Masarykova univerzita
Fakulta informatiky

Nástroje pro migraci webového archivu

Bakalářská práce

Martin Prokop

Brno, jaro 2012

Prohlášení

Prohlašuji, že tato bakalářská práce je mým původním autorským dílem, které jsem vypracoval samostatně. Všechny zdroje, prameny a literaturu, které jsem při vypracování používal nebo z nich čerpal, v práci řádně cituji s uvedením úplného odkazu na příslušný zdroj.

Vedoucí práce: Mgr. Václav Rosecký

Poděkování

Na tomto místě chci poděkovat Mgr. Václavovi Roseckému a Ing. Petrovi Žabičkovi za pomoc při řešení této bakalářské práce.

Shrnutí

Bakalářská práce se zabývá možností migrace webového archivu vytvořeného v rámci projektu WebArchiv. Představuji projekt a jeho vývoj, dále podávám přehled nástrojů a prací s ním souvisejících.

Po přehledu a sumarizaci informací o WebArchivu pokračuji přehledem archivačních formátů. Předvádím výhody jednotlivých formátů a ukazuji důvody, proč je potřeba provést migraci stávajícího archivu.

V poslední části práce se zabývám samotnými nástroji sloužícími k migraci. Cílem práce je představit výhody a nevýhody jednotlivých nástrojů a tím připravit podklad pro další studii možnosti migrace a její realizaci.

Abstract

Bakalářská práce se zabývá možností migrace webového archivu vytvořeného v rámci projektu WebArchiv. Představuji projekt a jeho vývoj, dále podávám přehled nástrojů a prací s ním souvisejících.

Po přehledu a sumarizaci informací o WebArchivu pokračuji přehledem archivačních formátů. Předvádím výhody jednotlivých formátů a ukazuji důvody, proč je potřeba provést migraci stávajícího archivu.

V poslední části práce se zabývám samotnými nástroji sloužícími k migraci. Cílem práce je představit výhody a nevýhody jednotlivých nástrojů a tím připravit podklad pro další studii možnosti migrace a její realizaci.

Klíčová slova

WebArchiv, warc, arc, migrace, archivace webu.

Obsah

Kapitola 1

Úvod

V dnešní době je internet jedním z hlavních zdrojů a nositelů informací. Každým dnem vzniká velké množství nových elektronických dokumentů. Takové dokumenty nemají charakter stálých informací, jejich obsah se mění prakticky neustále. Je typické, že staré verze dokumentů jejich autoři neuchovávají, a proto dochází ke ztrátě cenných informací. Z toho důvodu je důležité, aby se webové zdroje dlouhodobě uchovávaly a bylo možné je zpětně rekonstruovat. Tato práce je cílem projektu WebArchiv.

Řešitelé projektu WebArchiv se archivaci věnují již dvanáctým rokem. S postupujícím zdokonalováním technologií se rychle mění trendy a parametry, které musí sledovat a splňovat. Dá se říci, že každým dalším dnem vznikají další požadavky pro archivaci a dlouhodobé uchování webových zdrojů. Jedním ze základních problému je samotné uchovávání datového archivu. Každým rokem narůstá jeho obsah a tomu se musí přizpůsobit práce s archivovanými zdroji. Dále je potřeba zajistit, aby archivovaná data byla čitelná dlouhodobě a dalo se s nimi snadno manipulovat. V současné době se projevují limity archivování zdrojů pomocí formátu arc. Tento formát je pro archivaci sice vhodný, ale je již zastaralý a málo robustní.

Ukazuje se, že by bylo vhodné nahradit arc formátem warc. Formát warc je relativně nový a prozatím ne příliš používaný. V budoucnu se však pravděpodobně stane standardem pro uchovávání webových zdrojů. Vzniká tedy otázka možnosti přechodu WebArchivu na práci výhradně s tímto formátem. Před samotným přechodem je potřeba vyřešit spoustu problémů a všechny používané nástroje musí být na práci s novým formátem připraveny.

Předmětem této bakalářské práce je jedna z dílčích akcích vedoucích k přechodu WebArchivu na uchovávání dat ve formátu warc archivů.

Jedná se o migraci stávajícího archivu do nového formátu. Tato problematika je nová a specifická přímo pro projekt WebArchiv.

Ve své práci se pokusím nalézt nástroje, které by samotnou migraci umožnili. Představím možná řešení převodu a ukáži jejich výhody a nevýhody. Výstupem práce by pak měly být ukázky jednotlivých nástrojů určených pro migraci, nikoliv samotná migrace. Závěry práce tedy budou tvořit podklady řešitelům projektu WebArchiv, pro plánování migrace webového archivu.

Kapitola 2

WebArchiv

2.1 O WebArchivu

WebArchiv je projekt, jehož cílem je archivace Českého internetu. Jeho zřizovatelem je Národní knihovna ČR, která spolupracuje s Moravskou zemskou knihovnou a Ústavem výpočetní techniky Masarykovy univerzity. Projekt vznikl roku 2000 v rámci projektu Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet. Jeho cílem je uchovat české webové zdroje v rámci zachování českého kulturního dědictví. WebArchiv klade důraz na takové elektronické zdroje, které nejsou dostupné v tištěné podobě. Jejich archivace ve WebArchivu jeden z nejspolehlivějších způsobů jejich uchování do budoucna.

Zachování těchto dokumentů je podstatné hlavně pro zachycení a možnost hodnocení vývoje českého kulturního dědictví. Projekt si klade za cíl uchovávat dlouhodobě české webové stránky a umožnit jejich zpětné vyhledávání.

Archivace není důležitá jen pro případ ztráty dat, ale i pro předejití problému zastarání technologií. Jak víme multimediální technologie se v dnešní době velmi dynamicky vyvíjí a dokument, který byl před několika lety běžně zobrazitelný pro většinu uživatelů internetu, již dnes není podporovaný většinou počítačů. Jelikož WebArchiv zálohuje velké množství dokumentů v průběhu let musí řešit i tento aspekt archivace dat. Má tudíž výborný potenciál k zachování informační hodnoty dokumentů.

Databáze WebArchivu obsahuje :

- Digitální dokumenty volně dostupné prostřednictvím sítě internet
- Publikace odborného, uměleckého a zpravodajsko-publicistického

zaměření

- Periodika, monografie, konferenční příspěvky, výzkumné a jiné zprávy, a akademické práce

nadpis1 Digitální dokumenty [1] volně dostupné prostřednictvím sítě internet

nadpis2 Publikace odborného, uměleckého a zpravodajsko-publicistického zaměření

nadpis3 Periodika, monografie, konferenční příspěvky, výzkumné a jiné zprávy, a akademické práce

2.2 Vývoj projektu

Pokusím se předvést cíle a vývoj projektu v průběhu jeho řešení. V počátcích projektu bylo nejprve potřeba stanovit základní parametry projektu a dlouhodobé cíle. V průběhu řešení pak docházelo hlavně k formulaci nových cílů, výzkumu nových technologií a vznikla i potřeba pro adaptování se na nové trendy a technologie. [1]

WebArchiv pravidelně publikuje zprávy o své činnosti a dalších záměrech. Zprávy jsou samozřejmě vystavovány na webových stránkách projektu. Přehled vývoje v jednotlivých letech, který uvedu níže, čerpal vždy výhradně ze zpráv pro příslušný rok. [36]

2.2.1 Rok 2000

V počátcích projektu bylo důležité připravit podmínky, které by umožnili samotné zpracovávání české národní bibliografie a zajistit její dlouhodobé ukládání. Následně bylo potřeba vyřešit organizační otázky týkající se získávání nových dokumentů. Zde se jedná hlavně o legislativní rámec a politiku přijímání nových dokumentů. A poté upřesnění možností přístupu k elektronickým zdrojů v souvislosti s autorským právem.

1. dokument no

2.2.2 Rok 2001

Důležitou prací bylo mapování situace s archivováním internetových zdrojů v rámci jiných projektů a institucí, protože čerpání zkušeností od jiných řešitelů podobných projektů je výhodné a může ušetřit spoustu práce do budoucna. Začalo samotné shromažďování internetových zdrojů. Zásadní byla otázka archivace sklizených dat, řešitelé potřebovali, pro dlouhodobou archivaci a práci s daty, zvolit vhodný formát uchovávání metadat stažených souborů. Zvolili Dublin Core Metadata Element Set1, který byl lokalizován pro české zdroje. Probíhal vývoj nových nástrojů: Dublin Core Metadata Generator, Generátor URN, Kalkulátor MD5, Nedlib Harvester.

Rok 20012

Řešitelé upravili kritéria výběru nových webových zdrojů. Jak jsem zmínil výše, došlo k vývoji v oblasti legislativy. Opět vycházeli ze strategie zahraničních kolegů. Dále byla věnována pozornost sklizni elektronických seriálů.

pomlčka - spojovník – dlouhán —
italika *neco*
tucne **neco**
strojopis `neco - link a tak`
uvozovky? „noco“
co je ve vebarchivu: 2.1
sdgew
& % $ # _ { }

- neodsadí odstavec

rok sfg Další údaje budou popsány v sekci 2.2.1 na straně 5.
Na uvod 1 na straně 2.

Tabulka 2.1: Obsah WebArchivu

Sklizeň	**Počet souborů**	**Rozsah (MB)**
2001	3 017 058	106 520
2002	10 272 093	315 756
2004	32 161 396	1 058 305
2005	9 336 123	253 785
2006	70 741 016	3 465 016
2007	81 300 000	3 600 000
2008	78 203 483	3 900 000
2009	Není známo	9 300 000
2010	Není známo	16 800 000
2011	Není známo	7 800 000
Celkem	Není známo	46 599 382

Literatura

[1] Co je WebArchiv?. WEBARCHIV. *WebArchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz.

[2] Charakteristika Webarchivu. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/wainfo/.

[3] Dokumenty. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/dokumenty/.

[4] CELBOVÁ, Ludmila. *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet*. 2000. Dostupné z: http://www.webarchiv.cz/files/dokumenty/zpravy/zprava2000.pdf.

[5] CELBOVÁ, Ludmila. *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet*. 2002. Dostupné z: http://www.webarchiv.cz/files/dokumenty/zpravy/zprava2001/zprava2001.pdf.

[6] CELBOVÁ, Ludmila. *WebArchiv - vytvoření podmínek pro zpřístupnění českých webových zdrojů: knihovnické, legislativní a technické aspekty*. 2003. Dostupné z: http://www.webarchiv.cz/files/dokumenty/zpravy/zprava2002.pdf.

[7] STOKLASOVÁ, Bohdana. *Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřednictvím Jednotné informační brány*. 2004. Dostupné z: http://webarchiv.cz/files/dokumenty/zpravy/Zamer2004Zpravatextrev.doc.

[8] STOKLASOVÁ, Bohdana. *Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřed-*

nictvím Jednotné informační brány. 2005. Dostupné z: http://webarchiv.cz/files/dokumenty/zpravy/Zamer2005Zpravatext.doc.

[9] CELBOVÁ, Ludmila. *Ochrana a trvalé zpřístupnění webových zdrojů jako součásti národního kulturního dědictví*. 2006. Dostupné z: http://webarchiv.cz/files/dokumenty/zpravy/zprava-VaVō2006-final.rtf.

[10] WEBARCHIV. *Zpráva WebArchiv - obnova dat - 2007*. Brno, 2007. Dostupné z: https://docs.google.com/Doc?docid=0AbRV47jJIQggZG5qOHJtZF8yNmRjanM1dgüh

[11] CELBOVÁ, Ludmila. *Ochrana a trvalé zpřístupnění webových zdrojů jako součásti národního kulturního dědictví*. 2007. Dostupné z: http://webarchiv.cz/files/dokumenty/zpravy/zprava2007.pdf.

[12] COUFAL, Libor. *Ochrana a trvalé zpřístupnění webových zdrojů jako součásti národního kulturního dědictví*. 2008. Dostupné z: http://www.webarchiv.cz/files/dokumenty/zpravy/zprava2008.pdf.

[13] WEBARCHIV. *Zpráva WebArchiv - VISK - 2009*. Brno, 2009. Dostupné z: https://docs.google.com/Doc?docid=0AbRV47jJIQggZG5qOHJtZF80NngybjZ3aGYüh

[14] WEBARCHIV. *Zpráva WebArchiv - Věda a výzkum - 2010*. Brno, 2010. Dostupné z: https://docs.google.com/Doc?docid=0AbRV47jJIQggZG5qOHJtZF8xOGZ4YzNoamRy

[15] WEBARCHIV. *Zpráva WebArchiv - VISK - 2010*. Brno, 2010. Dostupné z: https://docs.google.com/Doc?docid=0AbRV47jJIQggZG5qOHJtZF8xN2Q2cnpxcmZm

[16] Celoplošné sklizně. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/celoplosne-sklizne/.

[17] CC info. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/ccinfo/.

[18] GRUBER, Lukáš. *Licence Creative Commons a perspektiva jejich zavedení do českého prostředí*. Ikaros [online]. 2008, roč. 12, č. 3 [cit.

21.03.2012]. Dostupný z: http://www.ikaros.cz/node/4612. URN-NBN:cz-ik4612. ISSN 1212-5075.

[19] JANSA, Petr. *Právní aspekty implementace projektu „Creative Commons“ v České republice*. Praha, 2008. Dostupné z: http://www.creativecommons.cz/wp-content/uploads/dp_petr_jansa_komplet_xmp.pdf. Diplomová práce. Univerzita Karlova.

[20] Partneři projektu. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/partneri-projektu/.

[21] Pro vydavatele. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/vydavatele/.

[22] Partneři. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/partneri.

[23] Doporučit zdroj. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/formular-url/.

[24] Kritéria. WEBARCHIV. *Webarchiv* [online]. 21.3.2012 [cit. 2012-03-21]. Dostupné z: http://www.webarchiv.cz/kriteria/.

[25] BARTOŠEK, Miroslav. *Systém pro správu procesu archivace webových informačních zdrojů*. Brno, 2009. Dostupné z: http://is.muni.cz/th/173018/fi_b/. Bakalářské práce. Masarykova univerzita.

[26] KUSALÍK, Filip. *Identifikace a omezení přístupu k „nevhodným“ stránkám ve webovém archivu*. Brno, 2009. Dostupné z: http://is.muni.cz/th/173018/fi_b/. Bakalářské práce. Masarykova univerzita.

[27] VLČEK, Ivan. *Rozpoznání a archivace českého webu mimo národní doménu*. Brno, 2008. Dostupné z: http://is.muni.cz/th/172585/fiōb/. Bakalářské práce. Masarykova univerzita.

[28] BELLA, Martin. *Implementace OAI-PMH pro český WebArchiv*. Brno, 2008. Dostupné z: http://is.muni.cz/th/98989/fiōb. Bakalářské práce. Masarykova univerzita.

[29] JELÍNKOVÁ, Lenka. *Bibliografický popis elektronických online zdrojů v zahraniční a domácí katalogizační praxi*. Praha, 2006. Dostupné z: http://www.webarchiv.cz/files/dokumenty/ostatni/DPjelinkova2006.pdf. Diplomová práce. Univerzita Karlova.

[30] MATĚJKA, Lukáš. *Zpřístupnění archivu českého webu*. Brno, 2006. Dostupné z: http://is.muni.cz/th/49968/fiōm. Diplomová práce. Masarykova univerzita.

[31] ŠKODOVÁ, Markéta. *Strategie archivace elektronických online zdrojů a politika jejich výběru do digitálního archivu (se zaměřením na český systém WebArchiv)*. Praha, 2005. Dostupné z: http://www.webarchiv.cz/files/dokumenty/zpravy/skodova.doc. Diplomová práce. Univerzita Karlova.

[32] Arc File Format Reference. INTERNET ARCHIVE. *Internet Archive* [online]. [cit. 2012-03-21]. Dostupné z: http://www.archive.org/web/researcher/ArcFileFormat.php.

[33] ARC to WARC (to ARC). IA Webteam Confluence. *IA Webteam Confluence* [online]. [cit. 2012-03-21]. Dostupné z: https://webarchive.jira.com/wiki/display/Heritrix/ARC+to+WARC+%28to+ARC%2

[34] ISO/DIS 28500. Information and documentation ? The WARC File Format. *Information and documentation – The WARC File Format* [online]. New Zealand: ISO, 2008. [cit. 2012-04-14]. Dostupné z: http://archive-access.sourceforge.net/warc/WARC_ISO_28500_final_draft%20v018%20Zentveld%20080

[35] PROKOP, Martin. *Mailová korespondence s pracovníky British Library*. 2012.

[36] THE REGENTS OF THE UNIVERSITY OF CALIFORNIA, Ithaka Harbors, Inc., and The Board of Trustees of Leland Stanford Junior University. *Next-Generation Characterization: An Update on the JHOVE2 Project*. The Regents of the University of

California, Ithaka Harbors, Inc., and The Board of Trustees of Leland Stanford Junior University., 2011. [cit. 2012-04-14] Dostupné z: http://bitbucket.org/jhove2/main/wiki/documents/JHOVE2-Users-Guide_20110222.pdf.

[37] Hanzo / warc-tools / overview. BITBUCKET. *BitBucket* [online]. 2012 [cit. 2012-04-14]. Dostupné z: http://code.hanzoarchives.com/warc-tools/overview.

[38] Kpk09 / warc-tools / overview. BITBUCKET. *BitBucket* [online]. 2012 [cit. 2012-04-14]. Dostupné z: https://bitbucket.org/kpk09/warc-tools/overview

Kapitola 3

Přílohy

A. Gramatika arc souboru

```
arc_file == <<version_block>><<rest_of_arc_file>>
version_block == See definition below
rest_of_arc_file == <<doc>>|<<doc>><<rest_of_arc_file>>
doc == <<nl>><<URL-record>><<nl>><<network_doc>>
URL-record == See definition below
network_doc == whatever the protocol returned
nl == Unix-newline-delimiter
sp == (ascii space) comma is inappropriate because it can be
in an URL.
```

B. Arc version block verze 2

Vysvětlivky

```
version-number == integer in ascii
reserved == string with no white space
origin-code == Name of gathering organization with no white
space
URL-record-definition == names of fields in URL records
```

Příklad

```
version-2-block == filedesc://<<path>><<sp>><<ip_address>>
<<sp>><<date>><<sp>>text/plain<<sp>>200<<sp>>
-<<sp>>-<<sp>>0<<sp>><<filename>><<sp>><<length>><<nl>>
2<<sp>><<reserved>><<sp>><<origin-code>><<nl>>
```

```
URL<<sp>>IP-address<<sp>>Archive-date<<sp>>Content-type
<<sp>>Result-code<<sp>>Checksum
<<sp>>Location<<sp>>Offset<<sp>>Filename
<<sp>>Archive-length<<nl>> <<nl>>
```

C. Arc url record verze 2

Vysvětlivky

```
url == ascii URL string (e.g., "http://www.alexa.com:80/")
ip_address == dotted-quad (eg 192.216.46.98 or 0.0.0.0)
archive-date == date archived
content-type == "no-type"|MIME type of data (e.g., "text/html")
length == ascii representation of size of network doc in bytes
date == YYYYMMDDhhmmss (Greenwich Mean Time)
result-code == result code or response code, (e.g. 200 or 302)
checksum == ascii representation of a checksum of the data.
The specifics of the checksum are implementation specific.
location == "-"|url of re-direct
offset == offset in bytes from beginning of file to beginning
of URL-record
filename == name of arc file
```

Příklad

```
URL-record-v2 == <<url>><<sp>>
<<ip-address>><<sp>>
<<archive-date>><<sp>>
<<content-type>><<sp>>
<<result-code>><<sp>>
<<checksum>><<sp>>
<<location>><<sp>>
<<offset>><<sp>>
<<filename>><<sp>>
<<length>><<nl>>
```

D. Příklad arc souboru verze 2

```
filedesc://IA-001102.arc 0.0.0.0 19960923142103 text/plain 200
- - 0
IA-001102.arc 122
2 0 Alexa Internet
URL IP-address Archive-date Content-type Result-code Checksum
Location Offset Filename Archive-length
http://www.dryswamp.edu:80/index.html 127.10.100.2 19961104142103
text/html 200 fac069150613fe55599cc7fa88aa089d - 209 IA-001102.arc
202
HTTP/1.0 200 Document follows
Date: Mon, 04 Nov 1996 14:21:06 GMT
Server: NCSA/1.4.1
Content-type: text/html Last-modified: Sat,10 Aug 1996 22:33:11
GMT
Content-length: 30
<<HTML>>
Hello World!!!
<</HTML>>
```

E. Arc soubor vytvořený pomocí crawleru Heritrix

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `IAH-20110905093545-00000-kovar-laptop-8090.arc.gz`.

F. Výstup z programu JHOVE pro přiložený arc soubor

Výstup pro soubor dostupný jako elektronická příloha bakalářské práce pod názvem `IAH-20110905093545-00000-kovar-laptop-8090.arc.gz`.

Příklad

```
Jhove (Rel. 1.6, 2011-01-04)
Date: 2011-09-05 15:14:13 CEST
RepresentationInformation:
```

```
./IAH-20110905093545-00000-kovar-laptop-8090.arc.gz
   ReportingModule: BYTESTREAM, Rel. 1.3 (2007-04-10)
   LastModified: 2011-09-05 12:40:13 CEST
   Size: 74389018
   Format: bytestream
   Status: Well-Formed and valid
   MIMEtype: application/octet-stream
```

G. Výstup z programu JHOVE2 pro přiložený arc soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `IAH-20110905093545-00000-kovar-laptop-8090.arc.gz_jhove2output.xml`.

H. Ukázka CDX indexu pro arc soubor

Celý CDX soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `index.cdx`.

Příklad

```
127.0.0.1/amk_new/
20110905093547
http://127.0.0.1/amk_new/
text/html
200
O5DWG2NMAMAAW7JSGZE3BQ64J7VWCYO3 - -
1056 IAH-20110905093545-00000-kovar-laptop-8090.arc.gz
```

I. Warc – metadata record

```
WARC/0.18
WARC-Type: metadata
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Date: 2006-09-19T17:20:24Z
WARC-Record-ID: <<urn:uuid:16da6da0-bcdc-49c3-927e-57494593b943>>
WARC-Concurrent-To: <<urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>>
```

```
Content-Type: application/warc-fields
WARC-Block-Digest: sha1:UZY6ND6CCHXETFVJD2MSS7ZENMWF7KQ2
Content-Length: 59

via: http://www.archive.org/
hopsFromSeed: E
fetchTimeMs: 565
```

J. Warc – request record

```
WARC/0.18
WARC-Type: request
WARC-Target-URI: http://www.archive.org/images/logoc.jpg
WARC-Warcinfo-ID: <<urn:uuid:d7ae5c10-e6b3-4d27-967d-34780c58ba39>>
WARC-Date: 2006-09-19T17:20:24Z
Content-Length: 236
WARC-Record-ID: <<urn:uuid:4885803b-eebd-4b27-a090-144450c11594>>
Content-Type: application/http;msgtype=request
WARC-Concurrent-To: <<urn:uuid:92283950-ef2f-4d72-b224-f54c6ec90bb0>>
GET /images/logoc.jpg HTTP/1.0
User-Agent: Mozilla/5.0 (compatible; heritrix/1.10.0)
From: stack@example.org
Connection: close
Referer: http://www.archive.org/
Host: www.archive.org Cookie:
   PHPSESSID=009d7bb11022f80605aa87e18224d824
```

K. Warc – warcinfo record

```
WARC/0.18
WARC-Type: warcinfo WARC-Date: 2006-09-19T17:20:14Z
WARC-Record-ID: <<urn:uuid:d7ae5c10-e6b3-4d27-967d-34780c58ba39>>
Content-Type: application/warc-fields
Content-Length: 381
software: Heritrix 1.12.0 http://crawler.archive.org
hostname: crawling017.archive.org
ip: 207.241.227.234
```

```
isPartOf: testcrawl-20050708
description: testcrawl with WARC output
operator: IA_Admin
http-header-user-agent:
Mozilla/5.0 (compatible; heritrix/1.4.0 +http://crawler.archive.org)
format: WARC file version 0.18
conformsTo:
http://www.archive.org/documents/WarcFileFormat-0.18.html
```

L. Warc soubor vytvořený z přiloženého arc souboru pomocí nástroje WARC-TOOLS

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `WARCTOOLS_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz`.

M. Warc soubor vytvořený z přiloženého arc souboru pomocí nástroje WARC-TOOLS / hanzo

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `HANZO_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz`.

N. Warc soubor vytvořený z přiloženého arc souboru pomocí nástroje WARC-TOOLS / kpk09

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `KPK09_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz`.

O. Výstup z programu JHOVE2 pro přiložený warc soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem
`WARCTOOLS_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz_jhove2output.xml`

P. Popis standardu WARC

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `WARC_Guidelines_v1.pdf`.

Q. Mailová korespondence s pracovníky British library

Soubory jsou dostupné jako elektronická příloha bakalářské práce. Pod názvem `british_library_mail1` jsou informace, které dostal Ing. Libor Coufal. Pod názvem `british_library_mail2` jsou informace, které psali přímo mě. Pod názvem `british_library_mail3` jsou informace, které jsem od pracovníků požadoval ohledně deduplikace souborů.

R. WARCTOOLS – Functional Requirements Specification v8

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warc-tools_phase_III_frs_v8.pdf`.

S. WARCTOOLS – Functional Requirements Specification

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warc_tools_frs.pdf`.

T. WARCTOOLS – Non-Functional Requirements

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warc_tools_nfr.pdf`.

U. WARCTOOLS – Software Requirements Specification

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warc_tools_srs.pdf`.

V. WARCTOOLS – výstup z nástroje warcdump pro warc soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `WARCTOOLS_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz_warcdump`.

W. WARCTOOLS / hanzo – README soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warctools_hanzo_README`.

X. WARCTOOLS / hanzo – výstup z nástroje warcdump pro warc soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `HANZO_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz_warcdump`.

Y. WARCTOOLS / hanzo – výstup z programu JHOVE2 pro přiložený warc soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `HANZO_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz_jhove2output.xml`.

Z. WARCTOOLS / kpk09 – README soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warctools_kpk09_README`.

AA. WARCTOOLS / kpk09 – výstup z nástroje warcdump pro warc soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `KPK09_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz_warcdump`.

AB. WARCTOOLS / kpk09 – výstup z programu JHOVE2 pro přiložený warc soubor

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem
`KPK09_IAH-20110905093545-00000-kovar-laptop-8090.warc.gz_jhove2output.xml`.

AC. WARCTOOLS – archiv obsahující nástroj

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warc-tools.tar.gz`.

AD. WARCTOOLS / hanzo – archiv obsahující nástroj

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warc-tools_hanzo.tar.gz`.

AE. WARCTOOLS / kpk09 – archiv obsahující nástroj

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `warc-tools_kpk09.tar.gz`.

AF. JHOVE2 – uživatelská příručka nástroje

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `JHOVE2-Users-Guide_20110222.pdf`.

AG. JHOVE2 – informace o aktualizace nástroje

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `NDIIPP-2010-JHOVE2.pdf`.

AH. JHOVE2 – archiv obsahující nástroj

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `jhove2-2.0.0.tar.gz`.

AI. Jpype – archiv obsahující nástroj

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `JPype-0.5.4.2.zip`.

AJ. Warc – ISO specifikace

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem
`WARC_ISO_28500_final_draft v018 Zentveld 08061.doc`. Soubor je také dostupný online na adrese `http://archive-access.sourceforge.net/warc/`.

AK. JHOVE2 – vlastní implementace nástroje

Celý soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `RunFromARC2WARC.java`. V rámci projektu nástroje JHOVE2 musí být součásti package org.jhove2.app. Zde ukazuji pouze metodu třídy (runJHOVE2), která obstarává přímo spuštění.

```
/**
* Metoda dostane na vstup seznam souborů, které předá ke zpracování
* aplikaci JHOVE2 a nakonec na vypíše do zadaného souboru výstup
z JHOVE2.
* @param contents Pole cest k souborům
* @param path Cesta k souboru, do kterého se mají vypsat informace
* @throws IOException
* @throws JHOVE2Exception
*/
public void runJHOVE2(String contents[], String path) throws
IOException, JHOVE2Exception
{
   OutputStream output = new FileOutputStream("/dev/null");
   PrintStream nullOut = new PrintStream(output);
```

```
   System.setErr(nullOut);
   System.setOut(nullOut);
   String[] data = "-o", path, "-d", "XML";
   String[] seznam = arrayMerge(data, contents);

JHOVE2CommandLine.main(seznam);
}
```

AL. WARCTOOLS / hanzo – vlastní implementace nástroje

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `arc2warc_jhove2.py`. Aby správně fungoval, je potřeba mít nainstalován nástroj Jpype (`příloha AI 3`) a umístit soubor do kořenové složky WARCTOOLS / hanzo (`příloha AD 3`).

AM. Warc – gramatika

```
Warc-file = 1*warc-record
warc-record = header CRLF
block CRLF CRLF
header =version warc-fields
version ="WARC/0.18" CRLF
warc-fields =*named- field CRLF
block = * OCTET
```

AN. Warc – seznam definovaných polí

```
WARC-Type, WARC-Record-ID, WARC-Date, Content-Length, Content-Type,
WARC-Concurrent-To, WARC-Block-Digest, WARC-Payload-Digest,
WARC-IP-Address, WARC-Refers-To, WARC-Target-URI, WARC-Truncated,
WARC-Warcinfo-ID, WARC-Filename, WARC-Profile, WARC-Identified-Payload-Type,
WARC-Segment-Origin-ID, WARC-Segment-Number, WARC-Segment-Total-Length
```

AO. Warc – Continuation record

První warc záznam

```
WARC/0.18
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
```

Záznam, který na něj navazuje

```
WARC/0.18
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
<<last 302 bytes of image/jpeg binary data here>>
```