FORMÁTOVÁ ANALÝZA SKLIZENÝCH DAT V RÁMCI PROJEKTU WEBARCHIV NK ČR

File Format Recognition of Data Harvested by Web Archiving Project of National Library of the Czech Republic

Jaroslav Kvasnica
Národní knihovna České republiky

Rudolf Kreibich
Národní knihovna České republiky

Recenzenti:
Mgr. Pavla Švástová
Mgr. Eva Měřínská

Abstrakt:
V současné době Národní knihovna České republiky začala ukládat shromážděná data z archivu českého webu do úložiště dlouhodobé ochrany digitálních dokumentů. Článek se věnuje výstupu projektu Institucionální vědy a výzkumu, který má za cíl vytvořit plán pro retrospektivní analýzu souborových formátů nad celým webovým archivem a zmapovat nástroje, které tuto identifikaci provádějí. Podrobná znalost archivovaných dat umožní jejich kontrolu, která poskytne možnost vytvořit budoucí strategii jejich dlouhodobé ochrany. V neposlední řadě výstupy analýzy mohou vést ke zlepšení podmínek zpřístupnění archivovaných dat koncovému uživateli.

***Klíčová slova**: souborové formáty, web archiv, dlouhodobá ochrana digitálních dokumentů, Heritrix, archivace, Národní digitální knihovna, ARC, WARC*

Abstract:
National Library of the Czech Republic just begun to ingest harvested data from web archiving project into Long-term Preservation System. This article is output of Institutional Science and Research project aiming to implement retrospective file format recognition framework for harvested data and map tools related to file format recognition. Precise knowledge of archived data is cornerstone for building Long-term Preservation Strategy. Such analysis may also improve conditions of end-user access.

***Keywords:** file formats, web archive, long term preservation, Heritrix, archiving, National digital library, ARC, WARC*

Úvod

V současné době Národní knihovna České republiky spolu s Moravskou zemskou knihovnou pracují na vytvoření Národní digitální knihovny. Jedním z hlavních pilířů tohoto projektu je

úložiště dlouhodobé ochrany digitálních dokumentů (dále jen LTP úložiště), které "poskytne prostor pro bezpečné umístění dosud digitalizovaných dokumentů i digitálních dokumentů vytvořených či získaných v projektu NDK i v rámci dalších projektů"[1]
Jedním z těch projektů je také WebArchiv. "WebArchiv je digitální archiv českých webových zdrojů, které jsou zde shromažďovány za účelem jejich dlouhodobého uchování."[2] WebArchiv již od roku 2000 archivuje celý český web. Českým webem v kontextu WebArchivu jsou myšleny české domény, webové zdroje v českém jazyce, od českých autorů nebo vztahující se nějakým způsobem k České republice nebo českému národu. WebArchiv k archivaci přistupuje třemi způsoby. Jednak archivuje plošně celou českou doménu, dále realizuje výběrové sklizně, u kterých jsou vybírány zdroje kurátory. A třetím způsobem je výběr tematických sklizní, které se vztahují k nějaké významné události (např. povodně, volby atd.).

WebArchiv využívá ke stahování webových stránek nástroj zvaný Heritrix. Heritrix je tzv. crawler, tedy nástroj sloužící ke sklízení webu pro archivní účely, které vytvořilo americké sdružení Internet Archive. S tím souvisí i formáty, do kterých je stažený webový obsah uložen. Využívají se tzv. kontejnerové formáty, které jsou nazvány ARC a WARC. Do nich jsou ukládány soubory všech typů (vč. obrázků, videí nebo třeba zdrojových kódů), které se crawleru podařilo stáhnout. Soubory navíc crawler opatří metadaty s informacemi o průběhu jejich stahování. Veškeré soubory jsou bezztrátově komprimovány.

ARC je předchůdcem formátu WARC. Byl vytvořen v roce 1996, sdružením Internet Archive, které potřebovalo archivní formát, kam by mohlo ukládat agregovaná data sklizená z webu.[3] Protože byl ARC průkopníkem mezi archivními formáty, tak měl několik nedostatků, zejména pak metadata nebyla ideálně strukturována. Proto vznikla revize, která dostala název WARC. WARC přinesl lepší podporu sběru dat, snadnější přístup k datům a podporu výměny dat mezi organizacemi.[4] S WARCem přišla i nová verze Heritrixe, která používá WARC jako primární formát pro sklizená data. Od začátku letošního roku přešel český WebArchiv na tuto novou verzi, ale při práci s archivem je potřeba počítat s tím, že data od roku 2000 jsou ve starším formátu ARC.

Dlouhodobá ochrana dat z WebArchivu

Jak jsme již na začátku článku naznačili, tak veškerá data z WebArchivu budou uložena v novém digitálním úložišti dlouhodobé ochrany digitálních dokumentů. V první řadě je třeba vyjasnit, proč vlastně WebArchiv není soběstačný při archivaci dat. Do dnešní doby WebArchiv ukládal data pouze na úrovni běžné zálohy. Tím rozumíme uložení dat v několika kopiích, ale bez dlouhodobější strategie. Tuto formu archivace můžeme označit za krátkodobou.

---

[1] HUTAŘ, Jan. Podrobnější popis projektu NDK a jeho kontext. NÁRODNÍ KNIHOVNA ČR. *Národní digitální knihovna* [online]. 13.12.2011. Praha [cit. 2013-06-21]. Dostupné z: http://ndk.cz/narodni-dk/podrobnejsi-popis-projektu
[2] Co je WebArchiv?. *WebArchiv: archiv českého webu* [online]. [cit. 2013-06-21]. Dostupné z: http://www.webarchiv.cz/
[3] ARC_IA: Internet Archive ARC file format. *Sustainability of Digital Formats: Planning for Library of Congress Collections* [online]. 2008-02-14, 04-Apr-2013 [cit. 2013-06-25]. Dostupné z: http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml
[4] WARC: Web ARChive file format. *Sustainability of Digital Formats: Planning for Library of Congress Collections* [online]. 2009-08-31, 04-Apr-2013 [cit. 2013-06-25]. Dostupné z: http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

Zatímco dlouhodobá ochrana digitálních dat "hledá cesty, jak čelit nebo předcházet zastarávání či selhávání hardware, datových nosičů a zastarávání software nebo formátů souborů"[5]. Když mluvíme o archivaci dlouhodobé, máme na mysli desítky až stovky let. Není zde úkolem pouze zaručit, aby se neztratila žádná data, ale "rovněž zajistit technickou použitelnost a sémantickou srozumitelnost archivovaných digitálních dokumentů"[6].

LTP úložiště není jen technickým problémem, ale musí mít vyřešené i "zajištění organizace, financování, kvalifikovaný personál a efektivní řízení"[7]. Provozovat dlouhodobé úložiště je velmi časově, finančně i personálně náročné, a proto není možné, aby si WebArchiv budoval své vlastní úložiště dlouhodobé ochrany a stal se tak plnohodnotným dlouhodobým archivem.

Aktuální stav

V současné době Národní knihovna končí přípravy pro ukládání dat z WebArchivu do LTP úložiště. Dlouhodobá ochrana těchto dat je nyní zabezpečena na logickou úroveň kontejnerových formátů. Dnes to, co je schováno uvnitř kontejnerů, nevíme. Důvodem jsou nedostatečné kapacity výkonu na to, abychom byli schopni udělat kompletní analýzu všech souborů uvnitř archivních souborů (Více v samotné analýze).

Kvůli nedostatečné výpočetní kapacitě jsme v první fázi zvolili přístup, kdy jsme se zaměřili na zajištění provenience a autenticity a na ochranné aktivity na úroveni zabalených dat v kontejnerových formátech. Stejně tak jsme na tuto oblast zaměřili metadatový popis, který kompletně mapuje události, při kterých vznikaly jednotlivé sklizně, kontext a akce, které s těmito daty byly provedeny. Část metadat vytváří sám crawler Heritrix ve formě logů a další část bude vytvářena ve workflow LTP před uložením na úložiště.

Přestože náš přístup, tedy kombinace bitové ochrany a metadatového popisu na úroveň kontejnerů, je nejlepší v rámci našich současných možností, je třeba myslet do budoucnosti. Protože "v souvislosti s rostoucí komplexitou a diverzitou (...) formáty představují závažnější problémy než elektronické nosiče."[8] Proto je důležité vědět, jaké souborové formáty jsou obsaženy uvnitř českého webového archivu. Tato potřeba iniciovala vznik naší analýzy. Největší problém představuje rozsah dat, proto má analýza pomoci připravit strategii pro budoucnost. Problémy technického prostředí, zejména elektronických nosičů, jsou již vyřešeny v rámci projektu NDK.

Cíle analýzy

Hlavním cílem analýzy bylo zmapování možností retrospektivní identifikace formátů kompletního archivu webových stránek až do roku 2000. Zejména jsme chtěli odpovědět na otázky, jak časově náročná by identifikace byla a jakou metodiku zvolit. Jestli archiv zmapovat již před uložením do LTP úložiště nebo identifikaci odložit.

---

[5] HUTAŘ, Jan, Marek MELICHAR a Bohdana STOKLASOVÁ. Národní digitální knihovna. *Knihovna*. 2009, roč. 20, č. 1, s. 6-21.

[6] Tamtéž.

[7] ROSENTHAL, Colin, Asger BLEKINGE-RASMUSSEN a Jan HUTAŘ. *Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER)*. 1. vyd. Praha: Národní knihovna České republiky, 2009, 51 s. ISBN 978-807-0505-694.

[8] CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010, 154 s. ISBN 978-80-7050-588-5.

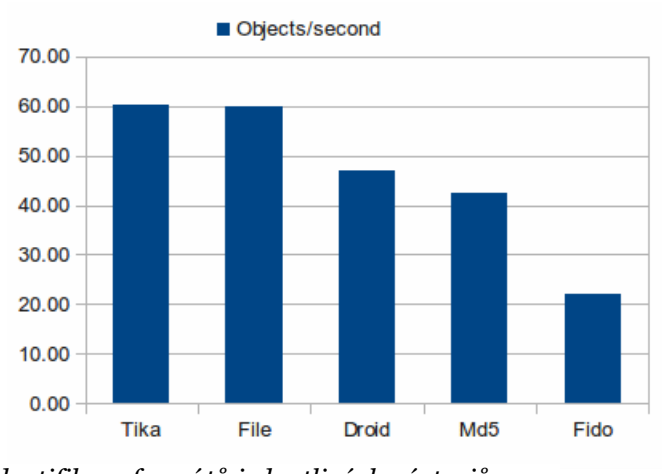
Dalším cílem bylo nastavení workflow, které by mohlo být aplikováno na nově vznikající sklizně. Tím by se zátěž na výkon rozložila v čase a nebyla zahlcena IT struktura jednorázovými extrémními úkony. Zároveň jsme chtěli na zvoleném vzorku nahlédnout do rozložení souborových formátů na českém webu.
Ještě před samotnou analýzou je potřeba se seznámit se základními pojmy, se kterými pracujeme. V první řádě se jedná o pojem long tail, který v obecném pojetí představuje mocninné rozdělení, kdy "první část křivky obsahuje jednotky s velkou frekvencí výskytu, jak postupujeme dále, frekvence výskytu se zmenšuje, ale jednotek přibývá"[9]. U analýzy formátů jde o to, že relativně malý počet souborových formátů představuje drtivou většinu celého archivu a zbytek, který nazýváme právě long tail, tvoří velký počet formátů s malým výskytem. U našeho vzorku tvoří 98 % archivu 7 formátů a zbylá 2 % pak 82 různých formátů. Zmíněných 7 formátů nazýváme formáty dominantními.

Nástroje pro identifikaci souborových formátů

Celé analýze formátu předcházelo zmapování dostupných nástrojů pro identifikaci formátů. Hned ze začátku jsme vyřadili populární nástroj JHove[10], protože je to zároveň nástroj validační a kvůli tomu je jeho práce časově náročnější. Opět kvůli časové náročnosti jsme vyřadili i FITS[11], což je komplexní nástroj složený z různých aplikací. Při dalším výběru jsme vycházeli z evaluační studie, kterou provedla Open Planets Foundation[12]. Hlavními kritérii pro nás byla rychlost, rozsah rozpoznatelných formátů a úspěšnost identifikace. Nakonec jsme se rozhodovali mezi nástroji: Droid[13], Fido[14] a Apache Tikka[15].

Naším záměrem bylo vybrat dva nástroje, abychom mohli ověřit i úspěšnost identifikace pomocí porovnání jednotlivých výsledků. Protože ve studii vyšly jako nejrychlejší Droid a Tikka (viz Graf 1) přiklonili jsme se právě k těmto nástrojům. Na grafu lze vidět i porovnání s Unixovým program File[16], který jsme ale shledali jako nespolehlivý, protože identifikuje formát pomocí obsahu souboru a není napojen na žádnou databázi s formáty.

[9] ZBIEJCZUK, Adam. Long Tail (dlouhý chvost). *WEB 2.0: charakteristiky a služby* [online]. červen 2007 [cit. 2013-06-25]. Dostupné z: http://zbiejczuk.com/web20/03-5-long-tail-dlouhy-chvost.html
[10] http://sourceforge.net/projects/jhove/
[11] https://code.google.com/p/fits/
[12] BLEKINGE, Asger Askov. Identification tools, an evaluation: The Scape Characterisation Tool Testing Suite. OPEN PLANETS FOUNDATION. *Open Planets Foundation: A community hub for digital preservation*[online]. 23 February 2012 [cit. 2013-06-25]. Dostupné z: http://www.openplanetsfoundation.org/blogs/2012-02-23-identification-tools-evaluation
[13] http://digital-preservation.github.io/droid/
[14] http://www.openplanetsfoundation.org/software/fido
[15] http://tika.apache.org/
[16] HAAS, Juergen. Linux / Unix Command: file. *About.com: Linux* [online]. 2013 [cit. 2013-06-25]. Dostupné z: http://linux.about.com/library/cmd/blcmdl1_file.htm



Graf 1 Rychlost identifikace formátů jednotlivých nástrojů

Zatímco Droid je napojen na registr formátů Pronom[17], který obsahuje více než 800 různých souborových formátů[18]. Navíc Droid vykázal poměrně velkou úspěšnost v identifikování formátů z long tailu. Dalším argumentem pro byl fakt, že jeho vývoj je velmi dynamický a dnes je již v šesté verzi. Proto jsme jako primární nástroj pro identifikaci vybrali právě jej. Jako druhý nástroj pro porovnání výsledků jsme zvolili Apache Tikka, který sice není napojen na žádnou databázi, ale ve studii byl nejrychlejší a byl nejvíce úspěšný v identifikování 20 nejběžnějších formátů[19].

Metodika

Celý postup se skládal ze tří kroků. Nejprve bylo třeba vyextrahovat veškerá data sklizených webových stránek z kontejnerů WARC. Jak jsme již zmiňovali výše, tak crawler vše, co stáhne, ukládá do kontejnerového formátu WARC resp. ARC. Pokud bychom spustili identifikační nástroj přímo nad sklizní, tak by její výstup obsahoval pouze identifikaci souborového formátu WARC. Druhý krok byla identifikace formátů a třetím pak zpracování výstupu z identifikačního nástroje.

Veškeré testování se provádělo na počítačích v Národní knihovně. Pro identifikaci formátů jsme využili 8 GB RAM paměti a běžný dvoujádrový procesor. Více výkonu jsme si pro testování nemohli dovolit alokovat, neboť bychom přespříliš zatížili celou IT infrastrukturu knihovny. Pro srovnání ve výše zmiňované studii Open Planets Foundations využili 70 GB RAM.

Při výběru dat byla pro nás nejdůležitějším kritériem reprezentativnost, ale zároveň jsme si byli vědomi nedostatku výkonu a času, kterým jsme disponovali. Z těchto důvodů jsme zvolili jako základní jednotku vzorku jednu sklizeň, která představuje ukončený proces stahování všech webových stránek podle určitých kritérií. WebArchiv dnes má tři typy sklizní: tematické, celoplošné a pravidelné.

My jsme si zvolili sklizeň pravidelnou, protože obsahuje průřez českým webem, ale v menším rozsahu než celoplošné a není tolik monotónní jako sklizně tematické. Konkrétně jsme vybrali

[17] http://www.nationalarchives.gov.uk/PRONOM

[18] The technical registry Pronom: about. THE NATIONAL ARCHIVES. *The National Archives* [online]. 2013 [cit. 2013-06-25]. Dostupné z: http://www.nationalarchives.gov.uk/aboutapps/PRONOM/default.htm

[19] BLEKINGE, Asger Askov. Identification tools, an evaluation: The Scape Characterisation Tool Testing Suite. OPEN PLANETS FOUNDATION. *Open Planets Foundation: A community hub for digital preservation*[online]. 23 February 2012 [cit. 2013-06-25]. Dostupné z: http://www.openplanetsfoundation.org/blogs/2012-02-23-identification-tools-evaluation

dubnovou sklizeň z roku 2013, která měla zabalená velikost 100 GB a její kontejnerový formát už byl nový formát WARC. Jednalo se o menší sklizeň. Běžně se sklizně pohybují v jednotkách terabajtů. Jak se později ukázalo, formát kontejnerů (WARC nebo ARC) nebyl z hlediska formátů až tolik důležitý (viz níže).

1. Extrakce

Na extrakci jsme využili nástroj zvaný jwat-tools ve verzi 0.5.6. SNAPSHOT. Jedná se o standardní nástroj, který v Národní knihovně běžně využíváme a máme s ním pozitivní zkušenost. Po rozbalení všech kontejnerů celková velikost dat vzrostla na cca 350 GB. To nám přišlo neúměrně mnoho a nebyli jsme si jistí, jestli bychom byli v současné situaci schopní s tak rozsáhlým balíkem dat efektivně pracovat. Proto jsme se rozhodli použít pouze prvních 100 GB.

Díky nástroji jwat-tools, který umí pracovat s ARC i s WARC, není důležité, v jakém formátu jsou data uložena. Myšleno pouze z hlediska pracovního workflow identifikace formátů, jinak se sklizně budou lišit v počtu metadatových objektů a výstup bude odlišný. Nakonec nám tedy zbylo 100 GB rozbalených dat stažených z českého webu. Což bylo před rozbalením celkem 41 WARC souborů. Extrakce celé sklizně pomocí nástroje jwat-tools trvala 16 hodin a 47 minut.

2. Identifikace

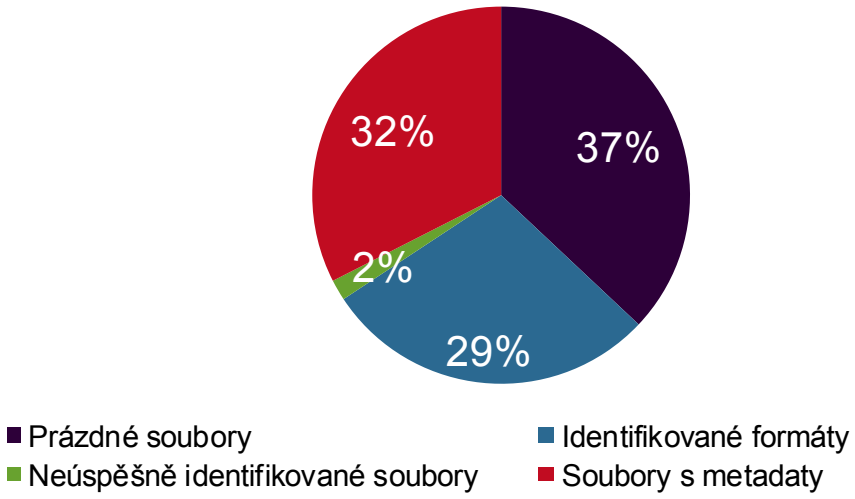
Druhý krok byla samotná identifikace formátů. Jak uvádíme výše, zvolili jsme Droid a Apache Tikka. Droid ve verzi 6.1 a Apache Tikka ve verzi 1.3, obě verze byly v době testování nejaktuálnější. V průběhu testování jsme zjistili, že Tikka je na našem hardware natolik pomalá, že náš vzorek by se identifikoval v řádech týdnů, tak jsme nakonec zůstali jen u Droidu.

Droid 100 GB dat zvládl za 14 hodin. Přestože byl řádově rychlejší než Tikka, tak naši kompletní testovací sklizeň, která je na poměry WebArchivu jedna z těch menších, by kompletně identifikoval více než dva dny.

3. Práce s výstupem

Další problém, který se objevil, až v průběhu analýzy, představoval výstup, který vygeneroval Droid. Droid vytvořil tzv. profil, který měl v komprimované formě 700 MB a pokud se profil extrahoval např. do CSV souboru, pak měl velikost 1,5 GB. Tak obrovský soubor nebyl schopný zpracovat žádný tabulkový procesor. Proto jsme museli celý výstup upravit, tak abychom s ním mohli dále pracovat.

Agregaci a úpravu dat jsme provedli pomocí programovacího jazyku AWK v systému Unix. Díky tomu jsme získali výstup, který měl odpovídající velikost a obsahoval pouze informace, které jsme potřebovali. Ještě před samotnou agregací jsme museli roztřídit soubory podle základních typů, které jsou znázorněny v Grafu 2.



Graf 2 Rozložení digitálních objektů

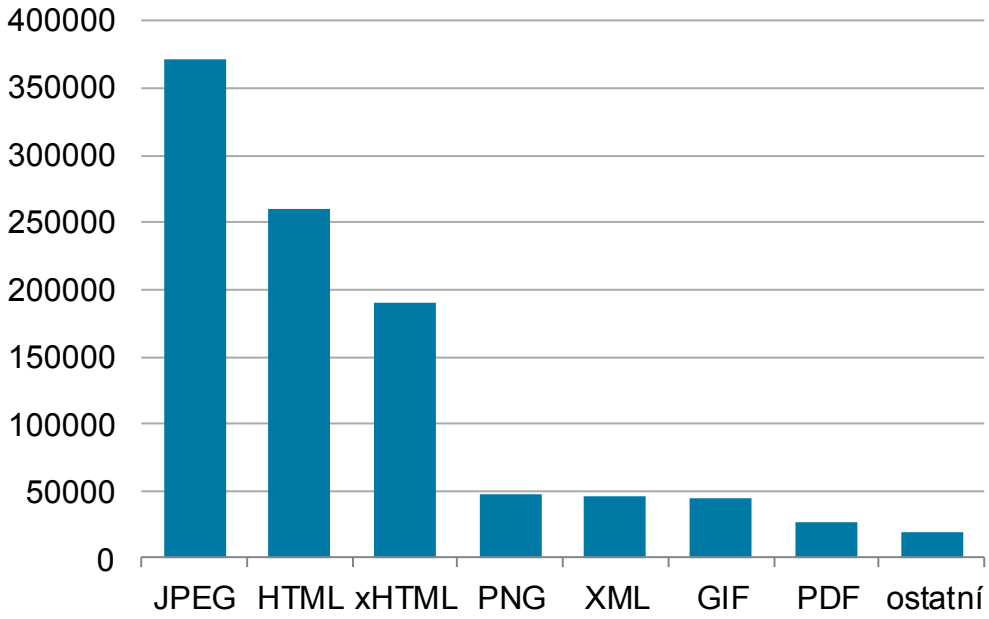
Výsledky analýzy

Celá identifikace formátů byla poměrně úspěšná. Droid nedokázal rozpoznat formát jen u 2 % digitálních objektů. Velkou část objektů tvořily prázdné soubory (soubory, které měli 0 B). U těchto souborů jsme nebyli schopni přesně určit, kdy vznikly. S největší pravděpodobností to jsou metadatové soubory, které vznikly a neobsahují žádnou informaci (např. zaznamenání http komunikace, která neměla žádný obsah). Jejich původ může být u samotného Heritrixe nebo u nástroje pro extrakci jwat-tools.

Další velkou část pak tvořily metadatové soubory, které vytváří crawler Heritrix při samotném stahování webu. A poslední část tvoří objekty, u kterých Droid dokázal identifikovat jejich souborový formát. Celkem 100 GB extrahovaných dat obsahovalo přes 3,5 milionu digitálních objektů a to vč. prázdných souborů.

Po zjištění, kolik digitálních objektů tvoří soubory, u kterých se podařilo identifikovat jejich formát, jsme se zaměřili právě na tuto množinu dat. Jak se ukázalo, český web se příliš neliší od světového internetu. Podle Michala Daye zkušenosti světových webových archivů ukazují, že většinu povrchového webu tvoří relativně malé množství formátů.[20]

[20] DAY, Michal. The Long-Term Preservation of Web Content. MASANÈS, Julien. *Web archiving*. Online-Ausg. New York: Springer, c2006, s. 177-199. ISBN 3540233385-.

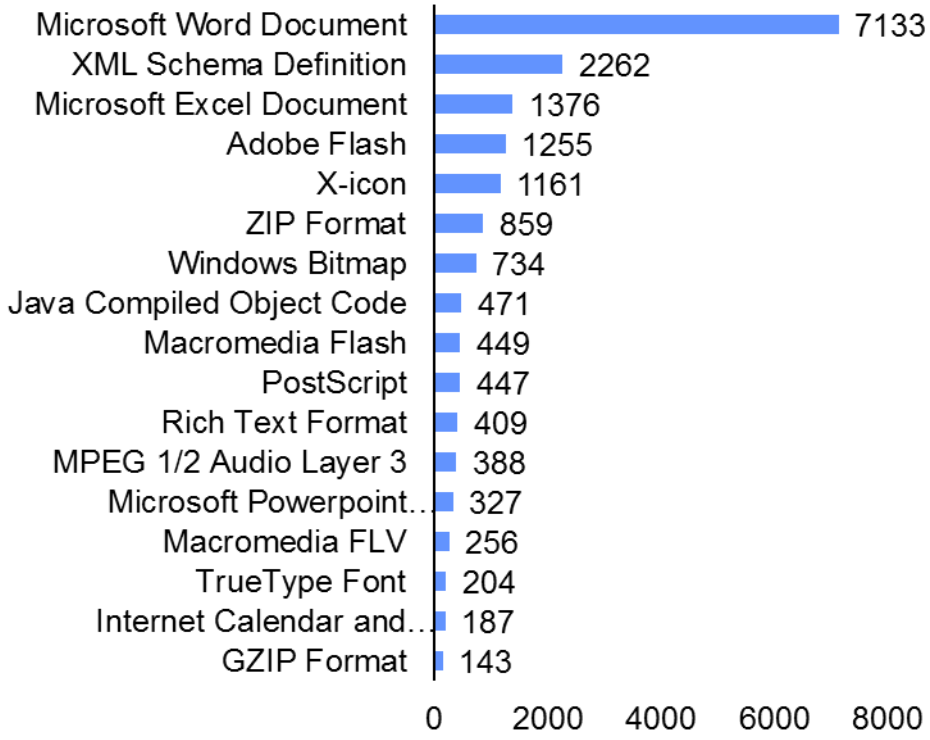


Graf 3 - Dominantní formáty

Jak je vidět na Grafu 3, drtivou většinu českého webu tvoří 7 dominantních formátů a to konkrétně celých 98 %. Jako dominantní označujeme formáty, které mají na webu několikanásobně větší zastoupení než ostatní formáty. V grafu můžeme vidět, že i nejslabší dominantní formát (v našem případě PDF) má téměř dvakrát větší zastoupení než všechny zbylé formáty dohromady. Tyto zbylé formáty pak tvoří 2 % z identifikovaných objektů. Těch je celkem 82 různých druhů. Nesmíme zapomenout na další 2 % digitálních objektů z předchozího grafu, u kterých není známo, v jakém jsou formátu. Určitě se mezi nimi nacházejí i poškozené soubory, přesto můžeme předpokládat, že se jedná o další desítky až stovky více či méně obskurních souborových formátů.

Ještě je potřeba doplnit, že jsme stejné souborové formáty, které byly zastoupeny v různých verzích, brali jako jeden. Například JPEG je v databázi PRONOM (ze které Droid vychází) zastoupen v pěti různých verzích a všechny verze se objevily v našich datech, ale my jsme všechny považovali za stejný formát. K tomuto jsme přistoupili jednak pro větší přehlednost a také, aby se méně časté verze běžných formátů neobjevovaly v long tailu. Pokud bychom na základě analýzy chtěli přistoupit např. k migraci, pak je nutné počítat s různými verzemi jednoho formátu. Nicméně pro účely analýzy je vhodnější agregace verzí formátů.

Když se podíváme podrobněji na long tail, pak dvě třetiny formátů jsou zastoupeny méně než stokrát. Výraznější první třetinu formátů můžeme vidět na Grafu 4. Přestože v long tail dominuje Microsoft Word Document, tak oproti nejméně častému dominantnímu formátu je ho téměř pětkrát méně



Graf 4 Nejčastější formáty v long tail

Závěr

Analýza ukázala, že výkon, který jsme schopni v současné době alokovat pro práci s archivem webových stránek, není dostatečný pro celkovou retrospektivní analýzu souborových formátů. Bohužel také ukázala, že není dostatečný ani pro identifikaci nově vznikajících sklizní. Proto v tuto chvíli nejsme schopni nastavit workflow pro průběžnou identifikaci. Časová a výkonová náročnost znemožňuje identifikovat formáty ještě předtím, než budou data uložena do LTP úložiště. Již není možné nadále zdržovat příjem dat z WebArchivu do LTP kvůli riziku jejich ztráty a nemožnosti je nahradit.

Zároveň jsme ale zjistili, jaké má český web dominantní souborové formáty a že 2 % formátů tvoří long tail. Díky tomu bychom mohli do budoucna zlepšit zpřístupnění archivu koncovému uživateli (např. extrakcí textu, obrázků apod.).

V budoucnosti bychom se chtěli zaměřit na identifikaci formátů v celém archivu, protože díky tomu budeme moci nastavit nejlepší strategii dlouhodobé ochrany a nebudeme se muset spoléhat jen na bitovou ochranu, která je v tuto chvíli jediná možná. Musíme si uvědomit, že formáty na internetu také podléhají stárnutí. A je možné, že už dnes některé soubory v archivu z roku 2000 nebudeme schopni přečíst.

Znalost souborových formátů v archivu samozřejmě nevyřeší všechny problémy a otázky, jak ochránit data z webu, ale rozhodně může pomoci v dalším vývoji účinné strategie. Již dnes je jasné, že nás čeká rozhodnutí, jakou cestou se vydáme, jestli zvolíme migraci, emulaci nebo kombinaci

obou přístupů. Povědomí o formátech nám také může ukázat zajímavé informace o samotném českém webu - jaký obsah český web obsahoval a jak se měnil v čase.

Seznam použité literatury

1. ARC_IA: Internet Archive ARC file format. *Sustainability of Digital Formats: Planning for Library of Congress Collections* [online]. 2008-02-14, 04-Apr-2013 [cit. 2013-06-25]. Dostupné z: http://www.digitalpreservation.gov/formats/fdd/fdd000235.shtml
2. BLEKINGE, Asger Askov. Identification tools, an evaluation: The Scape Characterisation Tool Testing Suite. OPEN PLANETS FOUNDATION. *Open Planets Foundation: A community hub for digital preservation*[online]. 23 February 2012 [cit. 2013-06-25]. Dostupné z: http://www.openplanetsfoundation.org/blogs/2012-02-23-identification-tools-evaluation
3. Co je WebArchiv?. *WebArchiv: archiv českého webu* [online]. [cit. 2013-06-21]. Dostupné z: http://www.webarchiv.cz/
4. CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vyd. Praha: Národní knihovna České republiky, 2010, 154 s. ISBN 978-80-7050-588-5.
5. *DAY, Michal. The Long-Term Preservation of Web Content. MASANÈS, Julien. Web archiving. Online-Ausg. New York: Springer, c2006, s. 177-199. ISBN 3540233385-.*
6. HAAS, Juergen. Linux / Unix Command: file. *About.com: Linux* [online]. 2013 [cit. 2013-06-25]. Dostupné z: http://linux.about.com/library/cmd/blcmdl1_file.htm
7. HUTAŘ, Jan, Marek MELICHAR a Bohdana STOKLASOVÁ. Národní digitální knihovna. *Knihovna*. 2009, roč. 20, č. 1, s. 6-21.
8. HUTAŘ, Jan. Podrobnější popis projektu NDK a jeho kontext. NÁRODNÍ KNIHOVNA ČR. *Národní digitální knihovna* [online]. 13. 12. 2011. Praha [cit. 2013-06-21]. Dostupné z: http://ndk.cz/narodni-dk/podrobnejsi-popis-projektu
9. *WARC: Web ARChive file format. Sustainability of Digital Formats: Planning for Library of Congress Collections [online]. 2009-08-31, 04-Apr-2013 [cit. 2013-06-25]. Dostupné z: http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml*
10. *ROSENTHAL, Colin, Asger BLEKINGE-RASMUSSEN a Jan HUTAŘ. Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER). 1. vyd. Praha: Národní knihovna České republiky, 2009, 51 s. ISBN 978-807-0505-694.*
11. *The technical registry Pronom: about. THE NATIONAL ARCHIVES. The National Archives [online]. 2013 [cit. 2013-06-25]. Dostupné z: http://www.nationalarchives.gov.uk/aboutapps/PRONOM/default.htm*
12. *ZBIEJCZUK, Adam. Long Tail (dlouhý chvost). WEB 2.0: charakteristiky a služby [online]. červen 2007 [cit. 2013-06-25]. Dostupné z: http://zbiejczuk.com/web20/03-5-long-tail-dlouhy-chvost.html*