

Zpráva: Migrace na *formát WARC*

Práce na výzkumném záměru Koncepce rozvoje Národní knihovny České republiky jako výzkumné organizace na léta 2010 – 2012, oblast 3 – testování nástrojů a možností migrace na formát *WARC*.

Martin Prokop

martin.prokop@mail.muni.cz

1. Úvod

V této zprávě se budu zabývat možností migrace webového archivu *WebArchiv*. Konkrétně migrací uložených dat z formátu *ARC* do formátu *WARC*. Většina informací obsažených v této zprávě bude čerpána z mé bakalářské práce¹, kterou jsem odevzdával na jaře 2012 při studiu Fakulty Informatiky Masarykovy University. Bakalářská práce tedy tvoří součást mé zprávy. Ve zprávě budu zpravidla uvádět jednotlivé problémy a budu k nim předkládat odkazy na místa v mé práci, kde se jim podrobněji věnuji.

Ve zprávě uvedu pouze informace, které považuji pro zprávu nejdůležitější. V mé bakalářské práci je samozřejmě větší objem ucelených informací. Ty však nemusí být vždy pro zprávu relevantní. A naopak: některé informace, které jsou ve zprávě, v práci neuvádím.

V bakalářské práci jsem se zabýval specifickou částí migrace. Mým úkolem bylo najít vhodné nástroje pro migraci archivu a poté do nich integrovat nástroj *JHOVE2*², který slouží k identifikaci a validaci souborů. Cílem bylo provést při migraci jednotlivých archivů analýzu dat, která jsou v něm obsažená. Značná část bakalářské práce se tedy věnuje právě nástroji *JHOVE2*, jeho efektivitě a integraci do migračních nástrojů.

Nástroj *JHOVE2* však není pro migraci archivu zásadní. Důvodem proč ho využít při migraci je možnost objevit možné nekonzistentní soubory nebo nesprávné informace o archivovaných souborech. *ARC* archiv totiž nese informace o uložených souborech. Pokud jsou tyto informace nepřesné, může docházet k chybám při interpretaci dat. Užitím nástroje *JHOVE2* by se takové vadné informace povedlo najít a následně zamezit jejich přenesení do nových *WARC* souborů.

Zásadnější však je samotná migrace. Ve své zprávě se tedy budu věnovat migračnímu nástroji, který jsem zvolil jako nejvhodnější. Pokusím se popsat detaily práce s ním a uvést informace, které jsem do své bakalářské zprávy neuvedl.

1 Práci příkládám k této zprávě. Součástí bakalářské zprávy jsou také přílohy, které jsou stejně jako text práce dostupné na adrese <<https://github.com/MartinProkop/MojeWebarchivBakalarskaPrace>>.

2 Nástrojem *JHOVE2* se zabývám v bakalářské práci na s. 35.

2. Formát *WARC*

Formát *WARC*³ je popsán jako *ISO standard 28500:2009*⁴ již od roku 2009. Pro *WebArchiv* představuje možnost zkvalitnění a zefektivnění práce s archivovanými zdroji. V bakalářské práci se podrobněji zabývám specifiky tohoto formátu a uvádím i jeho výhody a nevýhody, se zřetelem na projekt *WebArchiv*⁵.

Většina nevýhod pramení z toho, že málokdo používá při archivaci webových zdrojů pouze *WARC* formát. Formát většina institucí prozatím testuje, ale neužívá ho v plné provozu. To je dáno hlavně tím, že není ještě plně podporován všemi standardními nástroji. Překážky k jeho plnému využívání jsou tedy spíše otázkou implementace funkcionality u jednotlivých nástrojů.

Výhody formátu jsou oproti tomu v jeho flexibilitě, univerzálnosti a přizpůsobivosti⁶. Je evidentní, že jeho využití, povede k zefektivnění práce s archivovanými webovými zdroji.

3 Formátem *ARC* i *WARC* se zabývám ve své bakalářské práci na s. 23 a stránkách následujících.

4 *ISO standard* za poplatek dostupný na adrese <<http://www.iso.org/iso/catalogue/catalogue de- tail.htm?csnumber=44717>>.

5 Informace jsou k dohledání v bakalářské práci na s. 26 a stranách následujících.

6 Jednou z nejlákavějších možností, které formát *WARC* nabízí je *deduplikace* souborů v archivu. *Deduplikací* se zabývám v bakalářské práci na s. 31.

3. Přejít na práci s *WARC* soubory

Přejít *WebArchivu* na práci výhradně s *WARC* archivy je komplexní záležitost. Samotná migrace je pouze jedna z dílčích úkonů. Sám si nejsem vědom všech dalších úkonů, které bude potřeba provést, aby *WebArchiv* pracoval pouze s *WARC* archivy. Namátkou se bude muset jednat například o implementaci podpory formátu *WARC* ve všech nástrojích, které *WebArchiv* používá.

Při zkoumání možnosti migrace jsem kontaktoval pracovníky *Britské knihovny*⁷, abych od nich zjistil, jak v *Britské knihovně* přistupovali k migraci a práci s *WARC* archivy. Domnívám se však, že mnohem větší množství informací, než jsem získal já, mají pracovníci *WebArchivu* z různých zpráv a seminářů *IIPC*.

Kdy by se mohlo pracovat v rámci projektu *WebArchiv* pracovat výhradně s *WARC* soubory mi není známo. Domnívám se však, že samotná migrace starých *ARC* archivů, by měla být provedena až v závěrečné fázi přechodu na práci s novým formátem.

⁷ Informace o *Britské knihovně* jsou v mé bakalářské práci na s. 33.

4. Migrační nástroje

Nástroje pro migraci *ARC* archivů jsou vcelku specifickým typem nástrojů, které převážně vyvíjí jednotlivé organizace, jenž se zabývají archivací webových zdrojů. Nástrojů je relativně malé množství. Problém je v tom, že každá organizace nástroj implementuje tak, aby byl „na míru“ a kompatibilní s ostatními nástroji, které používá.

Například při mé komunikaci s *Britskou knihovnou* jsem zjistil, že využili nástroje, které by byly pro *WebArchiv* zbytečné. To je dáno především tím, že pro archivaci webových zdrojů používají i jiné typy formátu archivu či jiné nástroje pro správu archivu.

Při výběru nástroje pro *WebArchiv* jde tedy hlavně o to, najít takový nástroj, který bude co nejsnadněji zařaditelný mezi *WebArchivem* používané nástroje.

5. Výběr vhodného nástroje

V bakalářské práci jsem, po konzultaci s Mgr. Václavem Roseckým, testoval tři různé nástroje na migraci *ARC* archivů. Jejich výběr byl relativně snadný, jelikož jich opravdu není moc. Tyto tři jsem vyhodnotil jako nejvhodnější. Některé další současné nástroje na migraci *ARC* souborů jsou odvozeny právě z těchto třech nástrojů.

Základním nástrojem, který jsem vybral, byl nástroj *warc-tools*⁸, který sice již není vyvíjen, přesto se jevil jako vhodný. Jeho hlavní výhodou bylo to, že byl implementován ve více jazycích a patřil mezi vůbec první nástroje, které tuto problematiku řešily.

Dalším nástrojem byl nástroj *warc-tools / hanzo*⁹, který je stále vyvíjen. Navíc přímo navazuje na předchozí nástroj. Vývojáři ho začali implementovat poté, co ukončili práci na výše zmíněném nástroji *warc-tools*. Jde navíc o nástroj, za jejímž vývojem stojí společnost *Hanzo Archives*¹⁰, která se zabývá problémem archivace dat na profesionální úrovni.

Poslední nástroj, který jsem testoval se jmenuje *warc-tools / kpk09*¹¹. Vznikl jako boční projekt nástroje *warc-tools / hanzo*. Tento výběr byl motivován především snahou prověřit, zda je nástroj, který je mutací oficiálního nástroje, v něčem výrazně odlišný od původního. Jak se později ukázalo, a jak konstatuji v bakalářské práci, jsou oba nástroje velmi podobné.

Jednotlivé nástroje jsem testoval¹² tak, že jsem si zvolil vzorová data – 9 *ARC* archivů z databáze *WebArchivu* a jeden mnou vytvořený¹³ – a zkoušel je jednotlivými nástroji převést na *WARC* archivy a poté jsem je porovnával. Dále jsem testoval i ostatní

8 Dostupný na adrese <<http://code.google.com/p/warc-tools/>>. V bakalářské práci se nástroji věnuji na s. 39.

9 Dostupný na adrese <<http://code.hanzoarchives.com/warc-tools/wiki/Home>>. V bakalářské práci se nástroji věnuji na s. 40.

10 Webové stránky organizace <<http://www.hanzoarchives.com/>>.

11 Dostupný na adrese <<https://bitbucket.org/kpk09/warc-tools/wiki/Home>>. V bakalářské práci se nástroji věnuji na s. 42.

12 Testování jsem prováděl na osobním počítači: Intel Core 2 Duo CPU T7100 @ 1.80GHz * 2, 2 GB RAM, Linux 3.0.0-19-generic 32-bit.

13 Všechny o velikosti 100 MB.

nástroje, které jsou součástí balíčků *warc-tools*. Jednalo se zejména o nástroje, které testují validitu *WARC* souborů (*warcvalid*), vypisují obsah *WARC* archivů (*warcdump*) či umí filtrovat jejich obsah (*warcfilter*).

Zjistil jsem, že původní nástroj *warc-tools* není kompatibilní s ostatními¹⁴. Podporuje totiž starší verzi *WARC* souborů – *WARC file format v 0.18*. Jelikož nástroj není dále vyvíjen, nemá smysl uvažovat o jeho využití. Využít by se dal jedině v případě, že by *WebArchiv* pokračoval v jeho vývoji.

Poté jsem testoval podobnost nástroje *warc-tools / hanzo* a *warc-tools /kpk09*. Jak jsem již zmínil, *warc-tools /kpk09* je z *warc-tools / hanzo* prvního odvozen. Jedná se o mutaci prvního nástroje¹⁵. Mým cílem bylo tedy porovnat oba nástroje¹⁶. Došel jsem k závěru, že nástroje na mém testovacím vzorku dat¹⁷, vytváří totožné *WARC* soubory.

Po konzultaci s Mgr. Václavem Roseckým jsem se tedy rozhodl pro další testování migrace využít pouze nástroj *warc-tools / hanzo*. Bočních projektů *warc-tools / hanzo* je v současné době již pět¹⁸. Jejich vzájemné porovnání by mohlo ukázat nejvhodnější nástroj pro *WebArchiv*. Rozdíly v nich jsou zas tak velké. Každá z mutací je vyvíjena pro specifické potřeby jednotlivých projektů, ale základní kód – samotný algoritmus migrace¹⁹ *ARC* archivu a jeho přepisu do *WARC* – je stejný.

¹⁴ Tuto skutečnost popisují ve své bakalářské práci na s. 39.

¹⁵ Nástroje jsou vyvíjeny pomocí distribuovaného verzovacího systému <<http://bitbucket.com>>. Druhý je tedy boční větev prvního.

¹⁶ Podrobnosti o porovnání jsou uvedeny v mé bakalářské práci na s. 43.

¹⁷ Použil jsem relativně malý vzorek testovacích dat.

¹⁸ Jednotlivé mutace *warc-tools / hanzo* jsou k nalezení na adrese <<http://code.hanzoarchives.com/warc-tools/descendants>>.

¹⁹ Algoritmem migrace se budu zabývat níže ve zprávě.

6. *Warc-tools* / *hanzo*

Poté co jsem zvolil tento nástroj, jsem se mohl již podrobně zabývat jeho specifiky a vlastnostmi. Jak jsem zmínil již v úvodu, hlavní část své bakalářské práce jsem věnoval integraci nástroje *JHOVE2* do *warc-tools* / *hanzo*. Proto jsem musel pracovat i se samotným migračním nástrojem.

V následujících kapitolách se pokusím popsat nástroj *warc-tools* / *hanzo*.

6.1 *Parametry*

Vybírám zde nejdůležitější informace o vlastnostech *warc-tools* / *hanzo*.

- Nástroj je implementován v programovacím jazyce *Python*. Nejsem schopný posoudit jestli je to výhoda nebo nevýhoda – záleží na ostatních nástrojích, které *WebArchiv* používá. Propojení nástroje s nástroji v jiných jazycích by mohlo být problematické.²⁰
- Pro použití nástroje je potřeba mít nainstalované tyto nástroje: *setuptools* (slouží pouze k instalaci *Python packages*), *unittest2*, *python 2.6* (testoval jsem nástroj i s *python 2.7* a podle všeho není s nástrojem ani s užitím této verze *Pythonu* žádný problém).
- Tvoří ho sada nástrojů pro práci s *WARC* soubory a samozřejmě nástroj na převod migraci *ARC* archivu.
- Při převodu *ARC* souborů nepřibálí do výsledného souboru data, je nutné převádět *ARC.GZ* archivy.
- Nástroj nerealizuje *deduplikaci* – tu nerealizuje žádný z nástrojů, které jsem testoval, ani žádný z těch, se kterými jsem se setkal. *Deduplikace* předpokládá více nástrojů a úpravu politiky archivace a interpretace archivovaných zdrojů²¹.

²⁰ Například při mé práci, kdy jsem ho propojoval s *JHOVE2*, bylo nutné najít způsob, jak propojit nástroj implementovaný v *Pythonu* s nástrojem implementovaným v programovacím jazyce *Java*. Detaily jsou k dohledání v bakalářské práci na s. 45 a stránkách následujících.

²¹ *Deduplikaci* se zabývám ve své bakalářské práci na s. 31.

- Nástroj nemá dokončenou dokumentaci. Stojí v ní, že nástroj vytváří „*crappy warc file*“. To je dáno tím, že samotný algoritmus migrace je stejný jako u původního nástroje *warc-tools*. Nástroj je totiž neustále vyvíjen a proto je jasné, že se v něm mohou ještě nalézt chyby, s čímž také vývojáři počítají. Sám jsem jednu chybu odhalil – informace uvedu později.

6.2 Další nástroje, které jsou součástí *warc-tools* / *hanzo*

Nejdůležitější nástroje, které obsahuje balíček *warc-tools* / *hanzo*, kromě nástroje *arc2warc*, který slouží k migraci, jsou nástroje *warcindex* a *warcvalid*.

Nástroj *warcindex* slouží k vytvoření *CDX indexu*²² pro *WARC* archivy. *CDX indexy* se dají použít k porovnání obsahu původního *ARC* archivu a výsledného *WARC* archivu vytvořeného migrací. Tím lze kontrolovat úspěšnost provedené migrace. Nástroj *warcindex* tedy doporučuji k vytváření *CDX indexů* *warc* souborů.

Dalším důležitým nástrojem je *warcvalid*. Slouží k validaci *WARC* archivů, jeví se jako vhodný pro kontrolu validity nových *WARC* archivů vytvořených při migraci. Díky nástroji jsem odhalil problémy s migračním nástrojem, který popisuji níže. Nástroj doporučuji k ověřování validity *WARC* archivů.

Další nástroje, které jsou v balíčku obsaženy, ale se kterými jsem přímo nepracoval jsou:

- *warcdump* – vypisuje informace o *WARC* archivu.
- *warcfilter* – vyhledává ve *WARC* souboru hlavičky *WARC záznamů*²³, podle zadaných parametrů.
- *warc2warc* – slouží pro převod *WARC záznamů*.

6.3 Použití nástroje

S balíčkem nástrojů se pracuje velmi jednoduše. Všechny nástroje jsou psány pro spouštění z příkazového řádku. Během práce jsem s ním měl pouze dva problémy.

²² Informace o *CDX indexu* lze dohledat v bakalářské práci na stranách 21, 26 a 50.

²³ *WARC záznamy* se zabývám podrobně ve své bakalářské práci na s. 27 a stranách následujících.

První problém je pouze záležitostí implementace. Jde o to, že neobsahuje nástroj pro hromadnou migraci souborů. Je tedy nutné spouštět nástroj pro každý vstupní *ARC* soubor. To však nepovažuji za zásadní problém, implementace takového nástroje by nebyla nikterak obtížná.

Zásadnější problém jsem objevil při použití migračního nástroje. Nástroj totiž umožňuje více druhů spouštění. Podle dokumentace se nástroj má spouštět z příkazové řádky tak, že jako parametr za přepínačem '*-o*' se zadá cesta k výslednému výstupnímu *WARC.GZ* archivu:

- `arc2warc -Z -o output.warc.gz input.arc.gz`

Tento způsob však u většiny mnou testovaných archivů vedl právě k nevalidním výsledným archivům. Dále jsem tedy zjistil, že jde spustit z příkazové řádky s užitím přesměrování na standardní výstup ('>') do cílového *WARC.GZ* archivu:

- `arc2warc -Z input.arc.gz > output.warc.gz`

Zde ovšem také vznikaly chybné *WARC* archivy. Tento problém jsem konzultoval s vývojáři softwaru, následně vývojáři nástroj opravili. Nyní doporučuji využívat první variantu spouštění nástroje.

6.4 Odhalení chyby v nástroji

Jak zmiňuji výše, odhalil jsem chybu v práci nástroje. Nástroj *warcvalid* hlásil pro některé z testovacích souborů chyby. Proto jsem chybu nahlásil vývojářům. Poté co jsem jim zaslal testovací data, konstatovali chybu nástroje *warcvalid* v práci s *GZipovanými* archivy.

Chyba, kterou validátor hlásil, byla buď:

- `('incorrect trailing newline', '\n')`

anebo:

- `('incorrect trailing newline', '\r')`

Popřípadě nástroj hlásil i obě tyto chyby zároveň.

Jednalo se o chybu v kódování nového řádku ve výsledných souborech. Tuto chybu opravili. V poslední verzi nástroje, kterou jsem následně testoval, již k této chybě nedochází.

6.5 Migrace

Pro testování migrace jsem použil vzorek dat z *WebArchivu*. Použil jsem soubory z různých sklizní – bylo potřeba otestovat archivy, ve kterých bude zastoupeno co největší množství různých souborů.

Při testování jsem se zaměřil na to, jestli budou výsledné *WARC* archivy validní a jestli bude jejich obsah totožný s původními *ARC* archivy. K tomu jsem použil výše zmíněné nástroje *warcvalid* a *warcindex*.

Dále jsem sledoval rychlost migrace, mým hlavním úkolem bylo zjistit rozdíl časové náročnosti při migraci s užitím nástroje *JHOVE2* a porovnat s časem migrace bez jeho použití.

V následující tabulce nabízím přehled rychlosti migrace. Rychlost ovlivňuje i obsah archivů – typy a velikosti jednotlivých souborů v archivu. Testy rychlosti jsem prováděl na svém osobním počítači, reálný převod na strojích *WebArchivu* by byl samozřejmě rychlejší.

Rychlost validace se pohybovala v závislosti na množství souborů v archivu v rozmezí 5 až 15 sekund.

Pokus	Počet souborů v archivu	Bez užití JHOVE2	S užitím JHOVE2
1	3 682	31 s	2 m 50 s
2	4 045	32 s	2 m 25 s
3	6 382	31 s	2 m 10 s
4	5 754	38 s	2 m 5 s

5	10 372	42 s	3 m 9 s
6	11 681	36 s	3 m 16 s
7	7 472	29 s	2 m 40 s
8	10 113	22 s	3 m 0 s
9	8 307	33 s	2 m 39 s
10	707	22 s	56 s
Průměr	X	32 s	2 m 31 s

Čas, který by trvala migrace celého archivu se dá jednoduše dopočítat podle celkového obsahu archivu, který má v současné době více než 47 TB dat. Výsledný čas bude ale pouze orientační.

6.6 Algoritmus migrace

Již výše jsem zmínil, že algoritmus převodu *ARC* archivu na *WARC* archiv je implementovaný přesně podle vzoru, jenž byl použit v původním nástroji *warc tools*. Navíc přímo v manuálu nástroje stojí, že nástroj vytváří „*crappy warc files*“.²⁴ Osobně se domnívá, že se jedná o jakési „alibi“ vývojářů – pro případ, že by se objevila chyba a bylo nutno ji opravit.²⁵

Před samotným použitím nástroje je nutné od vývojářů nástroje zjistit, jestli mají v plánu algoritmus upravovat či nikoliv. Pokud by se použil pro migraci *WebArchivu* algoritmus s chybou, mělo by to velmi nepříjemné následky, jejichž odstranění by stálo spoustu dalších prostředků.

Algoritmus funguje – velmi zjednodušeně – takto:

1. Nástroj načte vstupní *ARC.GZ* soubor (je nutné použít *GZipované* vstupní soubory,

²⁴ Popis nástroje *arc2warc* v dokumentaci: „*creates a crappy warc file from arc files on input*“. Dokumentace je dokladatelná v přílohách bakalářské práce jako příloha W na straně 65.

²⁵ Vyjádření vývojáře v poznámkách k *warc-tools* v dokumentaci: „*arc2warc uses the conversion rules from the earlier arc2warc.c ; as a starter for converting the headers ; I haven't profiled the code yet (and don't plan to until it falls over)*“. Dokumentace je dokladatelná jako příloha W bakalářské práce na straně 65.

jinak by nástroj do výsledného výstupního *WARC* souboru nepřibalil data²⁶).

2. Vytvoří prázdný výstupní *WARC.GZ* soubor. Dále do něj vloží první hlavičkový *WARC záznam*, který nese základní informace o archivu²⁷ – tyto informace získá upravením hlavičkového záznamu vstupního *ARC* souboru.
3. Zpracovává jednotlivé *ARC záznamy* v původním souboru – konvertuje je po jednom do *WARC záznamu*, ten zapíše do *WARC.GZ* výstupního souboru a připojí samotný soubor ze vstupního archivu.

Do jakéhokoliv kroku algoritmu se dá relativně snadno zasahovat²⁸. Domnívám se tedy, že pokud by to bylo nutné, mohou ho upravovat pracovníci *WebArchivu*.

6.7 Doporučení k implementaci a nasazení nástroje

Mé doporučení k případné implementaci zazněli vlastně již výše při popisu nástroje. Domnívám se, že pokud by se *WebArchiv* rozhodl nástroj použít na migraci webového archivu, měl by nejdříve kontaktovat vývojáře a požádat je o všechny informace o nedostatecích nástroje, popřípadě o doporučení k implementaci. Je i možné, že by vývojáři mutace *warc-tools* / *hanzo* pro potřeby *WebArchivu* doporučili jinou dostupnou mutaci.

Co se týká implementace změn v nástroji, domnívám se, že jeho úprava do takové podoby, jaká by byla nutná nasazení v provozu, je realizovatelná pracovníky *WebArchivu*.

Nástroj *warc-tools* / *hanzo* je podle mého názoru vhodný k realizaci migrace archivu. Musím ale konstatovat, že jsem rozhodně netestoval všechny dostupné nástroje.

26 I samotné *WARC záznamy* bez dat se dají jistě nějak využít, například k zjištění obsahu původního archivu.

27 Viz bakalářskou práci na s. 27 a stranách následujících.

28 Například jsem při implementaci analýzy obsahu archivu nástrojem JHOVE2 pomocí regulárního výrazu filtroval datastream z *ARC* souboru, vytvářel kopie přenášených souborů a ukládal je na disk. Mohl bych následně zasahovat do toho, které soubory budou do výstupního archivu přibaleny a které ne. Nebo bych mohl měnit jejich obsah. Také bych mohl měnit obsah *WARC záznamů*.

7. Zdroje informací

Zde nabízím přehled důležitých informačních zdrojů o migraci *ARC* archivů na *WARC* archivy. Další zdroje jsou vyhledatelné v mé bakalářské práci, domnívám se ale, že tyto jsou nejpodstatnější.

- IA Webteam JIRA: <<https://webarchive.jira.com/wiki/pages/viewpage.action?pageId=4865>>. Společnost se zabývá problematikou archivace webových zdrojů.
- Hanzo Archive: <<http://www.hanzoarchives.com/>>.
- Thomas Figg: <thomas.figg@hanzoarchives.com>. Vývojář *warc-tools* / *hanzo*.
- Roger Coram: <Roger.Coram@bl.uk>. Pracovník *Britské knihovny*.
- Bitbucket: <<http://code.hanzoarchives.com/warc-tools/descendants>>. Další mutace nástroje *warc-tools* / *hanzo*.
- Scape: <<http://wiki.opf-labs.org/display/SP/ISI2+ARC+to+WARC+migration>>
Potenciální zdroj informací o migraci. Další organizace, která se zabývá uchováváním webových zdrojů. Sám jsem s ní nebyl v kontaktu.

8. Závěr

Ve své zprávě jsem se věnoval migraci webového archivu *WebArchiv*. Představil jsem obsah své bakalářské práce, která je také součástí mé zprávy. Zpráva byla zacílena na představení nástroje *warc-tools* / *hanzo*, který považuji za vhodný pro realizaci migrace archivu *WebArchiv*.