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 za pomoc při řešení této bakalářské práce.

Shrnutí

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

Po přehledu a sumarizaci informací o WebArchivu pokračuje přehledem archivačních formátů. Předvádí výhody a nevýhody jednotlivých formátů. Zdůvodňuje, proč je potřeba provést migraci stávajícího archivu.

Poslední část práce se zabývá konkrétními nástroji sloužícími k migraci archivů. Cílem práce je představit výhody a nevýhody jednotlivých nástrojů a připravit tak podklad pro další studii možnosti migrace webového archivu a její realizaci.

Abstract

Klíčová slova

WebArchiv, warc, arc, migrace, archivace webu.

Obsah

Kapitola 1

Úvod

V dnešní době je internet jedním z hlavních zdrojů 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 umožnily. 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. [2] 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. [3]

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 [2]:

- 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

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.

WebArchiv pravidelně publikuje zprávy o své činnosti a dalších záměrech. Zprávy jsou samozřejmě vystavovány na webových stránkách projektu [4]. Přehled vývoje v jednotlivých letech, který uvedu níže, čerpal vždy výhradně ze zpráv pro příslušný rok.

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. [5]

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 staže-

ných souborů. Zvolili `Dublin Core Metadata Element Set`[1], který byl lokalizován pro české zdroje. Probíhal vývoj nových nástrojů: `Dublin Core Metadata Generator`, `Generátor URN`, `Kalkulátor MD5`, `Nedlib Harvester`.

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. [6]

2.2.3 Rok 2002

Činnost řešitelů byla především zaměřena na vývoj používaného softwaru. Za cíl měli vytvoření vyhledávací struktury a tím zpřístupnění zálohovaných webových zdrojů. Řešitelé neustále mapovali situaci v legislativě týkající se archivace zdrojů a v souvislosti tím došlo k vytvoření vzorových smluv s majiteli webových zdrojů. Pro nedostatek pracovních a finančních kapacit se odstoupilo od ukládání metadat ve formátu `Dublin Core` a řešitelský tým se rozhodl pro využití formátu `UNIMARC`[2] [7].

2.2.4 Rok 2003

Zpráva pro tento rok není zpracována, jelikož na ní řešitelé doposud nedostali grant.[3]

2.2.5 Rok 2004

Projekt byl řešen v rámci projektu s názvem Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřednictvím Jednotné informační brány.

Řešitelé se zabývali analýzou možností dalšího vývoje softwaru pro tvorbu metadat, jednoznačné identifikace dokumentů, stahování webových zdrojů a jejich ukládání. Dále možnostmi trvalého a efektivního

1. Více informací o `Dublin Core Metadata Element Set` dostupné online na adrese `http://dublincore.org/documents/dces/`.
2. Podrobné informace o formátu `UNIMARC` dostupné online na adrese `http://www.loc.gov/marc/unimarctomarc21.html`.
3. Osobní sdělení: mgr. Václav Rosecký – vedoucí práce.

přístupu k uloženým datům. Bylo potřeba monitorovat vývoj technologií – zejména standardů pro metadata a formáty. Byly budovány tématické brány a konspekt. WebArchiv užíval nástroje jako `NEDLIB`, `Heritrix`. [8]

2.2.6 Rok 2005

Projekt byl i nadále řešen v rámci stejného projektu jako v roce 2004 a pokračoval v práci ve stejném okruhu témat jako v předešlém roce. Důraz byl kladen na analýzy přístupnosti dokumentů v archivu. Databáze WebArchivu byla převáděna do formátu `arc`. V tomto formátu jsou data uložená až do dnes. Právě `formát` arc a jeho převod do formátu `warc` je předmětem mé práce. [9]

2.2.7 Rok 2006

Řešitelé projektu pracovali současně na projektu `CULTURE 2000`[4] `(Web Cultural Heritage)`. Z toho důvodu provedli obsáhlou analýzu obsahu archivu. Dále proběhla analýza využívaného softwaru: `NutchWAX`, `WayBack`, `WAXToolbar`, `WERA`, `ARCRetriever`, `WebCurator`, `Heritrix`, `DeDuplicator`. [10]

2.2.8 Rok 2007

Řešitelské upravili webové stránky projektu. Důraz byl kladen hlavně na uživatelskou přívětivost. Z webu je od té doby přímý přístup k nástroji `WERA`, který umožňuje fulltextové vyhledávání ve veřejné části databáze WebArchivu. Dále je plně zpřístupněn nástroj `Wayback`, který umožňuje vyhledá ve zdrojích podle URL adresy. Vyhledávání je napojeno na možnost vyhledávání v databázi `Internet Archive`[5].

Dále byl vyvinut nástroj pro vyhledávání podle metodiky konspektu. Tým se zabýval možností zlepšování tematických sklizní. Také se pracovalo na lokalizaci užívaných nástrojů. Prozkoumávala se i možnost průběžné analýzy sklizně.

---

4. Podrobné informace o projektu `CULTURE 2000` dostupné online na adrese `http://www.webarchiv.cz/culture-2000/`.
5. Webové stránky projektu Internet Archive dostupné online na adrese `http://www.archive.org/`.

Velký význam měla analýza možnosti sklízení bohemikálních zdrojů mimo doménu CZ. Na toto téma byla zpracována i bakalářská práce s názvem Rozpoznání a archivace českého webu mimo národní doménu, kterou zpracoval Ivan Vlček. [11] V tomto roce došlo ke změně legislativy týkající se uchovávání elektronických informačních zdrojů. Tento problém úzce souvisí s licencí Creative Commons[6].

Došlo také k havárii diskového pole, která způsobila ztrátu dat v archivu. Ze zhruba 4,4 TB dat zůstalo nepoškozeno 3,8 TB dat. Ztracená data zastupovala archívy ze všech proběhlých sklizní. Některé z poškozených souborů se povedlo obnovit ze záloh či doplnit z databáze Internet Archive. [12] [13]

2.2.9 Rok 2008

Pokračovalo se v analýze možností konspektu a tematických sklizní. Dále probíhala lokalizace používaných nástrojů. Jako v předchozím roce se prozkoumávala možnost sklízení bohemikálních zdrojů mimo doménu CZ – stále v rámci bakalářské práce Ivana Vlčka. [11] Nadále se sleduje možnost analýzy průběžných sklizní.

Významně se projevila potřeba řešit otázku ochrany a trvalého zpřístupnění uložených dokumentů. Opět bylo potřeba řešit jak hledisko legislativní tak technologické. Z technologického hlediska je hlavně potřeba navýšit kapacitu datových úložišť a ochranu proti ztrátě dat. V minulosti totiž několikrát došlo ke ztrátě dat v důsledku poruchy datového úložiště. [14]

2.2.10 Rok 2009

Řešitelé měli na starost hostování a správu serverů a diskového pole v ÚVT MU a Centrálním depozitáři NK v Hostivaři. To obnášelo pozorování a správu softwaru a instalaci nových serverů VMWARE.

Pro zabezpečení před ztrátami dat, které by nastaly při havárii úložišť probíhala záloha severů a revize politiky zálohování. S narůstající velikostí archivu se jeho zpráva stává extrémně obtížnou.

V průběhu roku bylo provedeno větší množství sklizní. Devět výběrových sklizní, dvě celoplošné sklizně domény CZ a dvě celoplošné

6. Podrobné informace o licenci Creative Commons uvedu v další části práce 2.4.1.

sklizně bohemikálních zdrojů mimo doménu CZ. Oproti předešlému roku bylo sklizeno více než dvojnásobné množství dat. Tento nárůst se dá vysvětlit výše zmíněnou instalací nových výkoných serverů. [15]

2.2.11 Rok 2010

Pokračovalo spravování serverů a diskového pole v ÚVT MU a Centrálním depozitáři NK v Hostivaři. Byly Doinstalovány dva nové VMWARE servery. Pokračovalo pravidelné zálohováni archivu. Probíhala revize přístupové politiky k serverům. Došlo k několika závažných výpadků způsobených hardwarovými chybami.

Z hlediska výzkumu se pokračovalo ve vývoji konspektu, lokalizace nových verzí softwaru a analýzy možností sklizní bohemikálních zdrojů mimo doménu CZ. Dále bylo potřeba navýšit robustnost infrastruktury. Řešitelé se snažili zlepšit dohled nad probíhajícími sklizněmi. [16]

V tomto roce bylo provedeno devět výběrových sklizní. A samozřejmě proběhla i celoplošná sklizeň nad doménou CZ. Při sklizních byl nasazen nástroj DeDuplicator. Podrobnosti o nástroji zmiňuji v části práce věnované nástrojům. [17]

2.2.12 Rok 2011

V době psaní této bakalářské práce nebyla zpráva dostupná.

2.3 Současný obsah databáze

Zde nabízím přehled množství stažených webových zdrojů v celoplošných sklizních. Kompletní přehledy jsou dostupné na webových stránkách projektu [18]. Mimo celoplošné sklizně, určené pro sběr dat na doméně CZ tvoří část databáze i tématické sklizně a konspekt. Celkově obsahuje databáze cca 47 TB dat.

Údaje v tabulce odpovídají zprávám, které WebArchiv pravidelně publikuje. Čísla však nemusí být úplně přesná[7].

7. například z důvodu havárie v roce 2007

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

2.4 Další informace o projektu

Zde nabízím přehled dalších důležitých tématických okruhů, které se týkají projektu.

2.4.1 `Creative Commons`

Pro zvyšování obsahu WebArchivu je zásadní licence `Creative Commons`[8]. Jedná se o mezinárodní licenční standard, pomocí kterého může autor elektronického díla poskytnout své dílo k užití jiným osobám. [19] Pokud autor webový zdroj neposkytne pod licencí `Creative Commons` musí s ním WebArchiv, pokud chce webový zdroj archivovat a umožnit vyhledávat, uzavřít smlouvu.
`Creative Commons` je podřízeno autorskému zákonu, je to jednoduchý způsob jak může autor elektronického díla poskytnout toto dílo dalším osobám. Výhodou `Creative Commons` je, že autor nemusí podepisovat smlouvu s každým uživatelem jeho díla. Autor jednoduše označí svůj dokument značkou `Creative Commons`, a následně uživatel daného dokumentu vstupuje automaticky do smlouvy. Majitel díla navíc přesně

8. Webové stránky projektu `Creative Commons` dostupné na adrese `http://creativecommons.org/`.

určí jak se s jeho dílem smí nakládat. [20] Sledování podmínek užití licence `Creative Commons` je jedna z velmi důležitých úkolů WebArchivu. [21] K velké změně došlo v roce 2009, kdy se Česká republika přidala k zemím s lokalizovanou verzí licence[9]. [19] WebArchiv na svých stránkách přímo poskytuje návod[10] jak může majitel webového zdroje přidat svůj web pod licenci `Creative Commons` a samozřejmě podporuje informovanost o licenci `Creative Commons`.

2.4.2 Partneři projektu

Hlavní partneři projektu WebArchiv jsou: [22]

Moravská zemská knihovna [11] provádí výzkum, má za úkol zveřejňovat výsledky výzkumu a publikuje informaci o projektu. Jejím zřizovatelem je Ministerstvo kultury ČR.

ÚVT Masarykovy univerzity [12] zabývá se výzkumem v oboru digitálních knihoven, zpracování a uchování multimédií.

CZ.NIC [13] Poskytuje WebArchivu seznam domén pro celoplošné sklizně v rámci `domény CZ`.

2.4.3 Nasmlouvané webové zdroje

Mimo licence Creative Commons má WebArchiv uzavřené smlouvy s autory webových zdrojů. Smlouva umožňuje archivovat tyto zdroje. Vlastníci webových informačních zdrojů by měli mít zájem o archivaci svých webových zdrojů. Důvodů pro to je mnoho. Hlavním je samozřejmě dostupnost starých verzí webů. Často se stává, že majitel webu ztratí nebo přijde o stará data. Je pro něj tudíž výhodné, že si je

9. Webové stránky projektu `Creative Commons` lokalizovaného do Českého jazyka dostupný online na adrese `http://www.creativecommons.cz/`.
10. Podrobný návod pro přidání stránky pod licenci `Creative Commons` dostupný online na adrese `http://www.webarchiv.cz/cc2`.
11. Bližší informace o Moravské zemské knihovně dostupné online na adrese `http://www.mzk.cz`.
12. Bližší informace o UVT Masarykovy univerzity dostupné online na adrese `http://www.muni.cz/ics`.
13. Bližší informace o organizaci CZ.NIC dostupné online na adrese `http://www.nic.cz`.

může dohledat v WebArchivu. Navíc jde podle mého názoru o vcelku prestižní záležitost – mít webové stránky uložené do budoucna pod hlavičkou zachování českého kulturního dědictví. [23]

V současné době má WebArchiv uzavřené smlouvy se zhruba 2932 autory webových stránek v prostředí domény CZ. Bližší informace o smlouvách jsou samozřejmě k dohledání na stránkách projektu[14]. [24] WebArchiv má samozřejmě i zpracovanou politiku přijímání nových zdrojů. Preferovány jsou „*především materiály významné kulturní či vědecké hodnoty, které vznikají jako původní digitální díla (tzn. nemají souběžnou tištěnou podobu)*“. [25] O politice přijímání nových zdrojů více v části práce věnující se kritériím sklizní.

2.4.4 Internet Archive

Internet Archive se zabývá řešením několika projektů související s archivováním webových zdrojů a uchováváním digitálních dokumentů. Jde o neziskovou organizaci sídlící ve městě San Francisco.

Mezi projekty organizace patří například nástroj WayBack, který využívá i WebArchiv, Open Library, který má za cíl archivovat knihy. Kompletní seznam projektů týmu Internet Archive je k dohledání na webu[15].

Je nasnadě, že spolupráce řešitelů projektu WebArchiv s týmem Internet Archive je velmi důležitá. Například z hlediska získávání zkušeností, technologií a informací. Projekty spolupracují i při vyhledávání webových zdrojů. Vyhledávání ve WebArchivu je přímo napojeno na možnost vyhledávat webové zdroje v databázi Internet Archive.

2.4.5 Další spolupráce

Řešitelé projektu samozřejmě spolupracují s řešiteli podobných projektů v zahraničí. Tato spolupráce je oboustranně velice výhodná. Dochází při ní k výměně zkušeností, informací a i nástrojů. Jelikož problematika archivace webových zdrojů je pořád relativně nová disciplína je potřeba neustále konzultovat další postup. Pracovníci WebArchivu se

14. Přehled smluv mezi projektem WebArchiv a majiteli domén dostupný online na adrese `http://www.webarchiv.cz/partneri`.
15. Projekty prováděné organizací Internet Archive dostupné online na adrese `http://www.archive.org/projects/`.

účastní odborných konferencí International Internet Preservation Consortium[16] (IIPC). Od roku 2007 je projekt WebArchiv členem IIPC, což řešitelům přináší řadu výhod. Například přístup k softwaru. [2]

2.4.6 Dostupnost informací

O projektu WebArchiv se lze dočíst na stránkách Národní digitální knihovny[17], která projekt řeší a samozřejmě na stránkách WebArchivu. K nalezení jsou tiskové zprávy, práce, které byli na toto téma již publikovány, informační letáky, příspěvky z konferencí. Popisem diplomových prací na téma WebArchiv se věnuji v příslušné části této práce.

Řešitelé se samozřejmě věnují i tomu aby WebArchiv vešel do povědomí veřejnosti. Řešitelé vystupují na konferencích, účastní se konferencí týkajících se archivace webu, publikují zprávy ainformační letáky, umožňují partnerům aby formou banneru informovali návštěvníky svých webů o existenci WebArchivu. Na webu projektu je dostupné velké množství dokumentů zabývající se problematikou archivace webových zdrojů a popularizací WebArchivu. Projektem WebArchiv se také intenzivně zajímá například Elektronický časopis o informační společnosti – Ikaros[18].

Bohužel se stává, že některé dokumenty nejsou dostupné nebo jsou porušené. Výroční zprávy jsou podány pro laickou veřejnost opravdu nezáživnou formou. Popularizační prezentace jsou zastaralé a nesplňují pravidla pro tvorbu kvalitních, sémantický a přístupných dokumentů. Na stránkách Národní digitální knihovny nelze nalézt prakticky žádné podstatné informace.

2.4.7 Kritéria výběru webových zdrojů

Při sklízení webových zdrojů se WebArchiv řídí předem stanovenými kriterii. Pravidelně se provádí několik druhů sklízí. Jedná se o plošnou sklizeň webových zdrojů na `doméně CZ` a o výběrové sklizně.

---

16. Webové stránky organizace International Internet Preservation Consortium dostupné online na adrese `http://netpreserve.org/about/index.php`.
17. Webové stránky Národní digitání knihovny dostupné online na adrese `http://www.ndk.cz/`.
18. Webové stránky Elektronického časopisu o informační společnosti Ikaros dostupný online na adrese `http://www.ikaros.cz/`.

V souladu s legislativou ČR jsou výběrové sklizně prováděny u zdrojů u nichž jejich autor udělal souhlas s jejich archivací a následným zveřejněním třetím osobám – jedná je buď o uzavření smlouvy s WebArchivem nebo o zdroje pod licencí `Creative Commons`.

Plošná archivace – harvesting

Plošná archivace má za cíl archivovat co největší počet bohemikálních zdrojů. Harvesting není náročný na filtrování a výběr sklízených zdrojů. Hlavní podmínkou pro plošnou sklizeň je doména webového zdroje. Typicky se jedná o zdroje umístěné na `doméně CZ`. Řešitelé ale už vyvinuli nástroj, který umožní sklízet bohemikální zdroje i mimo tuto doménu.

Další podmínky určují například formát sklízených dat, přístup ke zdroji dat, protokol atd. Z archivace jsou vyřazeny například streamované protokoly a obsah peer-to-peer sítí. Z důvody kapacity úložiště se stahují soubory s velikostí do 100 MB. Dalším kritériem je maximální počet souborů pro jeden zdroj. [26]

Výběrový přístup – konspekt

Výběrový přístup bere oproti plošné ohled na více parametrů. Zdroje jsou řazeny v tematických okruzích, těch není prozatím mnoho, jsou jednoduše přehledné. Přehled oborů sklizených v rámci konspektu je dostupný na stránkách WebArchivu.

Podmínky jsou podobné jako u plošné sklizně. Zdroj je umístěn na některé z národních domén, splňuje národnostní aspekty České republiky. Obsah zdroje musí být významný v kontextu české kultury. Zdroj musí být volně přístupný.

Důraz je kladen na dokumenty, které nejsou dostupné v jiné než elektronické formě. Jsou sklízeny časopisy, monografie, výzkumné zprávy, akademické práce, blogy, informační dokumenty, vládní dokumenty, tiskové zprávy. Nutné je sledování integrity zdroje a frekvence změn zdroje a podle toho určit frekvenci jeho sklízení. Nearchivují internetové aplikace, intranetové zdroje, weblogy, portály bez autorského obsahu a databáze. [26]

Tematické sbírky

Poslední přístup, který je využíván se orientuje na monotematické webové zdroje. Kritéria výběru jsou opět podobná jako u předchozích typů sklizní. Bohužel těchto, dle mého názoru, informačně velmi hodnotný sklizní je veřejně dostupných pouze devět1. Jedná se například o sbírku na téma: Výročí obsazení Československa 1968, České předsednictví EU, Prezidentské volby 2008. [26]

2.5 Starší práce na téma WebArchiv

V této kapitole se zaměřím na diplomové a bakalářské práce, které se týkají WebArchivu. Pokusím se ukázat jejich přínos a význam.

2.5.1 Systém pro správu procesu archivace webových informačních zdrojů

Diplomová práce, kterou zpracoval Adam Brokeš v roce 2009, měla za cíl vytvořit systém pro správu zdrojů, vydavatelů a smluv projektu. Autor ve své práci podává informace o projektu, provádí analýzu stavu WebArchivu, zabývá se historií projektu, pracovními postupy a stavem projektu, a právním rámcem archivace webových zdrojů.

Dále analyzuje používané nástroje. Konkrétně nástroj `Heritrix`, `WayBack` a `AutoContractMarker`. Následně analyzuje `WA Admin`. Výsledkem analýzy je popis nedostatků. Autor má zejména výhrady k datovému modelu, neflexibilitě systému a nemožnost hodnocení zdrojů. Protože nástroj `WA Admin` shledá nedostatečným pro další užívání, hledá za něj náhradu. `NetarchiveSuite`, který používá Dánská národní knihovna, nedoporučuje použít především kvůli rozdílnosti České a Dánské legislativy. Druhý nástroj, který připadá v úvahu, je `Web Curator Tool` vyvinutý národními knihovnami Velké Británie a Nového Zélandu. Ten také zamítá, protože pracovní postupy, které jsou v něm implementovány, jsou odlišné od postupů WebArchivu.

Rozhodne se tedy vytvořit vlastní nástroj. Tomu předchází analýza případů užití – stávajících pracovních postupů. Závěrečná část diplomové práce se zabývá návrhem a implementací nástroje `WA Admin v2`. Nástroj `WA ADMIN v2` je v současné době užíván týmem WebArchivu. [27]

2.5.2 Identifikace a omezení přístupu k „nevhodným“ stránkám ve webovém archivu

Autorem bakalářské práce je Filip Kusalík, který ji dokončil v roce 2009. Cílem bylo vytvořit nástroj, který identifikuje webové zdroje v archivu, jenž podle zákona není možné uchovávat. Autor ve své práci provede návrh nástroje a také ho implementuje. `ImpEval` je implementovaný v jazyce Java a je v současné době nasazen v běžném provozu.[28]

2.5.3 Rozpoznání a archivace českého webu mimo národní doménu

Bakalářskou práci vypracoval v roce 2008 Ivan Vlček v programovacím jazyce Java a jejím výstupem byl nástroj `WebAnalyzer`, který v současné chvíli WebArchiv využívá jako zásuvný modul nástroje `Heritrix`. Na začátku své práce autor provádí analýzu používaných nástrojů, se kterými bude muset `WebAnalyzer` spolupracovat. Zbytek práce je věnován analýze a návrhu `WebAnalyzeru`. Řešitel musel navrhnout aplikaci, i způsob její integrace do ostatních nástrojů. Autor při řešení práce kladl důraz na modularitu a flexibilitu nástroje a na jeho vývoji pokračoval i po dokončení bakalářské práce. [11]

2.5.4 Implementace `OAI-PMH` pro český WebArchiv

Martin Bella si ve své bakalářské práci, kterou dokončil v roce 2008, kladl za cíl navrhnout a implementovat rozhraní data-providera protokolu `OAI-PMH`. Toto rozhraní mělo být realizováno tak, aby se stalo součástí nástroje `WA Admin`. `OAI-PMH` protokol1 je určen pro získávání metadat z volně dostupných archívů. Protokol pracuje nad protokolem `HTTP`. Pro WebArchiv byl důležitý proto aby získal přesné informace o obsahu své databáze sklizených webových zdrojů.

Řešitel v práci podává podrobné informace o protokolu `OAI-PMH`, o možné aplikaci na WebArchiv, konstruuje datový model nástroje a následně nástroj implementuje. [29]

2.5.5 Zpřístupnění archivu českého webu

Diplomovou práci obhájil v roce 2006 Lukáš Matějka. Jejím výstupem byl nástroj `WA-CZ`. Nástroj umožňuje průběžnou sklizeň a indexaci

webových zdrojů. Autor v práci podává přehled archivačních formátů, používaných archivačních nástrojů, nástrojů pro zpřístupnění a mapuje vývoj projektu. Podává návrh aplikace a následně představuje svojí implementaci. Práce byla implementována v jazyce Java a začleněna mezi ostatní nástroje. [30]

2.6 Nástroje související s projektem WebArchiv

V této kapitolo představuji základní nástroje, které jsou používané v rámci WebArchivu.

2.6.1 APACHE TOMCAT

Apache Tomcat[19] jsem používal na svém počítači při práci na cvičných sklizních. Jedná se o kvalitní a dobře použitelný virtuální server.

2.6.2 ARCRetriever

„*Arcretriever je modul pro dodání dokumentu, jenž je součástí systému WERA. Java web aplikace na základe jména archivního souboru a pozice (offsetu) zobrazuje dokumenty vrácené z archivu.*“ [30] Tento nástroj byl používán ve WebArchivu v kombinaci s nástrojem WERA a NutchWAX.

2.6.3 DeDuplicator

Jedná se o modul programu Heritrix, který má na starost zachytávání duplicitních souborů v rámci série sklizní. Modul vyvíjí National and University Library of Iceland a je volně dostupný[20] ke stažení. Nástroj byl poprvé nasazen při sklizni v roce 2010, funguje tak, že kontroluje duplicitu archivovaných souborů, které mají jiný mimetype než html/text.

19. Webové stránky projektu Apache Tomcat dostupné online na adrese http://tomcat.apache.org/.
20. Webové stránky nástroje DeDuplicator dostupné online na adrese http://deduplicator.sourceforge.net/.

2.6.4 Generátor URN

Nástroj dostupný na stránkách projektu WebArchivu[21]. Slouží ke generování Dublin Core Metadata Element Set k zadané URL adrese. Nástroj byl vytvořen pracovníky WebArchivu.

2.6.5 HERITRIX

Nástroj Heritrix využívá WebArchiv přímo v provozu na sklizně webových zdrojů. Používal jsem ho při cvičných sklizních. Nástroj umožňuje široké množství možností pro konfiguraci parametrů sklizně. Jedná se například o maximální objem stažených dat, ale i o nastavení procházení webové stránky při jejím stahování. Jsou pro něj vytvářeny i moduly, které doplňují jeho funkcionalitu. Nástroj je volně dostupný ke stažení[22]. Je to jeden z nejpoužívanějších crawlerů.

Po spuštění nástroj prohledává určitou webovou adresu a stahuje dokumenty, které na ní nalezne. Vše ukládá do formátu arc (a nově i do formátu warc) a následně komprimuje nástrojem GZip[23]. WebArchiv preferuje soubory s velikostí 100 MB. Uživatel může během probíhající sklizně sledovat jejich průběh, Heritrix nabízí širokou paletu možností jak lze procházet sklizené soubory a filtrovat si v jejich seznamech. Samozřejmě podává i statistiku sklizeného zdroje.

Nástroj Heritrix nově od verze 1.12 dovede ukládat data do warc archivu a přitom provádět deduplikaci souborů. Nástroj při sklízení dat zjišťuje, zda se soubory od poslední sklizně změnily, pokud se tak nestalo, vytvoří zkrácený záznam, který slouží jako ukazatel na daný soubor - užívá při jednoznačné identifikaci SHA-1 algoritmus [24]. S takovými warc záznamy dále pracuje interpret staženého zdroje. Funkcionalitu implementuje add-on modul DeDuplicator. [38] [39]

Deduplikací se podrobněji zabývám v části práce věnované formátu warc na s. 31.

---

21. Nástroj Generátor URN dostupný onlina na adrese http://www.webarchiv.cz/generator/dc_generator.php.
22. Webové stránky projektu Heritrix dostupné online na adrese http://crawler.archive.org/.
23. Podrobné informace o formátu GZip dostupné online na adrese http://www.gzip.org/zlib/rfc-gzip.html.
24. Popis RFC standardu SHA-1 dostupný online na adrese http://tools.ietf.org/html/rfc3174.

2.6.6 ImpEval

Nástroj vytvořený v rámci bakalářské práce vypracované Filipem Kusalíkem v roce 2009 s názvem Identifikace a omezení přístupu k „nevhodným“ stránkám ve webovém archivu. Nástroj identifikuje webové zdroje v archivu, které podle zákona není možno uchovávat. Nástroj je v současné době používán v běžném provozu WebArchivu. [28]

2.6.7 NutchWAX

Nástroj používá WebArchiv pro fulltextové vyhledávání ve svém archivu. Je napojen na nástroj WayBack, který nenabízí možnost fulltextového vyhledávání v archivu. Nástroj byl v projektu WebArchiv používán v kombinaci s nástroji WERA a ARCRetriever. NutchWAX je součást projektu[25] Nutch, který zahrnuje nástroje pro archivaci a vyhledávání webových zdrojů včetně práce s arc soubory.Nutch nabízí komplexní řešení archivace webových zdrojů.

2.6.8 WA ADMIN

V roce 2009 realizoval Adam Brokeš WA Admin v2 v rámci své bakalářské práce s názvem Systém pro správu procesu archivace webových informačních zdrojů. WA Admin v2 nahradil nástroj vytvořený Lukášem Matějkou a stejně jako WA Admin sloužil pro správu zdrojů, vydavatelů a smluv WebArchivu. [27]

2.6.9 WA-CZ

Tento nástroj vytvořil jako v roce 2006 Lukáš Matějka v rámci své diplomové práce s názvem Zpřístupnění archivu českého webu. Nástroj sloužil pro průběžnou sklizeň a indexaci stažených webových zdrojů. Podrobněji se nástrojem zabývám v kapitole, která se týká starších prací. [30]

---

25. Webové stránky projektu NutchWAX dostupné online na adrese http://archive-access.sourceforge.net/projects/nutch/.

2.6.10 Wayback

Wayback je nástroj pro vyhledávání dokumentů v archivech. WebArchiv jej používá v běžném provozu. Je volně dostupný[26] na internetu. Používal jsem ho pro prohlížení mnou stažených webových zdrojů pomocí virtuálního serveru Apache Tomcat. Pro vyhledávání v archivu je potřeba vytvořit CDX index[27] všech souborů, které jsou archivované. To se provádí nástrojem, který je k WayBacku přidružený. WayBack obsahuje větší sadu nástrojů pro práci s CDX indexy.

Po nakonfigurování WayBack zobrazuje všechny archivované webové zdroje nalezené podle specifické URI adrese. Umožňuje jejich zobrazení podle data jejich archivace a nabízí procházení jednotlivých sklizených verzí.

Nástroj umožňuje též interpteraci stažených webových zdorojů formátu warc, které užívají deduplikaci. Dokáže tedy interpretovat warc záznamy, které slouží jako ukazatele na existující warc záznamy. [38]

Deduplikací se podrobněji zabývám v části práce věnované formátu warc na s. 31.

2.6.11 WebAnalyzer

Nástroj, který pro WebArchiv vytvořil Ivan Vlček v rámci své bakalářské práce s názvem Rozpoznání a archivace českého webu mimo národní doménu v roce 2008. Nástroj se v současné době používá. [11]

2.6.12 Web Curator Tool

Nástroj vyvinutý v roce 2006 National Library of New Zealand a The British Library. Nástroj slouží k řízení a stahování webových zdrojů a je stále podporován[28]. Stejně jako Dánský NetarchiveSuite používá Heritrix, na rozdíl od NetarchiveSuite je Heritrix jeho součástí.

---

26. Webové stránky projektu Waybak dostupné online na adrese http://archive-access.sourceforge.net/projects/wayback/> .
27. Jedná se standardní textový soubor, který obsahuje základní informace o souborech obsažených v internetovém archivu. Jednotlivé záznamy jsou v něm odděleny pouhým odřádkováním. Příklad CDX indexu je k nalezení v příloze H na s. 57.
28. Webové stránky projektu Web Curator Tool dostupné online na adrese http://webcurator.sourceforge.net/.

Kapitola 3

Archivační formáty

3.1 Úvod

Jak jsem zmiňoval již výše všechna sklizená data jsou ukládána na datová úložiště, která se fyzicky nacházejí v Brně a v Praze. Samozřejmě se provádějí pravidelné zálohy stažených dat. Pro shromažďování tak velkého objemu dat je potřeba zvolit vhodný formát pro jejich archivaci.

Řešitelský tým musel při volbě vhodného formátu počítat se spoustou nároků, které bude na archiv v budoucnosti klást. Zejména se jednalo o požadavky týkající se dostupnosti uložených dat. S archivovanými daty je potřeba v krátkém čase manipulovat.

Jelikož množství souborů[1], které se v rámci jednotlivých sklizní stáhne se stále zvětšuju, bylo nutné jednotlivé soubory sdružovat a ne je jednoduše umístit do adresáře v úložišti. S takto volně uloženými daty by se špatně manipulovalo a i jejich archivace by nebyla snadná.

Jako nejlepší řešení se tedy ukázaly speciální typy souborů, které slouží k archivaci webových zdrojů. V minulosti byl používán formát nedlib, v současné době je preferovaný formát `arc`. V budoucnosti se počítá s využitím formátu `warc`. Formát `warc` je nejnovější typ internetového archivu, a jako takový má mnoho výhod oproti starším typům archívů.

Z dlouhodobého hlediska je nutné udržovat stažená data v jednom formátu. Právě formát `warc` se zdá k tomuto účelu vhodný. Předmětem mé práce je právě prozkoumat možnost migrace starších archivů, uloženým ve formátu `arc`, do formátu `warc`.

V následující kapitole představím podrobněji archivační formáty a ukáži jejich výhody a nevýhody.

---

1. V posledních letech přesahuje desítky miliónů.

3.2 Formát `nedlib`

Archivy v tomto formátu již WebArchiv nepoužívá. Informace o něm jsou dostupné v diplomové práci Lukáše Matějky.

3.2.1 Popis standardu `nedlib`

Typický archiv obsahoval 2000 souborů. Polovinu z nich tvořily stažené soubory a druhou polovinu soubory obsahující metadata ke staženým souborům. Každému staženému souboru odpovídal jeden soubor s metadaty, ve kterém byly obsaženy informace o souboru – například typ souboru, velikost a další.

Názve souboru uvnitř archivu odpovídal `MD5`[2] součtu příslušného souboru. Vyhledávání požadovaného souboru v archivu se realizovalo pomocí indexu s `MD5` názvy souboru.

3.2.2 Výhody `nedlib` pro WebArchiv

- Metadata byla zvlášť oddělena pro každý archivovaný soubor
- Soubor přímo v názvu obsahoval `MD5` otisk svého obsahu

3.2.3 Nevýhody `nedlib` pro WebArchiv

- Špatná manipulace s jednotlivými soubory v rámci archivu
- Nutnost rozbalit celý archiv při přístupu k jednotlivým souborům

3.2.4 Užití v praxi ve WebArchivu

Tento typ souboru se již nevyužívá a archivy v tomto formátu byli převedeny do formátu `arc`. Značná část byla ztracena při havárii úložiště. [30]

---

2. `MD5` je standardní hashovací algoritmus, které se používá například k šifrování hesel, ale i pro vytváření jednoznačných identifikátorů – kontrolních součtů – souborů. Jelikož byly objeveny kolize šifrovací funkce, tak se od `MD5` začalo v posledních letech opouštět. Popis `MD5 RFC` standardu dostupný na adrese `http://www.ietf.org/rfc/rfc1321.txt`.

3.3 Formát arc

3.3.1 Popis standardu arc

Arc je bezeztrátový kompresní formát pro ukládání dat. Byl vyvinut v 80. letech a od roku 1996 je užíván WebArchivem pro ukládání stažených webových zdrojů. V současné době WebArchiv při sklízen webových zdrojů, pomocí nástroje Heritrix, používá právě arc soubory komprimované do formátu GZip. Tuto komprimaci zajišťuje přímo nástroj Heritrix. Archivy mají typicky velikost 100 MB, pouze při archivaci větších souborů je tento limit překročen.

Arc soubory obsahují kromě samotných stažených souborů ještě metadata, která nesou další informace o stažených souborech. Jedná se například o kontrolní součet souborů, jejich URL, datum stažení, délku souboru. Kompletní popis užívaných souborů je k dohledání na webové stránce projektu Internet Archive, popřípadě na stránkách IA Webteam JIRA[3]. [31] Pro další informace viz též přílohy A, B, C, D, E a G na s. 57.

3.3.2 Výhody arc pro WebArchiv

- Na rozdíl od formátu nedlib je soubor dobře čitelný – soubory stažených webových zdrojů jsou ukládány v archivech a členěny.
- Umožňuje vytváření kontrolních součtů archivovaných souborů. Respektive v jedno volné metadatové pole je určeno pro ascii checksum. V praxi se využívá MD5 kontrolních součtů.

3.3.3 Nevýhody arc pro WebArchiv

- Neumožňuje na rozdíl od warc zabránit duplicitnímu ukládání dokumentů.
- Neobsahují samoopravný kód – pokud se archiv poruší tak není možné rekonstruovat jeho obsah. To však není dáno vlastnostmi arc archivů, ale tím, že archivy se po vytvoření ještě komprimují nástrojem GZip.

3. Webové stránky IA Webteam JIRA dostupné online na adrese https://webarchive.jira.com/.

- Při načítání libovolného souboru z archivu je potřeba rozbalit celý archiv – při 100 MB archivech jde o velkou režii. Pokud ovšem není vytvořen index souborů uvnitř archivu, pomocí kterého se dozvíme na jaké pozici v archivu se který soubor nachází. Pak není potřeba soubor sekvenčně procházet, ale stačí pouze otevřít archiv a posunout ukazatel na správné místo. [30]

3.3.4 Užití v praxi ve WebArchivu

Právě metadat připojených v archivu se využívá při vyhledávání souborů v databázi WebArchivu. V praxi se vytvoří `CDX index` souboru, který obsahuje seznam souborů uložených v archivech společně s metadaty a cestou k danému archivu. Tento index se vytváří pomocí aplikace `CDX-indexer`, který je součásti nástroje `Wayback`. V `příloze H na s. 60` nabízím ukázku `CDX indexu`.

Při vyhledávání zdrojů v archivu se Wayback prochází `CDX index`, který ho odkáže k archivu s požadovaným souborem. Velmi přesný popis užití `arc` souborů v praxi vytvořil Adam Brokeš ve své diplomové práci: Integrace a automatizace systému v pracovních procesech projektu WebArchiv. [28]

3.4 Formát `warc`

3.4.1 Úvod

Přestože je formát `warc` již popsán jako `ISO standard 28500:2009`[4], je ve většině projektů jeho použití zatím pouze testováno. Před tím, než se jeho použití realizuje je potřeba podrobná analýza. Přesto tento formát již některé standardizované nástroje podporují. Například crawler `Heritrix` již umožňuje ukládání dat v tomto formátu.

Použití formátu `warc` se v rámci projektu WebArchiv prozatím jen plánuje. Před tím, než bude nasazen, je potřeba provést spoustu dílčích kroků. Je to proto, že formát je prozatím ještě nový a prozatím nebyly všechny testy provedeny. Při práci s velkými depozitáři dat je důležité každý krok, který bude zavádět systémovou změnu předem zvážit. Svou

4. `ISO standard` za poplatek dostupný online na adrese `http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717`.

roli v tom, že se o zavedení formátu uvažuje, je již výše zmíněná skutečnost: každá organizace zabývající se archivací webových zdrojů, používá jiné nástroje a postupy. Proto každá jednotlivě řeší rozdílné problémy s implementací nástrojů a jejich používáním.

V následujících kapitolách předvedu základní vlastnosti formátu warc a ukážu jeho výhody a nevýhody. Nevýhod související se zavedením formátu `warc`, neznamení, že by se mělo opustit od přechodu na tento formát. Nevýhody spíše představují problémy, které doposud nebyly vyřešeny nebo se vůbec neřešily. Před tím, než se přistoupí k zavedení tohoto formátu, je potřeba se se všemi vyrovnat.

V `přílohách I, J a K na s. 60` nabízím vzorové soubory typu `warc`, vydané k definici standardu. Dále v `přílohách L, M a N na s. 62` nabízím `warc` soubory vytvořené pomocí nástrojů pro migraci `arc` souborů, které jsem testoval. Také přikládám výstup z programu `JHOVE2` pro jeden z `warc` archivů v `příloze O na s. 63`. V `příloze P na s. 63` je dostupný dokument IIPC zabývající se formátem. Pro úplnost uvádím i ukázku `ISO standardu warc` v `příloze AJ`[5] `na s. 66`.

3.4.2 Popis standardu `warc`

Motivací pro definici standardu `warc` jsou tyto nároky na archivační formát: [33]

- Zpracovat všechny typy obsahů a kontrolních informací protokolů internetové aplikační vrstvy (jako `HTTP, DNS, FTP`).
- Zpracovat libovolná metadata související s jinými uložených daty.
- Podporovat kompresi a integritu dat.
- Ukládat všechny řídicí informace z týkající se těžby dat.
- Ukládat výsledky datových transformací spojených s jinými uloženými daty.
- Detekovat duplicitní uložení dat. Pomocí tohle lze odstranit zbytečná duplicitní data.

5. Nejedná se o současný oficiální standard, ale o poslední dostupnou specifikaci z roku 2009.

- Umožnit rozšiřitelnost bez toho, aby to ovlivnilo existující soubory a funkcionalitu
- Podporovat manipulaci a segmentaci s uloženými zdroji.

Soubor formátu `warc` tvoří jednoduché zřetězení jednoho nebo více záznamů typu `warc`. První záznam obvykle popisuje následující záznamy. Obecně platí, že obsah záznamu je buď záznam pokusu o vyhledání webového zdroje, nebo nese informace o archivovaném obsahu.

Warc záznamy se skládají z hlavičky, kterou následuje blok s obsahem. Hlacička se skládá z prvního řádku, který definuje, že záznam je typu warc, dále verze souboru, a flexibilní množství řádků oddělených symbolem `CRLF`, poté následuje již zmíněný blok obsahu. Záznam splňuje `BNF`[6], ukázka k nalezení v `příloze AM na s. 67`.

Seznam definovaných polí, která se mohou v záznamu zobrazovat lze nalézt v `příloze AN na s. 67`. Samozřejmě některá z polí jsou povinně[7] v každém záznamu a jiné jsou nepovinné[8]. Každé z polí má samozřejmě stanovená pravidla, která určují jejich možný obsah. [33]
`Warc` záznamy jsouo několika různých typů, podle toho k čemu záznam slouží: [33]

Warcinfo popisuje ostatní `warc` záznamy, které ho následují. Typicky je na začátku `WARC` souboru. Má tedy čistě informativní charakte. Lze v něm najít informace o tom, jaký software daný soubor vytvořil, `IP adresa` stroje na kterém byl vytvořen atd. Příklad záznamu v `příloze K na s. 63`.

Response nese odpověď na určitý dotaz. Obsahuje informace o tom jak daný dotaz probíhal. Lze v něm najít `HTTP` hlavičky dotazu atd.

Resource je určitý stažený zdroj, ale bez informací protokolu. Ty jsou, jak je zřejmé, neseny v jiném typu záznamu. Data jsou v něm uloženy v `Record block`. Záznam se bude lišit podle toho, jestli bude obsahovat data z určitého `HTTP` zdroje, `ftp` zdroje atd.

---

6. Backusova-Naurova forma je jedno z možných vyjádření bezkontextové gramatiky.
7. Povinné: `WARC-Record-ID, Content-Length, WARC-Type`
8. Nepovinné: `Content-Type, WARC-Block-Digest, WARC-Payload-Digest`

Request je žádost o webový zdroj. Obsahuje detailní informace o provedeném dotazu a o žádaném webovém zdroji. Příklad záznamu v `příloze J na s. 61`.

Metadata nese metadata sklizených zdrojů. Přesněji řečeno obsahuje informace o libovolném staženém jiném záznamu. Slouží tak pro reprezentaci stažených dat v repozitářích. Příklad záznamu lze nalézt v `příloze I na s. 60`.

Revisit slouží k znovunačtení určitého zdroje. Důvod proč znovunačíst určitý zdroj může být například porovnání uložených dat se vzdálenými daty. Pomocí tohoto typu záznamu se tedy dá například ošetřit ukládání redundantních dat.

Conversion je záznam, který obsahuje alternativní verzi uloženého záznamu. Slouží tedy k udržení informací o rozdílech mezi určitými záznamy. Tento typ záznamu tedy podává důležité informace o verzích či variantách již uložených záznamů.

Continuation je pokračování jiného záznama. Tento typ záznamu se používá pokud je určitý záznam třeba segmentovat do více záznamů. Příklad záznamu typu `Continuation` v `příloze A0 na s. 67`.

Záznam typu `Continuation` tedy umožňuje segmentaci `warc` souborů. Pokud by nějaký `warc` soubor měl překročit stanovenou maximální velikost, je možné ho rozdělit do více souborů. Informaci o segmentaci nese pole `WARC-Segment-Number`, ale k řízení a zajištění segmentace slouží i další pole. [33]
Standard dále doporučuje, jak nakládat s archivy typu `warc`: [33]

- Nedefinuje žádnou vnitřní kompresi, ale doporučený je nástroj `Gzip`.
- Doporučená velikost je 1 GB.
- Doporučený název archivu má tvar:
  `Prefix-Timestamp-Serial-Crawlhost.warc.gz`

3.4.3 Výhody warc pro WebArchiv

- Do budoucna pravděpodobně většina institucí přejde na archivaci webových stránek do warc archivů. Warc tedy bude nejspíše označen za jediný univerzálně podporovaný formát. Jedná se o ISO standard a bude jím s největší pravděpodobností i v budoucnosti.

- Ve warc archivech se neukládají archivované soubory duplicitně. Pokud již archivovaný soubor je jednom z warc archivů tak se na něj ostatní odkazují. To přináší úsporu kapacity. Zjistil jsem, že při převodu archivu Britské knihovně, zvažovali pracovníci, jestli tuto funkcionalitu využít. Nakonec se však rozhodli, že časová náročnost řešení, je pro ně důležitější než úspora místa, které by dosáhli deduplikací. [34]

- Na rozdíl od arc archivů umožňují warc archivy ukládat nejen HTTP odezvy, ale také původní požadavky.

- Warc umí pracovat s více protokoly: HTTP, FTP, NNTP a SMTP.

- Warc umožňují ukládat větší škálu metadat. Je možné přidávat metadata podle vlastních potřeby, což se může využít při speciálních požadavcích určených vnitřní politikou ukládání souborů.

- Na rozdíl od arc formátu je možné ukládat více druhů kontrolních součtů pro archivované soubory. To se dá jednoduše realizovat pomocí škálovatelných polí pro metadata. Arc archivy, které používá WebArchiv, provádí pouze MD5 kontrolní součty souborů. Na rozdíl arc archivů, je u warc archivu použít například i s SHA-1 hashovací funkcí pro jednoznačné určení stažených dokumentů.

- Při načítání libovolného souboru z archivu není potřeba rozbalovat celý archiv – při 100 MB archivech jde o velkou režii. Každý warc záznam má svůj vlastní offset, přes který se dá k libovolnému souboru v archivu. Pak není potřeba soubor sekvenčně procházet, ale stačí pouze otevřít archiv a posunout ukazatel na správné místo. [1]

3.4.4 Nevýhody `warc` pro WebArchiv

- Neobsahují samoopravný kód – pokud se archiv poruší tak není možné rekonstruovat jeho obsah. To však není dáno vlastnostmi `warc` archivů, ale tím, že archivy se po vytvoření ještě komprimují nástrojem `GZip`.
- Prozatím není vyřešeno, jak efektivně využít možnost ukládání větší škály metadat. Za tímto účelem bude nutno provést další studie.
- Prozatím nejsou implementovány nutné nástroje na práci s warc soubory. Před jejich zavedením v projektu WebArchiv bude potřeba tyto nástroje implementovat.
- Zpětná převoditelnost z `warc` do `arc` není prozatím triviálně možná. [32] Tím pádem se nedají případné nové sklizně ve formátu `warc` zařadit ke starým sklizním ve formátu `arc`.
- Převod celého archivu do formátu `warc` bude časově a výpočetně velmi náročný a jakákoliv chyba může znamenat ztrátu dat.

3.4.5 Plánované využití

Jak jsem již zmínil výše, s využitím formátu `warc` se v rámci projektu WebArchive počítá, ale prozatím není provedeno dostatek analýz. Nejdříve bude potřeba počkat na vyvinutí všech potřebných nástrojů.

Samotný převod celého stávajícího archivu bude pravděpodobně proveden až v poslední fázi přechodu k formátu `warc`. To plyne z citlivosti problému a nutnosti zajistit, aby nedošlo ke ztrátě dat.

3.4.6 Deduplikace souborů

Důležitým aspektem standardu `warc` je možnost `deduplikace` souborů. Jedná se o možnost odstranění těch souborů, které se v repozitáři objevují vícekrát.

Pokud se například v různých verzích určitého staženého zdroje nachází stejný soubor, je výhodné ho z důvodu žetření datové kapacity úložiště ponechat v archivu pouze jednou. `Deduplikace` souboru však

není jednoduchá operace. Nese s sebou spoustu problémů, které je třeba vyřežit.

Hlavním problém je, že takové duplicitní soubory se budou typicky nacházet v různých archivech. Proto je problematické při interpretaci určité verze webového zdroje implementovat paralelní načítání souboru z jiných archivů. Dále je problematické uchovávání konzistentních dat v jednotlivých verzích webových zdrojů – ztráta jednoho archivu by mohla znamenat ztrátu dat pro více verzí webového zdroje.

Samostatným problémem je převedení starých archivů ve formátu `arc` do formátu `warc` a přitom zajistit bezpečnou deduplikaci uložených dat. Například v British Library možnost deduplikace při migraci svého archivu ani nevyužili. [34] Řežitelé projektu WebArchiv by však rádi této možnosti využili rádi.

Mnou testované nástroje deduplikaci při migraci neumožňují, funkcionalita by se však dala implementovat. Nepodařilo se mi najít zmínku o tom, že by nějaký volně dostupný migrační nástroj tuto možnost využíval. Hlavním problémem při nasazení takového migračního nástroje by bylo to, že by takový migrační nástroj nemusel splňovat všechny požadavky, které by na ně WebArchiv kladl. [9] Bylo by proto pravděpodobně nutné, aby řežitelé takový nástroj sami implementovali nebo minimálně upravili jiné řešení.

Samotný formát `warc` nenabízí možnost `deduplikace`, ale díky široké žkále metadat, které uchovává, je realizace deduplikace umožněna mnohem lépe, než u staršího formátu `arc`. [33] `Deduplikace` je tedy řežena na úrovni nástrojů, které `warc` soubory vytvářejí. Typickým příkladem nástroje, který deduplikuje stažená data je nástroj `Heritrix`. Ale existují i jiné nástroje, které tuto funkcionalitu implementují[10]. Právě `Heritrix` je používán i ve WebArchivu. Problematikou deduplikace se zabývám v části věnované danému nástoji `Heritrix na s. 20` a v části, kde se věnuji nástroji `Wayback na s. 21`.

9. O rozdílnosti používaných nástrojů v různých projektech zabývajících se archivací se zabývám v jiné části této práce.

10. Například nástroj `ArchiveIt`. Webové stránky nástroje `ArchiveIt` dostupné onlina na adrese `https://webarchive.jira.com/wiki/display/ARIH/Archive-It+Feature+Release+3.0`

Kapitola 4

Migrační nástroje

4.1 O migraci

4.1.1 Úvod

Převod `arc` archivů do formátu `warc` je specifický problém. Různé instituce používají různé formáty na ukládání webových zdrojů. Většina těchto institucí bude mít právě snahu převést uložená data do formátu warc.

Informace o migraci nejsou snadno dostupné. To je způsobené především tím, že problematika je relativně nová. Formát `warc` byl standardizován v roce 2009. Migrační nástroje jsou sice vyhledatelné na internetu, ale většinou není jejich součástí kompletní dokumentace.

Problém dostupnosti informací o migraci je především v tom, že každá instituce, která se archivací zabývá, používá jiné nástroje. Každá instituce si během své existence přizpůsobila nástroje pro své potřeby. Tyto potřeby může tvořit například jazykové, dostupnost nástrojů, rozdílnost legislativy, rozdílnost v politice ukládání webových zdrojů. Proto, pokud určitá instituce analyzuje problematiku migrace, zabývá se problémy specifickými pro její podmínky.

Proto například informace získané z Britské knihovny jsou pro WebArchiv hodnotné, ale ve skutečnosti nejsou jejich nástroje okamžitě použitelné. Pokud by je chtěli pracovníci WebArchivu využít bylo by nutné je předělat pro jejich potřeby.

4.1.2 Britská knihovna

Řešitelé projektu WebArhiv se migrací prozatím příliš nezabývali. Nicméně ve světě se tím již lidé zabývají. Proto jsem se pokusil zjistit informace o migraci u pracovníků z Britské knihovny. Mojí mailovou

korespondenci uvádím v příloze Q na s. 63. V Britské knihovně vyzkoušel vývojový tým skript pro převod do archivního formátu warc už v roce 2009. Celkově bylo převedeno cca 4,4 TB dat.

O převodu, který provedli, zatím není dostatek informací. Pracovníci Britské knihovny však použili pro migraci nástroje, které nejsou přímo použitelné ve WebArchivu. Používali totiž jiné nástroje pro ukládání dat a práci s archivy – Web Curator Tool a PANDAS. Překvapilo mě, že nevyužili ani možnosti deduplikace souborů. [34]

4.2 Použité nástroje

Kromě samotných nástrojů pro migrci arc archivů jsem používal i jiné nástroje, zde podávám jejich přehled.

4.2.1 JHOVE

Slouží k identifikaci a validaci souborů. Program disponuje množstvím modulů pro rozpoznávání formátů souborů. V příloze F na s. 59 je dostupná ukázka výstupu programu JHOVE pro arc soubor. Program je používal jen k pokusnému testování souborů.

Konkrétně program umožňuje identifikaci, validaci a charakterizaci digitálních objektů. Identifikace je proces, který určuje formát objektu. Validace určuje stupeň korespondence daného objektu se specifikací příslušného formátu. Charakterizace navíc informuje o specifických vlastnostech daného objektu.

Tyto operace jsou potřebné při správě digitálního archivu. Jeho použití je nutné při uchovávání digitálních souborů. Řešitelé projektu Webarchiv musí nutně vědět, jaké typy souborů uchovávají v repositářích. Jedním z hlavních důvodů je možnost zpětné převoditelnosti uchovaných webových zdrojů a jejich interpretace[1].

JHOVE obsahuje moduly pro práci s bytestreamy, ASCII a UTF-8 texty, AIFF a WAVE audio soubory. Dále s obrázky typu GIF, JPEG, JPEG 2000, TIFF, také PDF soubory, textovými a XML output handlery. Je naprogramován v programovacím jazyce Java.

1. Různé verze elektronických dokumentů je možno zobrazit správně pouze specifickými nástroji. Proto je potřeba mít informace o tom, jaké verze dokumentů archiv obsahuje.

Jelikož je již zastaralý, používal jsem pro svou práci jeho novější verzi: JHOVE2. Přesto je tento nástroj stále vyvíjen a dostupný k volnému stažení[2].

4.2.2 JHOVE2

Úvod

Program je novější verzí již zmíněného programu JHOVE. Slouží k validaci a identifikaci souborů. Program je vyvíjen v mnoha variantách[3] – v rámci distribuovaného verzovacího systému Mercurial na serveru www.bitbucket.com. Je vyvíjen JHOVE2 Project Team – California Digital Library, Portico, Stanford University .

Protože je JHOVE2 vyvíjen mnoha různými vývojáři, vznikají různé verze tohoto programu se specifickými druhy přidané funkcionality. Pro mé účely jsem použil oficiální verzi[4] a dále verzi určenou pro validaci testování souborů[5], která má zásuvný modul pro analýzu arc souborů.

Program jsem použil pro zjištění obsahu převáděných arc archivů. Je totiž potřeba přesně zjistit obsah archivu. Jelikož při vytváření archivů dochází občas k tomu, že není správně identifikován mimetype archivovaných souborů. Tato informace je však zásadní pro zpětnou interpretaci obsahu stažených souborů. V příloze G a O ma s. 60 je ukázka výstupu z nástroje pro arc a warc soubor. Příloha AH na s. 66 obsahuje archiv s nástrojem JHOVE2.

Implementací a analýzou několika archivů z depozitáře WebArchivu se zabývám v jiné kapitole.

---

2. Webové stránky projektu JHOVE dostupné online na adrese http://sourceforge.net/projects/jhove.

3. Distribuované verzovací systémy umožňují při vývoji projektu v určitém checkpointu rozdělit zdrojový kód na dvě části (forky) a pracovat na nich nezávisle. Přičemž checkpoint reprezentuje všechny dostupnými komponentami projektu. Přehled různých variant projektu JHOVE2 dostupný online na adrese https://bitbucket.org/jhove2/main/descendants.

4. Webová stránka projektu JHOVE2 dostupná online na adrese https://bitbucket.org/jhove2.

5. Webová stránka projektu JHOVE2-BNF dostupná online na adrese https://bitbucket.org/lbihanic/jhove2-bnf/overview.

Podrobnější informace JHOVE2

Detailní informace o programu jsou obsaženy v uživatelské příručce [6], která je k nahlédnutí v `příloze AF na s. 66`. Další podstatné informace lze nalézt v dokumentu z roku 2010, který se věnuje aktualizaci nástroje, je přiložen k bakalářské práci v `příloze AG na s. 66`. Nástroj je napsán v programovacím jazyce Java a ke svému správnému běhu vyžaduje `The OpenSP SGML parser`[7].

Jak jsem zmínil výše, program `JHOVE2` je novější verzí programu `JHOVE` a jako takový přináší nové možnosti a funkcionality. Hlavní změnou je to, že u objektů provádí charakterizaci, která je oproti původním partikulárním testovacím procesům u JHOVE komplexnější. „*Characterization is the process of examining a formatted digital source unit and automatically extracting or deriving representation information about that source unit that is indicative of its significant nature and useful for purposes of classification, analysis, and use.*“ [35]

Charakterizace se skládá ze čtyř procedur [35]:

Identifikace určení formátu objektu.

Feature extraction (extrakce rysů) určení charakteristických rysů objektu.

Validace určení stupně korespondence objektu se specifikací formátu.

Assessment určení stupně korespondence objektu se specifickými nároky určenými uživatelem softwaru. Nástroj je tedy možno „personalizovat“ pro potřeby uživatele, což je výhodné hlavně proto, že téměř každá organizace zabývající se archivací webových zdrojů, používá odlišnou politiku archivace dat. Nástroj poté dovede například určit rizika při práci s různými objekty, nebo provést či doporučit další akce ke zpracování objektů.

Mezi další přednosti programu patří [35]:

---

6. V celé kapitole používám při popisu funkcí a vlastností nástroje vlastní překlady z této příručky.

7. Webová stránka `The OpenSP SGML parser` dostupná online na adrese `http://search.cpan.org/dist/SGML-Parser-OpenSP/`

- Modulární architektura a práce s plug-iny
- Jednoduché API a design základních modulů
- Bufferovaní I/O operací
- Internacionalizovaný výstup: podporuje více formátů – JSON, XML, text.
- Široká možnost konfigurace nástroje
- Kompletní dokumentace nástroje
- Široká škála rozpoznatelných objektů[8]

Hlavní předností nástroje je samozřejmě široká škála typů souborů, které dokáže identifikovat a validovat. Identifikovat dokáže JHOVE2 všechny objekty obsažené v databázi PRONOM[9], respektive objekty, které definují DROID signature files[10] – jedná se o více než 550 formátů. Výčet typů objektů, které dovede JHOVE2 validovat obsahuje například: ICC color profile, JPEG 2000, PDF, SGML, Shapefile, TIFF, UTF-8, WAVE, XML, Zip.

Nástroj tedy poskytuje velký potenciál pro využití v při řešení problému s archivací webových zdrojů. Je podporován a užíván velkým množstvím organizací a dá se říci, že patří mezi uznávaný a standardní nástroj. [35]

4.2.3 Jpype

Nástroj umožňující spuštění programů napsaných v programovacím jazyce Java v rámci programu napsaného v programovacím jazyce Python.

Použití tohoto nástroje bylo nutné kvůli tomu, že nástroje pro migraci arc archivů jsou psány v programovacím jazyce Python, ale JHOVE2 je programován v jazyce Java.

---

8. V následujícím odstavci podávám pouze neúplný výčet podporovaných objektů, podrobný výčet lze nalézt v příloze AG na s. 66.
9. Organizace zabývající se problematikou souborových formátů a softwarových produků s nimi souvisejících. Webové stránky databáze PRONOM dostupné online na adrese http://www.nationalarchives.gov.uk/PRONOM.
10. DROID (Digital Record Object Identificatio), je nástroj pro automatickou identifikaci formátu souborů. Webové stránky nástroje DROID dostupné online na adrese http://droid.sourceforge.net.

Nástroj v podstatě umožní spuštění `Java Virtual Machine`[11] uvnitř `Python` skriptu a následně práci s Java aplikací.

Je samozřejmé, že implementace takového programu, který musí využívat více programovacích jazyků najednou, je problematické z hlediska efektivity. Efektinější a elegantnější by samozřejmě byl nástroj, který je implementován jedním programovacím jazykem[12].

Ale vzhledem k tomu, že je většina vývojářů projektů, které pracují na archivaci webu, rozsáhlá a nestejnorodá, není možné tento požadavek splnit. Zpravdila si každý řešitelský tým jednoduše implementuje své nástroje. Pokud je pak chce jiný tým přejmout a užívat, musí se s z toho plynoucímy problémy vypořádat.

Celkově jde ale říci, že použitím nástroje `Jpype` pro spojení nástrojů `warctools` a `JHOVE2` nevede k nějakému razatnímu zesložitění techniky migrace archivu, jelikož je nástroj je implementován efektivně. Navíc v podstatě neexistuje jiná varianta, tedy mimo implementace vlastních nástrojů, což by ale bylo velmi náročné, a snad i zbytečné, jelikož například nástroj `JHOVE2` je standardní, efektivní a velmi robusní nástroj.

Nástroj je dostupný volně ke stažení na webových stránkách[13] jeho vývojářů. Jeho využití bylo zásadní pro implementaci analýzy obsahu převáděných `arc` archivů. V `příloze AI na s. 66` je archiv s nástrojem `Jpype`.

4.3 Testování migračních nástrojů

Migraci realizuji pomocí několika variant nastrojů `WARC-TOOLS`. Pokusím se převést vzorek arc archivů a porovnat výsledky. Z informací, které získám pokusím poukázat na výhody a nevýhody jednotlivých nástrojů.

---

11. Webové stránky `Java` dostupné online na adrese `http://www.java.com`.
12. Toto tvrzení není nutně pravdivé. Samozřejmě existují softwarová řešení, která dosahují vyšší efektivity právě tím, že kombinují různé programovací jazyky. To však není případ nástroje, který je potřeba implementovat pro potřeby WebArchivu
13. Webové stránky projektu `Jpype` dostupná online na adrese `http://jpype.sourceforge.net`.

4.3.1 WARC-TOOLS

Sada nástrojů pro práci s `arc` a `warc` soubory. Skripty jsou napsány v několika programovacích jazycích – `C, Python, Ruby`. Součástí sady skriptů jsou jak skripty pro převod archivů, tak skripty pro analýzy warc souborů. Jedná se konkrétně o `warcdump` a `warcvalidator`. Bohužel `WARC-TOOLS` neobsahuje i nástroje na práci s warc archivy.

Tento nástroj není již v současnosti dále vyvíjen a podporován, ale je stále dostupný ke stažení[14]. Na webových stránkách projektu je dostupná dokumentace, kterou nabízím v `přílohách R, S, T a U na s. 63`. `Příloha O na s. 63` obsahuje výpis z programu `JHOVE2` pro `warc` archiv generovaný tímto nástrojem. `Příloha V na s. 64` obsahuje výstup z nástroje `warcdump` pro `warc` archiv. V `příloze L na s. 62` je `warc` archiv generovaný nástrojem `WARCTOOLS`. `Příloha AC na s. 65` obsahuje kompletní nástroj.

Kompatibilita s ostatními nástroji

Nástroj `warcdump` není kompatibilní s ostatními mnou testovanými nástroji.

- pro `WARC-TOOLS / hanzo` hlásí: `Incompatible Warc Version`
- pro `WARC-TOOLS / kpk09` hlásí: `Incompatible Warc Version`
- pro soubory typu `arc` hlásí: `Incompatible Warc Version`

Nástroj `warcvalidator` není kompatibilní s ostatními testovanými nástroji.

- pro `WARC-TOOLS / hanzo` hlásí: `Incompatible Warc Version`
- pro `WARC-TOOLS / kpk09` hlásí: `Incompatible Warc Version`
- pro soubory typu `arc` hlásí: `Incompatible Warc Version`

Nástroj tedy není přímo kompatibilní s ostatními mnou testovanými nástroji.

---

14. Webové stránky projektu `WARCTOOLS` dostupná online na adrese `http://code.google.com/p/warc-tools/`.

Další parametry

- Výhodou nástroje je, že má napsané skripty pro hromadnou migraci.
- Zvládá převádět soubory `arc` i soubory `arc.gz`.
- Při převodu `arc` souboru do `warc` archivu přibalí i archivované soubory. Na rozdíl od ostatních nástrojů – ostatní nástroje přibalí archivované soubory pouze pokud dostanou na vstup `arc` archivy, které jsou zabalené nástrojem `GZip`.
- Nástroj není v současné době dále vyvíjen. Z toho se dá odvodit, že pravděpodobně nebude možno používat ho dlouhodobě – ostatní nástroje budou obohacovány o další funkcionalitu a tím reagovat na nové požadované parametry k převodu `arc` souborů.
- Vývojáři ve specifikaci uvedli, že chtějí implemntovat možnost `deduplikace` souborů v archivu v průběhu migrace. [40] Bohužel tato funkcionalita není implementována.

4.3.2 `WARC-TOOLS / hanzo`

Další nástroj obsahující skripty pro převod `arc` souborů na `warc` soubory, včetně nástrojů pro validaci a práci s `warc` soubory. Tento nástroj vyvíjí tým, který je odvozen od týmu vyvíjejícího původní `WARC-TOOLS`[15].

V `příloze W na s. 64` je k nalezení dokumentace nástroje. `Příloha X na s. 64` obsahuje výstup z nástroje `warcdump` a `příloha Y na s. 64` výstup z nástroje `JHOVE2` pro daný `warc` soubor. `Příloha AD na s. 65` obsahuje kompletní nástroj.

Kompatibilita s ostatními nástroji

- Nástroj `warcdump` je kompatibilní s ostatními testovanými nástroji.
- Z nástroje zcela vychází nástroj `WARC-TOOLS / kpk09`.

15. Webové stránky projektu `WARC-TOOLS / hanzo` dostupné online na adrese `http://code.hanzoarchives.com/warc-tools/wiki/Home`.

- Dokumentace nástroje s nástrojem WARC-TOOLS / kpk09 je totožná.

Další parametry

- Nemá napsané skripty pro hromadnou migraci.
- Při převodu arc souborů nepřibalí k souboru data. Je potřeba mu dávat na vstup arc.gz soubory.
- Po otestování na zkušebním vzorku je textový warc soubor, totožný se souborem z WARCTOOLS / kpk09.
- Nástroj není schopen provést při migraci i deduplikaci souborů obsažených v archivu.
- Nástroj je v současné době vyvíjen.
- Nástroj nemá dokončenou dokumentaci. V manuálu stojí, že vytváří „crappy“ warc soubory. Nástroj pro migraci je údajně převzat z nástroje WARCTOOLS. [36]
- Zjistil jsem, že nástroj havaruje při převodu archivu se specifickým záznamem.

4.3.3 WARC-TOOLS / kpk09

Třetí mnou testovaný nástroj na určený pro migraci arc souborů. Jedná se o projekt, který je vyvíjen jako fork k původnímu kódu WARC TOOLS / hanzo[16]. V příloze Z na s. 65 je k nalezení dokumentace nástroje. Příloha AA na s. 65 obsahuje výstup z nástroje warcdump a příloha AB 65 výstup z nástroje JHOVE2 pro daný warc soubor. Příloha AE na s. 65 obsahuje kompletní nástroj.

Kompatibilita s ostatními nástroji

- Nástroj warcdump je kompatibilní s ostatními testovanými nástroji.

16. Webové stránky projektu WARC-TOOLS / kpk09 dostupné online na adrese https://bitbucket.org/kpk09/warc-tools/wiki/Home.

- Nástroj zcela vychází z nástroje `WARC-TOOLS / hanzo`.
- Dokumentace nástroje s nástrojem `WARC-TOOLS / hanzo` je totožná.

Další parametry

- Nemá napsané skripty pro hromadnou migraci.
- Při převodu `arc` souborů nepřibalí k souboru data. Je potřeba mu dávat na vstup `arc.gz` soubory.
- Po otestování na zkušebním vzorku je textový `warc` soubor, totožný se souborem z `WARCTOOLS / hanzo`.
- Nástroj není schopen provést při migraci i `deduplikaci` souborů obsažených v archivu.
- Nástroj je v současné době vyvíjen.
- Nástroj nemá dokončenou dokumentaci. V manuálu stojí, že vytváří „crappy“ `warc` soubory. Nástroj pro migraci je údajně převzat z nástroje `WARCTOOLS`. [37]
- Zjistil jsem, že nástroj havaruje, stejně jako `WARCTOOLS / hanzo`, při převodu archivu se specifickým záznamem.

4.4 Podobnost `WARCTOOLS / hanzo` s `kpk09`

Jak jsem zmínil již výše nástroje `WARCTOOLS / hanzo` s `WARCTOOLS / kpk09` jsou si velmi podobné. Tato podobnost logicky vychází z toho, že `WARCTOOLS / kpk09` vychází z nástroje druhého. Pokud bychom chtěli důsledně porovnávat tyto dva nástroje můžeme to udělat dvěma způsoby. První způsob je porovnání výstupů pro stejný vstup a druhý je porovnání zdrojového kódu.

4.4.1 Porovnání výstupů pro stejné vstupy

Testoval jsem nástroje pro převod na vlastním vzorku `arc` archivů. Konkrétně jsem testoval podobnost příslušných dvou `warc.gz` archivů

a jejich výstupů při použití nástroje `warcdump`. Výstupy z nástroje `warcdump` jsem porovnával pomocí nástroje `diff`[17]. Výsledky byli až na názvy cest k archivům totožné. Testování samotných `warc.gz` souborů jsem prováděl pomocí nástroje `zdiff`[18], i tento nástroj potvrdil totožnost souborů.

Musím zdůraznit, že jsem použil relativně malý testovací vzorek. Proto abych mohl uspokojivě konstatovat, bych musel použít mnohem větší vzorek dat. Navíc bych musel testovat archivy s různým obsahem – dá se očekávat, že by mohlo dojít k iferenciaci výstupů, kdyby archivy obsahovali atypické typy souborů.

4.4.2 Porovnání zdrojového kódu

Porovnání zdrojového kódu je náročnější procedura, ale pravděpodobně může jasně odpovědět na otázku. Pravděpodobně by bylo nejlepší dotázat se na rozdílnost obou nástrojů přímo vývojářů. Server, na kterém jsou oba nástroje umístěny, umožňuje přímé porovnávání pomocí nástroje `diff`. Informace o rozdílnosti je možné zhlédnout z dvou hledisek.

- Změny ve `WARCTOOLS / hanzo`, které se neprojevily v druhém nástroji[19].
- Změny ve `WARCTOOLS /kpk09`, které se neprojevily v druhém nástroji[20].

4.5 Závěry z porovnávání nástrojů

Na základě porovnání jednotlivých nástrojů, které rozvádím v předcházející kapitole, jsem po poradě se svým vedoucím, Mgr. Václavem

17. Informace o nástroji `diff` dostupné online na adrese `http://unixhelp.ed.ac.uk/CGI/man-cgi?diff`.
18. Informace o nástroji `zdiff` dostupné online na adrese `http://resin.csoft.net/cgi-bin/man.cgi?section=1&topic=zdiff`.
19. Náhled změn mezi porovnávanými nástroji dostupný online na adrese `https://bitbucket.org/kpk09/warc-tools/compare/hanzo/warc-tools`. 11.4.2012 šlo o modifikace ve 34 souborech.
20. Náhled změn mezi porovnávanými nástroji dostupný online na adrese `https://bitbucket.org/kpk09/warc-tools/compare/..hanzo/warc-tools`. 11.4.2012 šlo o modifikace ve 7 souborech.

Roseckým, došel k závěru, že nejlepším nástrojem pro případný převod webového archivu bude nástroj `WARCTOOLS / hanzo`.

Hlavní výhoda tohoto nástroje je to, že je neustále vyvíjen a za jeho vývojem stojí organizace Hanzo Archives.

Kapitola 5

Implementace úprav v migračních nástrojích

5.1 Úvod

Mým úkolem bylo najít způsob, jak při procesu migrace na `warc` archiv zjistit obsah původního archivu. To znamená, že jsem měl v průběhu migrace provést analýzu obsahu archivu. V průběhu převodu znamená taková analýza nejmenší režii, oproti dodatečnému provádění takové analýzy. Navíc by analýza mohla potencionálně ovlivnit samotný převod, jelikož by v případě nutnosti bylo možno změnit obsah převáděného archivu, před tím než bude uložen do `warc` archovu.

Z toho důvodu jsem musel provést několik změn v programu `JHOVE2` i nástroji `WARCTOOLS`, respektive jsem změnu provedl jen v nástroji `WARCTOOLS / hanzo`. Změna v ostatních nástrojích `WARCTOOLS` by znamenala obdobné zásahy do zdrojového kódu, a pravděpodobně by vedla je stejným výsledkům, jak vyplývá z mého porovnání nástrojů.

Úprava umožňuje během testování obsahu archivu provádět další operace. Při migraci archivu dochází k tomu, že před tím, než nástroj `arc2warc` vloží do nového `warc` archivu soubor z původního archivu, provede se analýza daného souboru pomocí nástroje `JHOVE2`.

V praxi bude tato varianta mého programu umožňovat provedení rozhodování o tom, zda daný soubor přidat do nového `warc` archivu, nebo ho vyloučit, nebo s ním umožní provádět další operace. Výstup migrace tak tvoří nový `warc` archiv a jednotlivé analýzy všech souborů obsažených v archivu.

5.2 Úprava JHOVE2

Program JHOVE2 je implementován v programovacím jazyce Java. Jelikož je napsán tak, aby ho bylo možné spouštět z příkazového řádku, musel jsem provést změny, který by mi umožnili spouštět jej přímo z těla programu WARCTOOLS. Bylo tedy nutno implementovat v rámci programu JHOVE2 další třídu.

Moje nová třída je spuštěna přímo z programu arc2warc.py[1]. Poté co inicializuje nástroj JHOVE2 předává mu pomocí metody runJHOVE2Loop jednotlivé soubory k testování. Mnou implementovaná třída tedy opět umožňuje testování více souborů, přičemž není nutné spouště JHOVE2 vícekrát.

Při implementaci jsem objevil ne příliš vážnou chybu v nástroji: nástroj vypisuje některé chybové hlášky standardní výstup. Chyba by byla závažná, pokud by se na ní nepřišlo a během provádění migračního nástroje by se chybové hlášky vypisovali do těla warc archivu. V takovém případě by mohlo dojít ke zbytečnému poškození archivu. Tuto chybu jsem tedy jednoduše odstranil přejměrováním výstupů do souboru.

Dále jsem zjistil, že nástroj při používání modulu XmlFormat zpracovává některé soubory neúměrně dlouho. Jednalo se o soubory, které obsahovali například data ve formě CDATA. Proto mi vedoucí práce, mgr. Václav Rosecký, doporučil daný modul odstranit. jeho absence nebude mít výrazný vliv na validování obsahu souborů.

Poslední problém, který jsem objevil u nástroje JHOVE2 je to, že v něm vzniká výjimka při testování archivů. Bylo kvůli tomu nutné upravit nástroj arc2warc.py, tak aby výjimky neohrozili běh migrace.

Ukázku této třídy je možno najít v příloze AU a na s. 69, kompletní implementovaný nástroj je k nalezení v příloze AW na s. 70.

5.3 Úprava WARCTOOLS / hanzo

Při úpravě nástroje WARCTOOLS stačilo upravit pouze arc2warc.py, který je implementovám v programovacím jazyce Python. Bylo nutné provést takové změny, aby bylo možno přímo při běhu programu spus-

1. Součást nástroje WARCTOOLS, která je určena k migraci archivů. Do jiných programů v rámci WARCTOOLS jsem nemusel zasahovat.

tit aplikaci v programovacím Java. A upravit program tak, aby bylo možno přistupovat k jednotlivým souborům uvnitř arc archivu. Dalším problémem bylo, že program JHOVE2 neumožňuje vícenásobné spuštění. Opakovaná inicializace by však byla výpočetně velmi náročná[2].

Pro účely spouštění Java aplikace v rámci rámci skriptu v jazyce Python bylo třeba vybrat vhodný nástroj. Vybíral jsem mezi nástroji Jpype a Jython[3]. Nástroj Jpype umožňuje spuštění Java Virtual Machine uvnitř skriptu v jazyce Python, oproti tomu nástroj Jython spustí Java Virtual Machine ještě před samotným vykonáváním skriptu. Nakonec jsem zvolil nástroj Jpype, který mi přišel vhodnější. Hlavně protože v době volby nástroje jsem se domníval, že budu potřebovat spouštět Java Virtual Machine vícekrát během průběhu skriptu.

Původní program převádí arc archiv tak, že předělá metadata uvnitř archivu a jeho obsah – soubory v něm obsažené – jako bytestream převede do nového warc archivu. Můj výsledný upravený program pracuje tak, že během převodu archivu provede pomocí regulárního výrazu filtrování jednotlivých souborů uvnitř archivu. Ty uloží do dočasného adresáře a volá na ně nástroj JHOVE2 z těla skriptu arc2warc předtím, než je provedeno zařazení souboru do nového archivu. Nakonec jsou soubory z dočasného adresáře odstraněny.

Upravený program pod názvem arc2warc_loop_jhove2.py nabízím v příloze AV na s. 70, kde jsou i moje komentáře k úpravám zdrojového kódu.

5.4 Výstupy z analýzy obsahu převáděných archivů

Své programy arc2warc_jhove2.py jsem otestoval při převodu několika souborů arc.gz, které jsem pomocí nástroje Heritrix sám sklidil. V příloze AP na s. 68 nabízím warc soubor, který vznil převodem testovacího arc.gz archivu z přílohy E na s. 59.

Jak se dalo předem očekávat migrace spojená s testováním obsahu archivu je časově velmi náročná. V následující tabulce podávám přehled výsledků testování. Je evidentní, že migrace bez analýzy archivu pomocí

2. Typický archiv je má velikost 100 MB a obsahuje stovky až tisíce souborů.
3. Webové stránky projektu JYTHON dostupné online na adrese http://www.jython.org.

nástroje JHOVE2 bude mnohem rychlejší.[4].

Ke zefektivnění nástroje by se pravděpodobně dalo dojít tak, že by se v rámci nástroje JHOVE2 implementovala třída, která by umožnila přímou charakterizaci souborů, bez nutnosti vytváření dočasných souborů na disku. Taková změna však není vzhledem k rozsahu nástroj JHOVE2 jednoduchá. V současné implementace funguje nástroj tak, že i když dostane na vstup bytestream, vytvoří z něj dočasný soubor. Nástroj JHOVE2 je vyvíjen týmem lidí, který měl zřejmě důvod takouto užitečnou funkcionalitu neimplementovat.

V následující tabulce uvádím časy na zpracování zpracování archivů a poznámky z testovaných archivů. Použil jsem různé archivy z repozitáře WebArchivu. Všechny měli 100 MB. U některých neuvádím čas validace vzniklého warc archivu. Jedná se o ty, které migrační nástroj nepřevedl správně – haváriím převodu se věnuji v následující kapitole.

Tabulka 5.1: Časy migrace souborů

Pokus	**Souborů**	WARCTOOLS	S JHOVE2	**validace**
1	3 682	31 s	2 m 50 s	10 s
2	4 045	32 s	2 m 25 s	12 s
3	6 382	31 s	2 m 10 s	x
4	5 754	38 s	2 m 5 s	13 s
5	10 372	42 s	3 m 9 s	x
6	11 681	36 s	3 m 16 s	x
7	7 472	29 s	2 m 40 s	12 s
8	10 113	22 s	3 m 0 s	x
9	8 307	33 s	2 m 39 s	x
Průměr				

4. Musím zdůraznit, že jsem testování prováděl na svém osobním počítači s parametry: Intel Core 2 Duo CPU T7100 @ 1.80GHz * 2, 2 GB RAM, Linux 3.0.0-19-generic 32-bit. Reálná čísla při použití nástroje na strojích WebArchivu by byla jistě rozdílná. Důležitý je zde pouze poměr času s užitím nástroje JHOVE2 a bez něj.

5.5 Havárie nástroje WARCTOOLS / hanzo

5.6 Porovnání užitých verzí nástroje WARCTOOLS / hanzo

5.7 Další nástroje sady WARCTOOLS / hanzo

Jak bylo řečeno již výše, WARCTOOLS / hanzo je sada nástrojů. Kromě migračního nástroje obsahuje i nástroje pro práci s warc archivy. Některé znich jsou pro projekt WebArchiv důleží. V následující kapitole o nich podám několik základních informací a doporučení.

5.7.1 warcindex.py

Tento nástroj slouží k vytvoření CDX indexu pro warc archivy. Jak jsem již zmiňoval, CDX indexy slouží k interpretaci dat z archivů pomocí nástroje Wayback. Nástroj warcindex.py doporučuji k vytváření CDX indexů warc souborů.

5.7.2 warcvalid.py

5.8 Doporučení pro implementaci migrace archivu

Kapitola 6

Závěr

Ve své bakalářské práci jsem se zabýval migračními nástroji pro převod webových archivů v rámci projektu WebArchiv. Konkrétně pro převod archivů typu `arc` na archivy typu `warc`.

V první kapitole práce podávám základní informace o projektu WebArchiv, věnuji se starším pracím týkajících se projektu a nakonec představuji software, který je při realizaci projektu užívaný.

Ve druhé kapitole práce představuji některé možné archivační formáty. První formát je již zastaralý a v projektu se nevyužívá. Druhý formát, formát `arc`, je stále používán. Ukazuji tedy jeho výhody a nevýhody pro projekt. Poslední část kapitoly věnuji formátu nejnovějšímu formátu, formátu `warc`.

Tento formát je v rámci projektu užíván k ukládání nových sklizených dat. V minulosti sklizená data jsou však v repozitáři uložena ve formátu předcházejícím. Stejně jako u předchozích formátů ukazuji jeho výhody a nevýhody. Z kapitoly by mělo být zřejmé, že přechod na nový formát je potřeba realizovat – jednotlivé argumenty jsou k dohlednání v příslušné části práce.

Poslední kapitola je věnována migračním nástrojům. Mým cílem bylo najít a otestovat nástroje, pomocí kterých by bylo možno převést starší archivy do nového formátu. V kapitole ukazuji jaké další nástroje bude nutné použít, aby byla realizace migrace možná. Důležitý je zejména nástroj `JHOVE2`, který slouží k testování obsahu archivů. Při převodu bude totiž třeba odhalit možné chyby v uložených souborech a odstranit je. Dále je tedy v této kapitole obsažen i popis mé implementace testování obsahu archivů.

Pro samotnou migraci jsem si vybral tři volně dostupné nástroje a otestoval je na vzorku uložených dat. Z testování mi vyšel jako nejvhodnější nástroj `WARCTOOLS / hanzo`. Jeden nástroj jsem vyloučil protože

je již zastaralý a není dále vyvíjen. Druhý protože vychází z prvního a za jeho vývojem nestojí tak silná autorita jako v případě mnou zvoleného nástroje. I přesto, že jsem nástroj `WARCTOOLS / hanzo` zvolil jako nejvhodnější, jsem v nástroji odhalil chybu: nástroj havaruje při převodu specifických souborů v archivu. V kapitole též ukazuji o jaké soubory se může jednat.

Výstup práce tedy tvoří doporučení nástroje `WARCTOOLS / hanzo` k migraci dat, analýza jeho rozdílnosti a výhod oproti jiným testovaným nástrojů a analýza případů, kdy dochází k jeho haváriím.

Informace z mé bakalářské práce použije Národní knihovna ČR (zřizovatel projektu WebArchiv) jako podklady k realizaci migrace webového archivu.

Literatura

[1] MASANÉS, JULIEN (Ed.): *Web Archiving*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN 3-540-23338-5.

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

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

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

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

[6] 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`.

[7] 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`.

[8] 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.

[9] STOKLASOVÁ, Bohdana. *Budování vzájemně kompatibilních informačních systémů pro přístup k heterogenním informačním zdrojům a jejich zastřešení prostřednictvím Jednotné informační brány*. 2005. Dostupné z: http://webarchiv.cz/files/dokumenty/zpravy/Zamer2005Zpravatext.doc.

[10] CELBOVÁ, Ludmila. *Ochrana a trvalé zpřístupnění webových zdrojů jako součásti národního kulturního dědictví*. 2006. Dostupné z: http://webarchiv.cz/files/dokumenty/zpravy/zprava-VaV_2006-final.rtf.

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

[12] WEBARCHIV. *Zpráva WebArchiv – obnova dat – 2007*. Brno, 2007. Dostupné z: https://docs.google.com/Doc?docid=0AbRV47jJIQggZG5qOHJtZF8yNmRjanM1dg&hl=cs.

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

[14] 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.

[15] WEBARCHIV. *Zpráva WebArchiv – VISK – 2009*. Brno, 2009. Dostupné z: https://docs.google.com/Doc?docid=0AbRV47jJIQggZG5qOHJtZF80NngybjZ3aGY&hl=cs.

[16] WEBARCHIV. *Zpráva WebArchiv – Věda a výzkum – 2010*. Brno, 2010. Dostupné z: https://docs.google.com/Doc?docid=0AbRV47jJIQggZG5qOHJtZF8xOGZ4YzNoamRy&hl=cs.

[17] WEBARCHIV. *Zpráva WebArchiv – VISK – 2010*. Brno, 2010. Dostupné z: https://docs.google.com/Doc?docid=0AbRV47jJIQggZG5qOHJtZF8xN2Q2cnpxcmZm&hl=cs.

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

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

[20] 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.

[21] 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.

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

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

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

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

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

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

[28] 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.

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

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

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

[32] 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%29.

[33] 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%20080618.doc.

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

[35] 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`.

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

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

[38] ResourceIndex configuration options. INTERNET ARCHIVE. *Wayback* [online]. 2012 [cit. 2012-04-20]. Dostupné z: `http://archive-access.sourceforge.net/projects/wayback/resource_index.html`.

[39] DeDuplicator. NATIONAL AND UNIVERSITY LIBRARY OF ICELAND. *DeDuplicator* [online]. 2010 [cit. 2012-04-20]. Dostupné z: `http://deduplicator.sourceforge.net`.

[40] Hanzo Archives Limited. *WARC Tools Phase III Functional – Requirements Specification* [online]. 2009 [cit. 2012-04-20]. Dostupné z: `http://www.google.cz/url?sa=t&rct=j&q=&esrc=s&source=web&cd=19&ved=0CIMBEBYwCDgK&url=http%3A%2F%2Fwarc-tools.googlecode.com%2Ffiles%2Fwarc-tools_phase_III_frs_v8.pdf&ei=zk2QT__jNcPitQbSq-ihBA&usg=AFQjCNGbvOMqJiLAu5H4GbpZNYjGfHCZtA&sig2=fPbSa8zHvySi8ZZkI75F7Q`

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. NIC

AL. NIC

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
```

Navazující záznam

```
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>>
```

AP. Warc soubor vytvořený z přiloženého arc souboru pomocí mnou upraveného nástroje WARC-TOOLS / HANZO

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

AQ. NIC

AR. NIC

AS. NIC

AT. NIC

AU. `JHOVE2` – vlastní implementace nástroje – varianta pro krokové testování každého souboru v archivu

Celý soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `RunFromARC2WARCLoop.java`, zde ukazuji pouze metodu třídy (runJHOVE2Loop), která obstarává předávání jednotlivých souborů k testování.

```
   /**
* Metoda dá JHOVE2 na vstup cestu k souboru a cestu kam má
* uložit výstup z analýzy.
* @param sourcePath Cesta k vstupnímu souboru
* @param outputPath Cesta kam se má uložit analýza
* @throws JHOVE2Exception
* @throws IOException
*/
public void runJHOVE2Loop(String sourcePath,
   String outputPath) throws JHOVE2Exception, IOException
{
   // vytvoreni source z vstupního souboru
   Source source = (FileSource) jhove2.getSourceFactory().
      getSource(jhove2, sourcePath);
   //charakterizace
   jhove2 = (JHOVE2) jhove2.getModuleAccessor().
      startTimerInfo(jhove2);
   Input input = source.getInput(jhove2);
   try {
      source = jhove2.characterize(source, input);
```

```
   } finally {
      if (input != null) {
         input.close();
      }
   }
   jhove2 = (JHOVE2) jhove2.getModuleAccessor().
      endTimerInfo(jhove2);
   //vystup výstup
   displayer.display(source, outputPath);
}
```

AV. `WARCTOOLS / hanzo` – vlastní implementace nástroje – varianta pro krokové testování každého souboru v archivu

Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `arc2warc_loop_jhove2.py`. Aby správně fungoval, je potřeba mít nainstalován nástroj Jpype (`příloha AI na s. 66`) a umístit soubor do kořenové složky WARCTOOLS / hanzo (`příloha AD na s. 65`).

AW. `JHOVE2` – vlastní implementace nástroje

Archiv s upraveným nástrojem `JHOVE2`. Soubor je dostupný jako elektronická příloha bakalářské práce pod názvem `jhove2-2.0_martinprokop.tar.gz`.