. Dlhodobé dôveryhodné digitálne úložisko. Formátová stratégia. Formáty súborov. Kontajnerové formáty súborov. Kodeky.

ISBN 978-80-89303-53-3
ISSN 2453-9406

Webový archívny formát WARC

Andrej Bizík, Univerzitná knižnica v Bratislave

Abstrakt
Za posledné roky sa pamäťové organizácie snažia nájsť najvhodnejší spôsob, ako sledovať zmeny a zhromažďovať webové stránky a webový obsah, ktorý sa mení každý deň. Je dôležité, aby formáty umožňovali v jednom súbore jednoducho a bezpečne uschovať veľmi veľký počet dátových objektov rôznych formátov a typov pre skladovanie, riadenie a výmenu dát. Preto bol vytvorený medzinárodný štandard ISO, opisujúci webový archívny formát WARC. Článok obsahuje krátku históriu, základné údaje a stručný popis súboru WARC. Zároveň sa zaoberá popisom fungovania webového archívu slovacikálneho obsahu, ktorý začala realizovať Univerzitná knižnica v Bratislave v roku 2015 ako národný projekt „Digitálne pramene – webharvesting a archivácia e-Born obsahu“. Pilotná prevádzka projektu skončila k 31.12.2015, od januára 2016 na ňu nadväzuje prvý rok udržateľnosti. Pre potreby realizácie projektu vzniklo v rámci Národnej agentúry ISSN oddelenie Depozit Digitálne Pramene (DDP), ktoré sa stalo súčasťou rutinnej prevádzky. Za toto obdobie sa uskutočnilo niekoľko archívnych zberov, uložených vo formáte WARC, ktorých počet a veľkosť sú vyhodnotené v závere.

Kľúčové slová: WARC formát, archivácia webu, webový archív

História

Súbor ARC (Arc File Format) začala interne používať organizácia Internet Archive na záznam postupností obsahu, so stručným popisom zozbieraných súborov. Archívny súbor ARC zhromažďuje údaje vo veľkých súhrnných súboroch pre jednoduché uchovanie v konvenčnom systéme súborov. Formát ARC riadi veľké množstvo malých súborov vo veľkých systémových súboroch. Bol navrhnutý tak, aby súhrnné objekty boli identifikované bez použitia súboru. Zároveň ukladá súbory načítané pomocou rôznych sieťových protokolov a po prvom zápise je integrita súboru nezávislá na následnom

obsahu. Pre načítanie objektov z konkrétneho archívneho súboru je dôležité udržať informácie o názvoch súborov, o ich veľkosti a o tom, ako sú navzájom previazané. Indexovanie zaznamenáva spracovávanie súborov a informácie o nich ukladá do databázy pre jednoduché vyhľadávanie, no nesnaží sa štandardizovať formát súborov.

Motivácia pre rozšírenie formátu ARC vzišla z diskusie a skúseností medzinárodného konzorcia Internet Preservation Consortium (IIPC), ktorého členmi sú Internet Archive, národné knižnice a od roku 2015 aj Univerzitná knižnica v Bratislave. Formát WARC je rozšírením formátu ARC, ktorý používa Internet Archive od roku 1996. Formát WARC sa líši od ARC tým, že ponúka nové možnosti, najmä pokiaľ ide o zaznamenávanie hlavičiek HTTP a metadát, pridelenie identifikátora pre každý obsiahnutý súbor, deduplikáciu obsahu a podobne. Formát WARC spravuje štruktúru a ukladá miliardy zdrojov zhromaždených z webu. Formáty WARC a ARC sú dostatočne odlišné, aby ich softvér jednoznačne rozpoznal a správne spracoval oba typy záznamov. Vzhľadom k veľkému množstvu už existujúcich archívnych dát vo formáte ARC je dôležité, aby sa pri prechode do formátu WARC neporušili záznamy.

Medzinárodný štandard ISO 28500: 2009 [1] zdokumentoval súbor vo formáte *WARC*. Norma popisuje formát súboru webového archívu WARC, ktorý ponúka konvencie pre zreťazenie viacerých dátových objektov do jedného dlhého súboru. Formát možno použiť na vytváranie aplikácií pre zber, správu, prístup a archiváciu obsahov stránok [2].

Ide o dostatočne flexibilný formát, ale ani ten nie je dokonalý. Preto sa na workshope konferencie IIPC GA v Stanforde v roku 2015 o nedostatkoch normy prijal návrh WARC Format 1.1 a vznikol projekt o špecifikovaní WARC 1.1 [3]. Projekt eviduje všetky požiadavky workshopu, ako aj navrhované zmeny z rôznych zdrojov. Vzniknutá predloha bude nakoniec odovzdaná medzinárodnej organizácii ISO na kontrolu a schválenie [4].

Typy WARC súborov

Záhlavie súboru WARC, ako aj každý jeho typ, obsahuje záznam o formáte a číslo verzie. V každom zázname sú pomenované polia. Každé uvedené pole sa skladá z názvu, po ktorom nasleduje dvojbodka („:“) a hodnota poľa. Názvy polí sa skladajú z veľkých a malých písmen. WARC súbor obsahuje za názvom poľa záznamové parametre, obsahujúce dôležité informácie (napr. identifikátor záznamu, čas vytvorenia,

dĺžka obsahu, typ obsahu). Každý záznam má uvedený typ v poli WARC-Type. Existuje osem typov záznamov, ktoré rozširujú formát WARC („warcinfo“, „response“, „resource“, „request“, „metadata“, „revisit“, „conversion“, a „continuation“). Polia sú v rôznom poradí hodnôt v kódovaní UTF-8 (8-bitový Unicode Transformation Format) [5]. Tieto typy sú dôležité hlavne pre softvér na čítanie súborov WARC.

„warcinfo“
Typ opisuje informácie o celom súbore, ako dátum vytvorenia, názov a podobne.

V prípade webového archívu obsahuje informácie o iniciátorovi zberu webu, ktorý vytvoril nasledujúce záznamy. Všetky tieto informácie sú voliteľné a pri každom súbore sa môžu líšiť. Základné kontaktné informácie obsahujú meno, názov organizácie, kontaktnú adresu autora, názov a verziu softvéru použitú na vytvorenie súboru, politiku pre rešpektovanie súboru robots.txt na webových stránkach, pracovný názov a IP adresu stroja, ktorý tento WARC vytvoril. Tieto informácie sa nachádzajú zvyčajne raz na začiatku súboru.

```
WARC/1.0
WARC-Type: warcinfo
WARC-Date: 2016-08-30T12:26:14Z
WARC-Filename:
WEB-20160830122614895-00000-12939~worker104.webdepozit.sk~20138.warc.gz
WARC-Record-ID: <urn:uuid:56ff4821-7a40-4757-b585-69d4ac388bf5>
Content-Type: application/warc-fields
Content-Length: 546

software: Heritrix/3.2.0 http://crawler.archive.org
ip: 10.109.33.230
hostname: worker104.webdepozit.sk
format: WARC File Format 1.0
operator: Andrej Bizik, Peter Hausleitner
publisher: Univerzitna kniznica v Bratislave
audience: webdepozit.sk users
isPartOf: basic
description: Archivacia stranok pre dalsie generacie
robots: obey //rešpektovať
http-header-user-agent: Mozilla/5.0 (compatible; heritrix/3.2.0 +
http://www.webdepozit.sk)
http-header-from: admin@webdepozit.sk
```

Obrázok 1: Ukážka typu záznamu warcinfo

„response“
Obsahuje prijatú odpoveď HTTP zo stránky, v prípade dostupnosti aj informáciu o sieťovom protokole. Medzi základné údaje patrí doménové meno, dátum, identifikátor

a kompletná odpoveď. Presný obsah je určený podľa typu záznamu a tiež podľa schémy URI (jednotného identifikátora zdroja) záznamu.

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: dns:consilium.europa.eu
WARC-Date: 2016-08-30T12:26:14Z
WARC-IP-Address: 10.109.22.31
WARC-Record-ID: <urn:uuid:f93ebb05-9c2c-4d47-8785-8071997c12f8>
Content-Type: text/dns
Content-Length: 61

20160830122614
consilium.europa.eu.    86352   IN  A   91.194.202.11
```

Obrázok 2: Záznam typu response

„resource“
Obsahuje zdroj súboru. Zdroj možno špecifikovať podľa viacerých schém, môže odkazovať na úložisko archívu alebo internetu, bez informácií o protokole. Napríklad schéma „http“ alebo „https“ obsahuje záznam o cieľovej adrese URI a schéma „dns“ obsahuje typ obsahu (Content-Type), ktorý obsahuje cieľovú adresu URI (WARC-Target-URI).

```
WARC/1.0
WARC-Type: resource
WARC-Target-URI: file://var/www/htdoc/images/logoc.jpg
WARC-Date: 2006-09-30T16:40:32Z
WARC-Record-ID: <urn:uuid:23200706-de3e-3c61-a131-g65d7fd80cc1>
Content-Type: image/jpeg
WARC-Payload-Digest: sha1:DBXHDRBXLF4OMUZ5DN4JJ2KFUAOB6VK8
WARC-Block-Digest: sha1:DBXHDRBXLF4OMUZ5DN4JJ2KFUAOB6VK8
Content-Length: 1662
```

Obrázok 3: Záznam typu resource

„request“
Záznam obsahuje podrobnosti o úplnej žiadosti v zmysle § 5 HTTP / 1.1 (RFC2616) [6], vrátane informácií o sieťovom protokole. Blok by mal obsahovať správu o žiadosti, teda požiadavku HTTP odoslanú cez sieť, vrátane hlavičky. Pole WARC-IP-adress býva použité na zaznamenanie sieťovej IP adresy. Záznam nešpecifikuje údaje o „https“, ako sú napríklad certifikáty.

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: http://consilium.europa.eu/robots.txt
WARC-Date: 2016-08-30T12:26:15Z
WARC-Concurrent-To: <urn:uuid:cc6acbff-7418-48bb-b46a-63c9dbc81383>
WARC-Record-ID: <urn:uuid:580acff0-683c-4c0a-b3e7-d8243459bb64>
Content-Type: application/http; msgtype=request
Content-Length: 246

GET /robots.txt HTTP/1.0
User-Agent: Mozilla/5.0 (compatible; heritrix/3.2.0 +http://www.webdepozit.sk)
From: admin@webdepozit.sk
Connection: close
Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8
Host: consilium.europa.eu
```

Obrázok 4: Záznam typu request

„metadata“

Polia „metadata“ vytvárajú obsah, s cieľom ďalej popísať zozbierané zdroje. Záznam môže odkazovať na iný záznam s informáciou pôvodného alebo transformovaného obsahu. Akýkoľvek počet metadát môže odkazovať na jeden konkrétny záznam. Formáty záznamov sa môžu líšiť a všetky polia sú voliteľné. Pole „via“ konkretizuje stránku, z ktorej bol obsah archivovaný. Čas zberu stránky v milisekundách obsahuje pole „fetchTimeMs“.

```
WARC/1.0
WARC-Type: metadata
WARC-Target-URI: http://consilium.europa.eu/robots.txt
WARC-Date: 2016-08-30T12:26:15Z
WARC-Concurrent-To: <urn:uuid:cc6acbff-7418-48bb-b46a-63c9dbc81383>
WARC-Record-ID: <urn:uuid:f00edaee-c6de-48df-a437-c2bc3552a63b>
Content-Type: application/warc-fields
Content-Length: 310

force-fetch:
via: http://consilium.europa.eu/
hopsFromSeed: P
fetchTimeMs: 186
charsetForLinkExtraction: UTF-8
outlink: http://www.consilium.europa.eu/robots.txt R Location:
outlink: http://consilium.europa.eu/favicon.ico I =INFERRED_MISC
outlink: http://www.consilium.europa.eu/robots.txt L a/@href
```

Obrázok 5: Záznam typu metadata

„revisit“

Nepovinný záznam, ktorý porovnáva zbieraný obsah s už archivovaným obsahom s cieľom nájsť rovnaký obsah, hlavne z dôvodu šetrenia pamäti. Účelom tohto typu záznamu je znížiť ukladanie duplicitného obsahu pri opakovanom načítaní rovnakého alebo málo zmeneného obsahu. Obsahuje odkaz na predchádzajúci, úplne alebo čiastočne duplicitný záznam z archívu.

```
WARC/1.0
WARC-Type: revisit
WARC-Target-URI: http://consilium.europa.eu/
WARC-Date: 2016-08-30T12:26:15Z
WARC-Payload-Digest: sha1:ENEXF2C2JV62CLQJ5CWYKX2FLFSWSVHG
WARC-IP-Address: 91.194.202.11
WARC-Profile: http://netpreserve.org/warc/1.0/revisit/identical-payload-digest
WARC-Truncated: length
WARC-Refers-To: <urn:uuid:d18647cb-532d-4612-9d3a-47c3fae9d7fc>
WARC-Refers-To-Target-URI: http://consilium.europa.eu/
WARC-Refers-To-Date: 2016-07-01T13:05:18Z
WARC-Record-ID: <urn:uuid:273de3fe-7969-4e01-94f0-bfced099ed40>
Content-Type: application/http; msgtype=response
Content-Length: 310

HTTP/1.1 301 Moved Permanently
Content-Type: text/html; charset=UTF-8
Location: http://www.consilium.europa.eu/
Server: Microsoft-IIS/8.5
X-Powered-By: ASP.NET
Access-Control-Allow-Origin: http://register.consilium.europa.eu
Date: Tue, 30 Aug 2016 12:26:15 GMT
Connection: close
Content-Length: 154
```

Obrázok 6: Záznam typu revisit

„conversion“

Záznamy „conversion“ obsahujú alternatívnu verziu obsahu, teda iný záznam, vytvorený ako výsledok archívneho procesu. Používajú sa pre zastavenie transformácie obsahu, ktorá udržuje životaschopnosť obsahu, v opačnom prípade obsah v pôvodnom formáte zmizne. Pôvodný obsah sa transformuje do rentabilnejšieho formátu, s cieľom udržať informácie použiteľné s existujúcimi nástrojmi a zároveň minimalizuje stratu pôvodných informácií. Záznamy môžu byť vytvorené tak, aby odkazovali na konkrétny zdroj záznamu, ktorý obsahuje transformovaný obsah. Každá transformácia by mala viesť ku kompletnému záznamu bez závislosti na pôvodnom zázname. Pre popis transformácie môžu byť použité metadáta.

```
WARC/1.0
WARC-Type: conversion
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2016-09-19T19:00:40Z
WARC-Record-ID: <urn:uuid:16da6da0-bcdc-49c3-927e-57494593dddd>
WARC-Refers-To: <urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>
WARC-Block-Digest: sha1:XQMRY75YY42ZWC6JAT6KNXKD37F7MOEK
Content-Type: image/neoimg
Content-Length: 934
```

Obrázok 7: Záznam typu conversion

„continuation“

Navzájom prepojené záznamy na zodpovedajúci obsah z iných súborov WARC. Vytvárajú logicky kompletný záznam pri prekročení limitu súboru WARC. Používajú sa na pokračovanie záznamov rozdelených do viacerých segmentov. Záznam obsahuje pôvodné ID, ktoré definuje začiatok záznamu a posledný „continuation“ záznam musí obsahovať pole „WARC-Segment-Total-Length“. Prvý súbor WARC bude obsahovať prvý segment – záznam typu response.

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Block-Digest: sha1:2ASS7ZUZY6ND6CCHXETFVJDENAWF7KQ2
WARC-Payload-Digest: sha1:CCHXETFVJD2MUZY6ND6SS7ZENMWF7KQ2
WARC-IP-Address: 207.241.233.58
WARC-Record-ID: <urn:uuid:39509228-ae2f-11b2-763a-aa4c6ec90bb0>
WARC-Segment-Number: 1
Content-Type: application/http;msgtype=response
Content-Length: 1600

HTTP/1.1 200 OK
Date: Tue, 19 Sep 2006 17:18:40 GMT
Server: Apache/2.0.54 (Ubuntu)
Last-Modified: Mon, 16 Jun 2003 22:28:51 GMT
ETag: "3e45-67e-2ed02ec0"
Accept-Ranges: bytes
Content-Length: 1662
Connection: close
Content-Type: image/jpeg
```

Obrázok 8: Warc typ response začiatok záznamu pri rozdelení

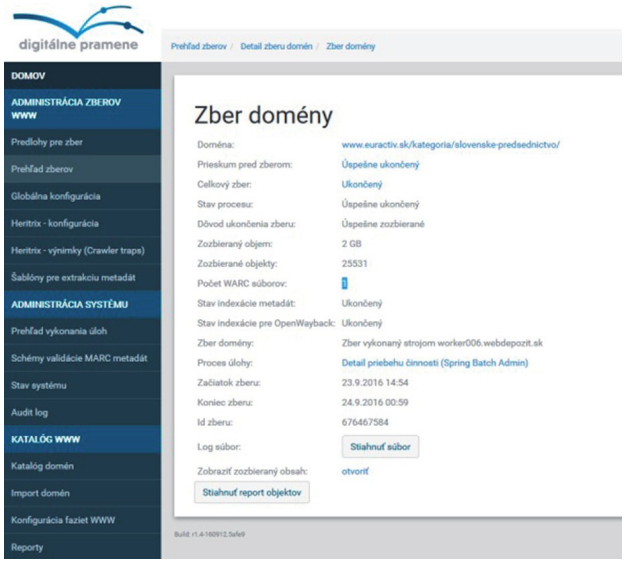
Budúci súbor WARC bude obsahovať pokračovanie záznamu, s poľami pre určenie začiatku segmentu (WARC-Segment-Origin-ID s pôvodným ID), číslo segmentu pre určenie poradia a v prípade posledného segmentu celkovú veľkosť záznamu (WARC-Segment-Total-Length).

```
WARC/1.0
WARC-Type: continuation
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Block-Digest: sha1:T7HXETFVA92MSS7ZENMFZY6ND6WF7KB7
WARC-Record-ID: <urn:uuid:70653950-a77f-b212-e434-7a7c6ec909ef>
WARC-Segment-Origin-ID: <urn:uuid:39509228-ae2f-11b2-763a-aa4c6ec90bb0>
WARC-Segment-Number: 2
WARC-Segment-Total-Length: 1902
WARC-Identified-Payload-Type: image/jpeg
Content-Length: 302
```

Obrázok 9: Záznam typu continuation

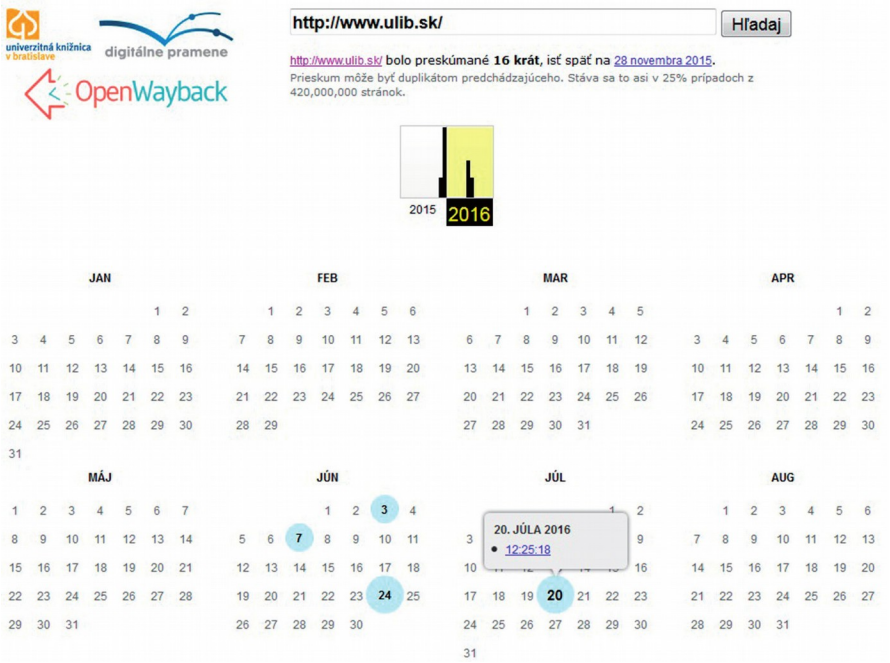
Slovenský webový archív – Webdepozit

Oddelenie DDP (Depozit Digitálnych Prameňov) využíva na zber a archiváciu vybraných webových stránok v podobe súborov WARC voľne dostupné softvérové riešenie Heritrix [7]. Pre účely archívu sú použité SATA disky v objeme 800TB, s predpokladom postačujúceho miesta na minimálne 5 rokov. Modul Web Curator Tool umožňuje cez webové užívateľské rozhranie konfigurovať parametre každého zberu. Heritrix následne zozbiera webový obsah z domén na základe pravidiel zadefinovaných v konfigurácii. Medzi jeho hlavné konfiguračné parametre pre každý webový zber patrí čas, veľkosť a počet dotazov na doménu. Webový obsah sa ukladá ako WARC súbor, pričom modul Deduplikátor kontroluje obsah zozbieraného obsahu a neukladá duplicitný obsah, ktorý už bol zozbieraný a uložený. Počas zberu sa zbierajú aj metadáta z webových stránok a ukladajú sa do katalógu s previazaním na WARC súbory. Na zobrazenie archívneho obsahu webu sa používa open source nástroj OpenWayback [8]. Zbierajú sa primárne html, php, css, js a image formáty. Pri zbere sa rešpektujú nastavenia v súbore robots.txt na strane servera. Pre každú webovú adresu URL zo zberu sa vytvoria vlastné WARC súbory. Jeden WARC má maximálnu veľkosť 2 GB a pri prekročení limitu sa pre danú doménu vytvorí viacero WARC súborov. Archivácia pracovných a log súborov sa komprimuje do súboru ZIP. Vznikne finálny WARC súbor s príponou súboru „warc.gz“.



Obrázok 10: Modul Web Curator Tool

Pri prezeraní archivovaného obsahu zadá používateľ do webového prostredia OpenWayBacku pôvodnú adresu URL stránky, kľúčové slovo alebo vyplní kritéria pre metadáta. Aplikácia zobrazí zoznam archivovaných snímok s časovým rozlíšením, kde používateľ vyberie jednu z nich pre zobrazenie archivovanej stránky.



Obrázok 11: Webové prostredie OpenWayback

Finálnym výsledkom archivácie je archivačný balík SIP (submission information package). SIP balíček je balík dát a metadát akceptovateľný pre LTP (Long Term Preservation) systém. Systém sám vyhľadá vytvorené WARC súbory, ktoré zabalí do SIP balíkov a uloží do cieľového adresára archivácie, kde si ich prevezme Centrálny dátový archív (CDA). CDA je nezávislý archív, spĺňajúci certifikáciu dôveryhodných digitálnych úložísk a informačnej bezpečnosti. Súčasťou SIP balíka je popisný súbor mets-md.xml, v ktorom sa nachádza zoznam všetkých priložených súborov a metadát. Pre archiváciu zozbieraného webového obsahu WARC súborov sa do mets-md.xml vkladajú popisy súborov do súborových tried, ktoré majú v identifikátore uložený názov domény v percentuálnom enkódovaní. Každá takáto skupina obsahuje len súbory patriace danej doméne, čo neskôr uľahčí vyberanie historických záznamov pre konkrétnu stránku. Nie je potrebné sťahovať celé SIP balíky, ale len konkrétny

WARC. Z hľadiska bezpečnosti archívu nemá prístup k vloženým balíkom žiadna iná inštitúcia.

```
crawl name: basic
crawl status: Finished
duration: 2h44m22s978ms

seeds crawled: 2
seeds uncrawled: 0

hosts visited: 50

URIs processed: 6077
URI successes: 5892
URI failures: 0
URI disregards: 185

novel URIs: 3333
duplicate-by-hash URIs: 2559

total crawled bytes: 4236491155 (3.9 GiB)
novel crawled bytes: 985031070 (939 MiB)
duplicate-by-hash crawled bytes: 3251460085 (3.0 GiB)

URIs/sec: 0.6
KB/sec: 419
```

Obrázok 12: Základný popis po skončení zberu

Štatistika súborov WARC vo Webdepozite

Štatistické údaje o počte Warc súborov vo Webdepozite UKB sa sledujú od januára 2016. K 3.10.2016 obsahuje archív Webdepozitu 1726 Warc súborov. Priemerná veľkosť jedného nekomprimovaného Warc je približne 348,5 MB. Ich celková komprimovaná veľkosť je 304 GB a nekomprimovaná 587 GB. Komprimácia teda ušetrí približne 48 % miesta z celkovej veľkosti WARC súborov. Viac informácií je uvedených na stránke www.webdepozit.sk.

Záver

V blízkej budúcnosti plánuje DDP dobudovať modul informačného systému pre extrakciu Open Graph objektov [9] z archívnych balíkov WARC. Projekt je zadaný na vypracovanie ako Diplomová práca na Slovenskej technickej Univerzite v Bratislave, odbor automatizácie a informatizácie v priemysle. Cieľom bude vytvorenie webovej aplikácie pre extrakciu Open Graph objektov z archívnych balíkov a ich transformáciu na databázové objekty archívu. V rámci práce sa bude riešiť vytvorenie autonómnej aplikácie, ktorá bude čítať archívne balíčky WARC a extrahovať z nich Open Graph objekty v zmysle Open Graph protokolu. V aplikácii sa vytvoria konfigurovateľné profily pre metadáta, ktoré budú mapovať Open Graph objekty na objekty v databáze. Aplikácia bude navrhnutá tak, aby ju bolo možné integrovať do existujúceho riešenia DDP ako samostatný modul. Metadáta budú prideľované podľa pravidiel RDA (upravených pre účely projektu Digitálne pramene – webharvesting a archivácia e-Born obsahu) a knihovníckeho formátu MARC 21. Metadáta tvoria hlavnú štruktúru, popis a následnú katalogizáciu pre bibliografické jednotky. Bohatý katalóg s presnou identifikáciou objektov umožní ďalej rozvíjať výskum v archíve Digitálnych prameňov, väčšiu dohľadateľnosť entít v aktuálnom, zmenenom alebo zaniknutom elektronickom obsahu.

Zoznam bibliografických odkazov

[1] ISO 28500: 2009 Information and documentation WARC file format

[2] Sigurðsson, Kristinn. The WARC Format 1.1 [online]. *Blogger*. 17 august 2015, [cit. 2016-09-26]. Dostupné na internete: https://kris-sigur.blogspot.sk/2015/08/the-warc-format-11.html

[3] Github: *IIPC WARC specifications* [online]. [cit. 2016-09-26]. Dostupné na internete: https://github.com/iipc/warc-specifications/

[4] Burner, Mike and Kahle, Brewster. *ARC File Format* [online]. Internet Archive [cit. 2016-09-27]. Dostupné na internete: http://www.archive.org/web/researcher/ArcFileFormat.php

[5] *HTML Unicode UTF-8 Reference* [online]. [cit. 2016-09-27]. Dostupné na internete: http://www.w3schools.com/charsets/ref_html_utf8.asp

[6] *HTTP/1.1 RFC2616* [online]. The Internet Society 1999 [cit. 2016-09-30]. Dostupné na internete: http://www.ietf.org/rfc/rfc2616.txt

[7] *Heritrix* [online]. Heritrix archival crawler project [cit. 2016-09-30]. Dostupné na internete: https://webarchive.jira.com/wiki/display/Heritrix

[8] *OpenWayback* [online]. IIPC [cit. 2016-09-28]. Dostupné na internete: http://netpreserve.org/openwayback

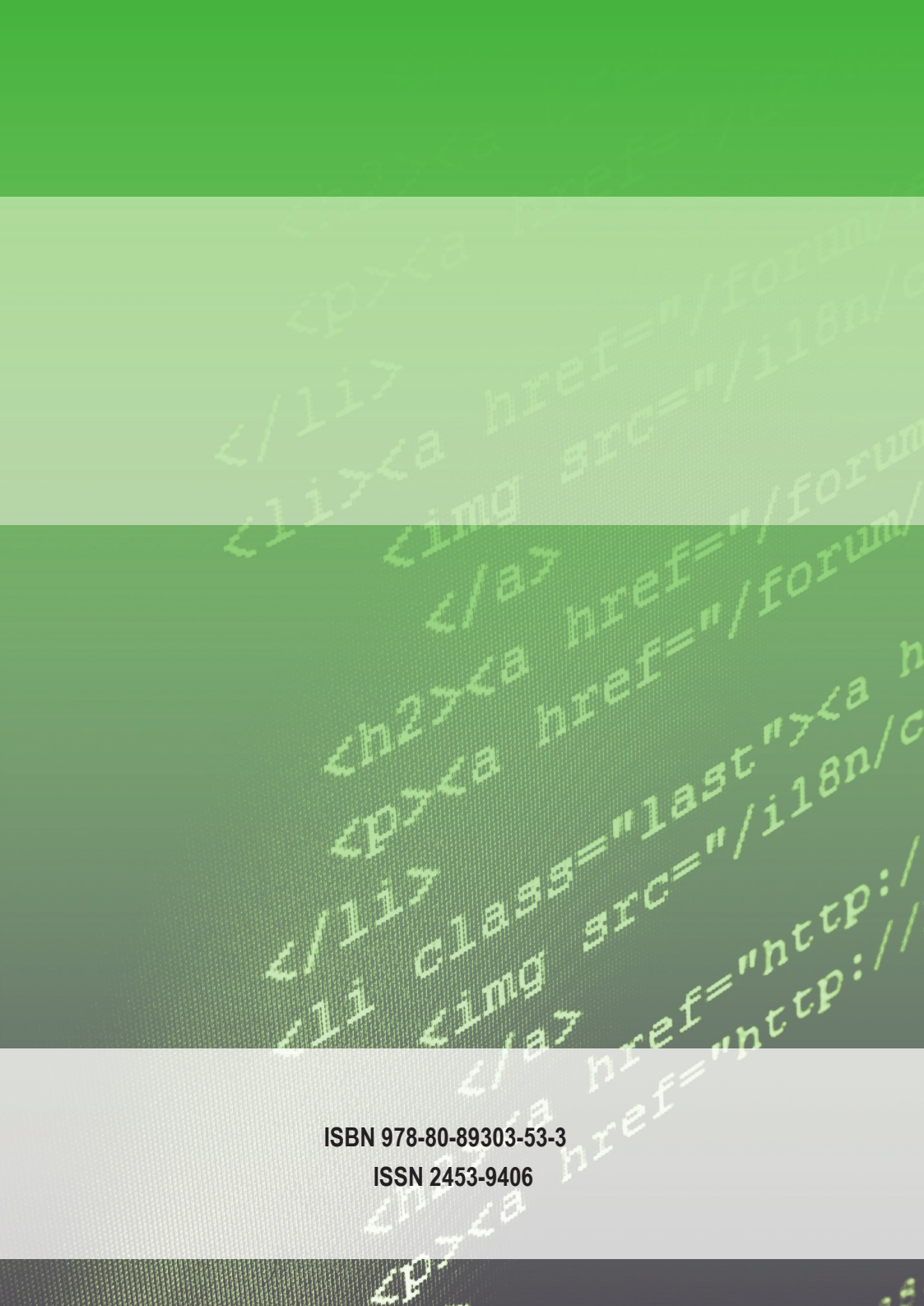
[9] *The Open Graph protocol* [online]. Open Web Foundation Agreement, Version 0.9 [cit. 30 septembra 2016]. Dostupné na internete: http://ogp.me/

CDA 2016

Formátové výzvy LTP

Vydala Univerzitná knižnica v Bratislave
Prvé vydanie. Počet strán 102.
Sadzba: DOLIS, s.r.o., Bratislava
Tlač: DOLIS, s.r.o., Bratislava

ISBN 978-80-89303-53-3
ISSN 2453-9406



ISBN 978-80-89303-53-3
ISSN 2453-9406