Profilování kolekce a stanovení určené skupiny WebArchivu Národní knihovny ČR

Collection Profilling and Defining the Designated Community in Web Archiving Project of National Library of the Czech Republic

Jaroslav Kvasnica

Národní knihovna ČR, oddělení archivace webu
Studia nových médií, Filozofická fakulta, Universita Karlova, Praha

Barbora Bjačková

Národní knihovna ČR, oddělení archivace webu

Recenzenti:

Mgr. Andrea Fojtů
Mgr. Marek Melichar

Abstrakt:
V příspěvku jsou popsány kroky, které povedou k vytvoření strategie dlouhodobé ochrany dat sklizených z internetu v rámci projektu WebArchiv, která je nutným předpokladem k jejich ochraně v řádu desítek let. Tato strategie obsahuje pravidla a postupy, které zajistí dlouhodobou použitelnost a přístup archivních dat. V současné době je datům sklizeným z internetu v rámci projektu WebArchiv poskytována pouze ochrana na úrovni bit-streamu, která není v delším časovém horizontu dostatečná. Článek pojednává nejen o dosud provedených krocích v Národní knihovně (analýza formátů), na základě kterých bude vytvořen profil kolekce, ale i o krocích budoucích jako je definování určené skupiny, implementaci některých z plánovacích nástrojů a evaluaci potenciálních ochranných řešení.

***Klíčová slova**: WebArchiv, archivace webu, profil kolekce, určená skupina, dlouhodobá ochrana digitálních dokumentů*

Abstract:
Current bit stream data preservation practice in Web Archive of National Library of the Czech Republic is insufficient in long term. Presented paper describes steps leading to the Long Term Preservation policy for our collection. It introduces strategies, best practices and actions required for long term usability and accessibility. It also mentions accomplished actions such file format analysis, enabling us to create the collection profile. Among planned steps is designated community description, planning tools implementation and evaluation of preservation solutions.

Keywords: *WebArchiv project, web archiving, collection profile, designated community, long term preservation*

Tento článek vznikl díky podpoře MK ČR na rozvoj Národní knihovny České republiky jako výzkumné organizace.

1. Úvod

Článek navazuje na dříve publikovanou studii s názvem Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR[1]. Cílem této studie bylo zmapování možností retrospektivní identifikace formátů webových stránek archivovaných v rámci projektu WebArchiv.
V současné době jsme díky projektu Národní digitální knihovna schopni zajistit bitovou ochranu dat na úrovni archivních formátů a zajistit jejich autenticitu pomocí metadat. "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."[2] Bitová ochrana představuje pouze první krok k dlouhodobé ochraně dat, která "se nejčastěji realizuje redundantním uložením a řízeným kopírováním na nové nosiče"[3], jedná se tedy pouze o ochranu na fyzické resp. technické úrovni. Pro zajištění autenticity dat jsou používána metadata ve formátu PREMIS, která jsou implementována při ukládání do úložiště dlouhodobé ochrany NK ČR.

Abychom byli schopni zajistit ochranu dat v archivu na vyšší úrovni než doposud, je třeba provést několik kroků, které povedou k vypracování podrobné strategie ochrany. Vybudování takové strategie (ang. *preservation policy)* je velmi náročný úkol, který představuje výzvu do dalších let. Strategie dlouhodobé ochrany pro webový archiv se stane v budoucnosti částí stragie úložiště dlouhodobé ochrany NK ČR. Bez této strategie nejsme schopni zajistit, aby data uložená v archivu, mohla být využita v budoucnosti. Archiv musí umět reagovat na vývoj technologií, software a hardware, ale také na nově vznikající potřeby uživatelů, pro které jsou data určena.

S tím souvisí první krok, který se nazývá anglickým termínem *archive profiling* nebo také *collection profiling*. Tímto termínem je myšlena analýza nad celým archivem a na jejím základě vypracování profilu dat, která jsou uložena v archivu. Dalším krokem je definování tzv. *designated community*, tedy určené skupiny uživatelů, pro který je obsah archivu určen. Velmi často se zapomíná na to, že veškerá aktivita dlouhodobé ochrany je určena k tomu, aby v budoucnosti uživatelé mohli k uloženým datům přistupovat a pracovat s nimi. Nemá smysl archivovat data, která nebudou použitelná a budou uložena jen pro archivaci samotnou.

[1] KVASNICA, Jaroslav a Rudolf KREIBICH. Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR. *ProInflow* [online]. 31.10.2013 [cit. 21.07.2014]. ISSN 1804–2406. Dostupný z: http://pro.inflow.cz/formatova-analyza-sklizenych-dat-v-ramci-projektu-webarchiv-nk-cr
[2] Tamtéž jako 1
[3] KVAŠOVÁ, Zuzana a Tomáš SVOBODA. Dlouhodobá ochrana elektronických publikací. *ProInflow* [online]. 31.10.2013 [cit. 2014-07-21]. ISSN 1804–2406. Dostupný z: http://pro.inflow.cz/dlouhodoba-ochrana-elektronickych-publikaci

2. Archive profiling

Definice a vymezení

Archive/collection profilling nebo také vytvoření profilu kolekce či archivu je základním stavebním kamenem pro práci s archivem, který je nehomogenní a není možné u něj plně kontrolovat formu archivovaného obsahu. Pro webové archivy obecně platí, že jejich obsah je determinován svým původem, objevuje se v nich nezměrné množství souborových formátů, data jsou mezi sebou nahodile propojena pomocí hypertextových odkazů. Profil takového archivu je pak "sada charakteristik, která popisuje obsah archivu. Profil je obecný popis na nejvyšší úrovni. Tento popis shrnuje obsah webového archivu."[4]

Takový profil pak poskytne kurátorům ucelenou představu o archivu, který spravují. Znalost charakteristiky archivu je pro kurátory jedním z hlavních předpokladů pro zavádění strategií dlouhodobé ochrany. Profil ale také slouží pro uživatele, kteří se mohou lépe rozhodnout, jaký archiv bude sloužit k naplnění jejich informačních potřeb. V první fázi jsme pro vytvoření profilu začali s charakterizací souborových formátů.

Stanovení profilu

S charakterizací souborových formátů jsme začali v roce 2013. Výsledky jsme publikovali na konci října roku 2013. Ukázalo se, že Národní Knihovna ČR dnes nedisponuje dostatkem výpočetního výkonu, pro provedení kompletní charakterizace veškerých dat z WebArchivu. Ta by za současné situace zabrala až desítky let.

Po konzultaci s kolegy z Rakouské národní knihovny (dále jen ÖNB) a z Vídeňské technické univerzity (dále jen TUW), kteří se zabývají problematikou *collection profiling* se ukázalo, že vhodné řešení by mohlo být vytvoření reprezentativního vzorku kolekce dat. Reprezentativní vzorek by v první fázi mohl představovat pouze 1 % všech dat, v druhé fázi 10 % atd. Vše závisí na dostupnosti výpočetního výkonu a časové náročnosti.

Kolegové z TUW nás také upozornili, že pro účely vytvoření kompletního profilu naší kolekce, nestačí pouze formátová identifikace, ale je nutné provést podrobnou charakterizaci, která je sice časově náročnější, ale přináší metadata, bez kterých není možné profil kolekce vytvořit.

Velmi názorný příklad, proč je důležitá kompletní charakterizace, uvedl na svém blogu Petar Petrov[5]. V případě identifikace formátů dostaneme výsledky, které obsahují pouze název formátu a jeho verzi (viz obrázek 1). U těchto výsledků by se mohlo zdát, že první soubor a druhý soubor jsou totožné a mělo by se s nimi takto pracovat.

[4] ALSUM, Ahmed, Michele C. WEIGLE, Michael L. NELSON a Herbert VAN DE SOMPEL. Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* [online]. 2014 [cit. 2014-07-15]. DOI: 10.1007/s00799-014-0118-y. Dostupný z: http://link.springer.com/10.1007/s00799-014-0118-y

[5] PETROV, Petar. To fits or not to fits. *Open Planets Foundation* [online]. 2012 [cit. 2014-07-21]. Dostupný z: http://openplanetsfoundation.org/blogs/2012-07-27-fits-or-not-fits

Property	File A	File B	File C
Format	PDF 1.2	PDF 1.2	PDF 1.4

Obr. 1 Identifikace formátů

V případě použití nástroje pro kompletní charakterizaci výsledky přináší zcela odlišný pohled (viz obrázek 2). Zde vidíme, že ačkoliv mají první dva soubory stejný souborový formát i stejnou verzi, tak první soubor je nevalidní a zároveň šifrovaný, zatímco druhý soubor je validní i bez šifrování. Z uvedených příkladů je zřejmé, že ačkoliv se první dva soubory z počátku jevily jako totožné, podrobná charakterizace ukazuje opak. Se šifrovanými soubory není možné pracovat stejně jako s nešifrovanými a proto se v našem příkladu více podobají první soubor s třetím. Z toho plyne, že bez přijatelné charakteristiky digitálních objektů nemáme dostatečné informace pro jejich dlouhodobou ochranu.

Property	File A	File B	File C
Format	PDF 1.2	PDF 1.2	PDF 1.4
Page Count	20	20.000	40
Encryption	Yes	No	Yes
File Size	1 MB	120 MB	2 MB
Valid	No	Yes	No
Well-formed	Yes	Yes	Yes

Obr. 2 Charakterizace formátů

V druhé fázi jsme se zapojili do studie[6], na které pracoval tým ze Stanfordské univerzity, který vyvíjí aplikaci Memento agregující data z různých webových archivů. Tato studie nám nejen přinesla srovnání s vybranými zahraničními webovými archivy, ale také nám pomohla definovat charakteristiky, které by měl profil našeho archivu obsahovat.

[6] ALSUM, Ahmed, Michele C. WEIGLE, Michael L. NELSON a Herbert VAN DE SOMPEL. Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* [online]. 2014 [cit. 2014-07-15]. DOI: 10.1007/s00799-014-0118-y. Dostupný z: http://link.springer.com/10.1007/s00799-014-0118-y

Ve studii byly definovány tyto charakteristiky[7]:

1. **Age** (stáří archivu) - staří archivu se počítá od prvního uloženého objektu. Nemusí se rovnat oficiálnímu začátku archivu. Například český web archiv má první sklizené objekty z roku 2001, ale pravidelné sklizně začaly až v roce 2005.

2. **Top-level Domain** (domény nejvyššího řádu) - tato charakteristika představuje poměr domén nejvyššího řádu (např. .cz, .com, atd.) ve sklizených datech. U českého webového archivu bude dominovat doména .cz, neboť je na ni český archiv specificky zaměřen

3. **Language** (jazyk) - jazyková charakteristika, která udává poměr zastoupených světových jazyků. Bohužel v současné době nejsme schopni tento poměr uvést, neboť metriku, která je použita ve studii nelze v našem případě aplikovat. Metrika ve studii je založena na komparaci webových archivů, kterým je předkládán vzorek webových zdrojů, u kterých je znám jejich jazyk a na základě reakce archivů je určen poměr zastoupení jazyka.

4. **Grow rate** (tempo růstu) - tempo růstu udává kapacitní nárůst (v našem případě v řádech terabytů) za roční období.

   Tyto čtyři základní charakteristiky jsme doplnili o další, které spolu dohromady tvoří kompletní profil našeho archivu. Počet položek nemusí být konečný a v budoucnu se obsah profilu může měnit. Profil by neměl být zakonzervovaný dokument, ale pravidelně rozvíjený a aktualizovaný dokument.

5. **Frekvence sklízení** - tato charakteristika udává, jak často webový archiv sklízí svoje zdroje.

6. **Hloubka sklízení -** počet objektů sklizených v rámci jedné sklizně jedné webové stránky. Díky kapacitním a výkonovým omezením musí být hloubka sklízení omezena, v opačném případě by se mohlo stát, že veškerou kapacitu zahltíme pouze několika rozsáhlými stránkami, které obsahují obrovské množství objektů, typicky inzertní servery, aukční portály, zpravodajské servery apod.

7. **Přístupnost -** tento údaj udává, jaká část archivu je zpřístupněna veřejně a jaká pouze v rámci instituce nebo vůbec

8. **Software** – názvy aktuální verze aplikací, které jsou využívány ke sklízení a zpřístupňování archivu uživateli.

9. **Formáty -** jaké souborové formáty obsahuje archiv a jaké je jejich procentuální zastoupení.

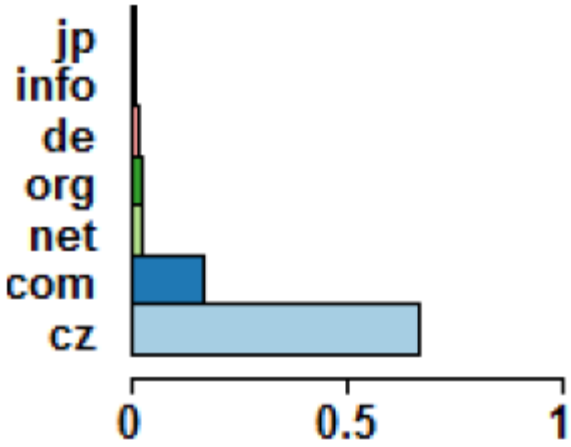
10. **Respektování souboru robot.txt** - charakteristika udává, zda archiv při sklízení respektuje nebo nerespektuje soubor robot.txt, který obsahuje instrukce pro roboty vyhledávačů. Každý archiv k tomu problému přistupuje jinak. Jedni zastávají názor,

[7] Tamtéž jako 5

že by měly být respektovány požadavky vlastníka stránky a sklízecí robot archivů je na stejné úrovni jako roboti vyhledávačů. Další skupina zastává názor, že webové archivy se snaží uchovat web, takový jaký je pro uživatele a ne tak jak jej vidí roboti.

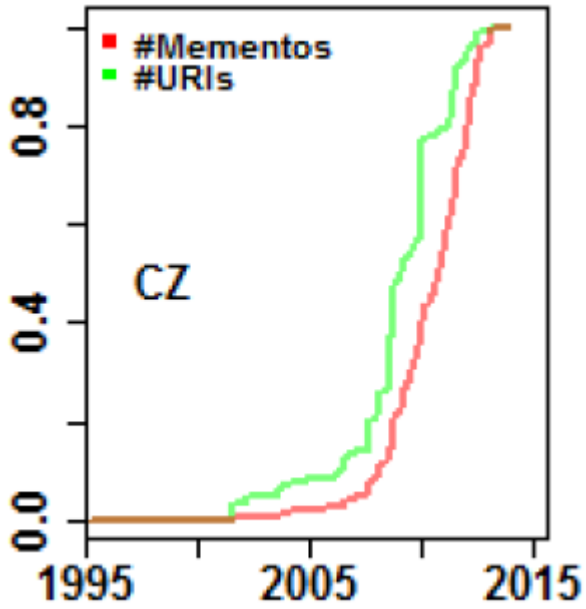
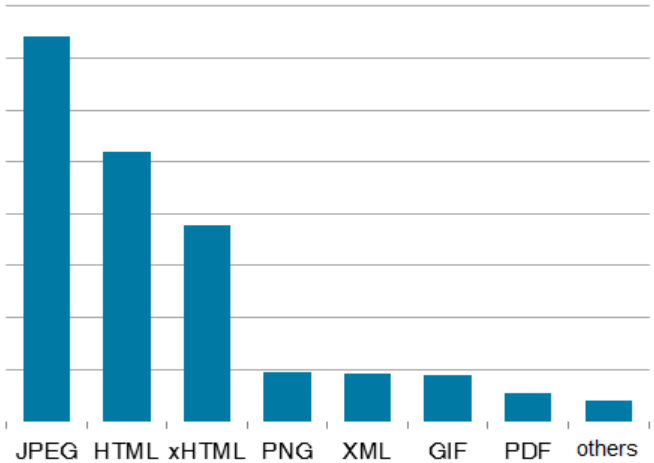
Návrh profilu

V tabulce níže je návrh profilu českého webového archivu. Tento návrh vychází z dat stanfordské studie, z našich interních statistik a z dat provedené charakterizace souborových formátů. Metodika, kromě formátové analýzy, byla kompletně přejata ze standforské studie[8]. Výpočet tempa růstu byl vypočítán nad kompletním archivem a poměr top-level domén na reprezentativním vzorku, stejně jako zastoupení souborových formátů.
V prvním obrázku v tabulce (obrázek 3) je graf, který znázorňuje poměr domén nejvyššího řádu v českém archivu. V dalším obrázku (obrázek 4) je graf znázorňující tempo růstu českého archivu v letech 1995-2014, kde zelená křivka znázorňuje počet archivovaných URI[9] a červená křivka znázorňuje počet archivovaných objektů.

Charakteristika	
Stáří archivu	3. 9. 2001
Top-level domény	 **Obr. 3 Top-level domény**

[8] ALSUM, Ahmed, Michele C. WEIGLE, Michael L. NELSON a Herbert VAN DE SOMPEL. Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* [online]. 2014 [cit. 2014-07-15]. DOI: 10.1007/s00799-014-0118-y. Dostupný z: http://link.springer.com/10.1007/s00799-014-0118-y

[9] URI (celým názvem Uniform Resource Identifier – „jednotný identifikátor zdroje“) je textový řetězec s definovanou strukturou, který slouží k přesné specifikaci zdroje informací

Tempo růstu	 **Obr. 4 Tempo růstu**
Frekvence sklízení	comprehensive: 1x/year; selective 1x|2x|6x|12x/year
Hloubka sklízení	comprehensive: 5000 objects; selective: 10000-15000 objects
Přístupnost	comprehensive: in house; selective more than 4000 URLs online access
Software	Heritrix Engine 3.1.2-SNAPSHOT-20130207.001528 Wayback-1.5.3-SNAPSHOT
Formáty	 **Obr. 5 Přehled formátů**
Robot.txt	Don't respect

Tab. 1 Návrh profilu archivu

V návrhu profilu je v poli pro souborové formáty pouze graf s jejich poměrným zastoupením. Pro podrobnou charakterizaci formátů budeme využívat aplikaci C3PO[10], kterou vyvíjí kolegové z TUW. Tato aplikace slouží ke zpracování metadat pocházejících z charakterizace souborových formátů. Pomocí webové aplikace pak zobrazuje tyto zpracovaná metadata ve formě, kdy s nimi lze lehce manipulovat a dále s nimi pracovat. Aplikace umožňuje vypsat souborové formáty obsažené v kolekci, jejich verze či software, kterým byly soubory vytvořeny. C3PO umožňuje jít i hlouběji a skládat dotazy, tím umožní vypsat např. všechny konkrétní verze PDF souborů vytvořené určitým programem atd.

Aplikaci C3PO máme v plánu zpřístupnit i pro badatele z řad široké veřejnosti. Tato přístupná verze bude obsahovat reprezentativní vzorek dat a měla by být přístupna na přelomu roku 2014/2015.

3. Určená skupina

Definice a vymezení

Před vydáním českého překladu normy OAIS se pro výraz *designated community* v české literatuře objevovaly různé termíny jako je uživatelská komunita (Fojtů), předem určená komunita (Giaretta), cílová komunita (Hutař), definovaná komunita (Wikipedie), popřípadě byl do českého jazyka překládán jako designovaná komunita (Melichar, Hutař). Český překlad normy OAIS tyto termíny sjednotil pod pojmem určená skupina.
Termín určená skupina je tedy jedním z mnoha pojmů, které přinesl vznik normy OAIS. Tato norma představuje referenční model otevřeného archivačního informačního systému, popisuje jeho architekturu – základní prvky a procesy, jejich vlastnosti a vztahy mezi nimi. Určená skupina je tedy definována v rámci referenčního modelu OAIS jako stanovená skupina možných koncových uživatelů, kteří by měli být schopni porozumět konkrétní množině informací. Určená skupina je vymezena daným archivem a toto vymezení se může časem měnit. Určená skupina se také může skládat z několika uživatelských komunit[11].
Určená skupina je důležitým prvkem pro naplňování závazných požadavků na model OAIS. Jedním ze základních úkolů důvěryhodného úložiště dle modelu OAIS je zajištění toho, že uchovávané informace budou srozumitelné samy o sobě (*independently understandable*) určené skupině. Dalším souvisejícím požadavkem je také stanovení skupin, které se stanou určenými skupinami a definování jejich znalostní základny a v neposlední řadě také zpřístupnění uchovávaných informací určené skupině[12].
Refenční model OAIS vymezuje určenou skupinu jako specifickou část uživatelů, která dokáže samostatně porozumět archivované informaci tak, jak je uchována a prezentována v rámci OAIS. Je tedy rozdíl mezi koncovými uživateli archivu a mezi jeho určenou skupinou. Uživateli archivu tak mohou být všechny osoby, které mají přístup k uchovávaným datům, zatímco jeho určenou skupinou jsou ti uživatelé, kteří dokáží na základě svých znalostí či schopností dané informace využít i bez pomoci odborníka[13]. Některé archivy ovšem mohou mít pouze jednu určenou skupinu, například každého uživatele knihovny, nebo naopak několik různých skupin, kdy může mít každá jiné informační potřeby a požadavky na archiv.

Stanovení určené skupiny

Definování rozsahu určené skupiny je klíčovým aspektem pro jakoukoliv dlouhodobou ochranu dat založenou na modelu OAIS.

[10] http://ifs.tuwien.ac.at/imp/c3po
[11] ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model.* Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, Srpen 2014.
[12] Tamtéž jako 1
[13] LAVOIE, Brian F. *The Open Archival Information System Reference Model: Introductory Guide* [online]. OCLC : Dublin, Ohio, 2004 [cit. 2014-7-14]. 20 s. Dostupný z: http://www.dpconline.org/docs/lavoie_OAIS.pdf

Je třeba si uvědomit, že rozsah a vlastnosti určené skupiny ovlivňují jak obsah archivu, tak i formu, ve které jsou informace uchovávány, tak aby mohly být určenou skupinou samostatně interpretovány. Není tak vhodné stanovit rozsah určené skupiny až na základě obsahu či formy dat v archivu[14].
S určenou skupinou je také spjat koncept vysvětlující informace (*representation information*), což je informace či soubor informací, který umožňuje pomoci pochopit obsah a smysl dokumentu. Například při uložení informace na disku CD-ROM je daná informace reprezentována bity a vysvětlující informace dává způsob, jak tyto bity přečíst. Vzhledem k požadavku na srozumitelnost uchovávaných informací pro určenou skupinu je podstatné porozumět znalostní základně (*knowledge base*) dané určené skupiny ke stanovení minimální vysvětlující informace, která musí být zachována. Znalostní základna určené skupiny se však může časem měnit, což by mělo vést k úpravě vlastností vysvětlující informace tak, aby byla stále zajištěna srozumitelnost obsahu daných informací[15].
Koncept určené skupiny je také významný pro navrhování a vytváření důvěryhodných archivů. Report RLG a OCLC[16] specifikuje důvěryhodný archiv jako takový, který poskytuje určené skupině spolehlivý, dlouhodobý přístup ke spravovaným digitálním zdrojům v současnosti i v budoucnosti[17]. Dokument CCSDS[18] zabývající se auditem a certifikací důvěryhodných digitálních archivů obsahuje doporučení a příklady k definování určené skupiny a její znalostní základny. Jako příklady určené skupiny uvádí anglicky mluvící veřejnost se středoškolským a vyšším vzděláním a přístupem k webovému prohlížeči (kompatibilnímu s HTML 4.0), astronomy s přístupem k programu se systémem FITS (Flexible Image Transport System), kteří dokáží pracovat s astronomickými spektrografickými nástroji a další[19].
Stanovení určené skupiny je také přínosné při vytváření strategie daného archivu, například je možné definovat různá přístupová práva na základě jednotlivých kategorií určených skupin.

Specifika cílové komunity v prostředí webových archivů

Přesto, že byl referenční model OAIS původně navržen jako součást iniciativy pro podporu dlouhodobého uložení dat získaných z vesmírných projektů, dále byl rozvíjen a nyní je chápán jako všeobecný rámec, který umožňuje definovat prvky a fungování organizace osob a systémů, která přijala zodpovědnost za uchovávání informací a jejich zpřístupňování své určené skupině. Model OAIS je tedy možné aplikovat na různé typy dlouhodobého uložení digitálních dat, včetně webových archivů[20].
Na data, která webové archivy uchovávají a mohou poskytovat své určené skupině, je možné nahlížet ve dvou rovinách. Tou běžnější je chápání archivace jako zobrazování informací z webu ve formě prohlížení uchovaných verzí jednotlivých webových stránek či dokumentů. Druhou rovinou je pak vnímání těchto dat jako celku obsahujícího souhrnné informace a různé souvislosti.
Specifickým problémem při plánování dlouhodobého uchovávání a zpřístupňování informací z webového archivu je dynamická povaha webu a množství typů dokumentů a jejich formátů, které web obsahuje. Množství formátů dnes již není možné zobrazit se současným technickým vybavením, příkladem mohou být softwarové aplikace, jako jsou počítačové hry,

[14] Tamtéž jako 3
[15] LAKSHMI, V a S. C JINDAL. *Digital libraries*. Delhi: Isha Books, 2004. ISBN 81-820-5109-6.
[16] Research Library Group a Online Computer Library Center
[17] *ČSN ISO 16363. Systémy pro přenos dat a informací z kosmického prostoru - Audit a certifikace důvěryhodných digitálních úložišť. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, Říjen 2014.*
[18] The Consultative Committee for Space Data Systems
[19] CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (CCSDS). *Audit and Certification of Trustworthy Digital Repositories* [online]. 2011 [cit. 2014-7-14]. Dostupný z: http://public.ccsds.org/publications/archive/652x0m1.pdf
[20] DAY, M. The Long-Term Preservation of Web Content. In: MASANÈS, J., ed. *Web archiving*. New York: Springer, c2006, s. 177-199. ISBN 978-3-540-23338-1.

které vyžadují určité technické parametry na zobrazovací zařízení pro jejich bezproblémovou srozumitelnost. I přes možnosti migrace dat nebo vytváření emulátorů znamená rychlé zastarávání technologií pro množství elektronických zdrojů definitivní ztráty.
Z hlediska této nehomogenní typové a formátové struktury webu je třeba při plánování archivace a dlouhodobé ochrany webových zdrojů provádět formátovou analýzu a také na jejím základě vytvářet plán dlouhodobého uchovávání a zpřístupnění s ohledem na určenou skupinu. Pro zachování požadavku na archiv dle OAIS musí být data uchovávaná a zpřístupňovaná webovým archivem srozumitelná pro stanovenou určenou skupinu. Pro zachování srozumitelnosti je tedy klíčová technická specifikace archivovaných objektů a popisné informace obsažené v metadatech. Z hlediska metadat je pro budoucí srozumitelnost webových zdrojů podstatné zmínit především kontextové údaje, které zejména ve webovém prostředí poskytují vazby na další zdroje.
Druhým významným rysem webového prostředí je jeho dynamická povaha. Web se neustále mění a vyvíjí, což je také nutné brát na vědomí při plánování strategie dlouhodobé ochrany a zpřístupnění webových zdrojů.

Návrh stanovení určené skupiny WebArchivu NK ČR

WebArchiv Národní knihovny ČR do současné doby neměl vytvořenou definici určených skupin a za cílovou uživatelskou skupinu byli považováni všichni uživatelé, tj. návštěvníci webových stránek a uživatelé vyhledávající archivované informace prostřednictvím počítačového terminálu v budově knihovny.
Vzhledem k významu konceptu určené skupiny pro uchovávání informací v rámci modelu OAIS a k plánu vypracování strategie WebArchivu byla stanovena potřeba definování určené skupiny WebArchivu.
Navrhovaný soubor určené skupiny obsahuje tři kategorie:

1. Individuální uživatelé
2. Institucionální uživatelé
3. Výzkumníci a vědci

První a největší kategorii tvoří individuální uživatelé. Vzhledem k tomu, že každý uživatel přistupuje do archivu s vlastním individuálním požadavkem, je možné chápat každého uživatele jako badatele. Touto kategorii určené skupiny se tedy rozumí obecná veřejnost s přístupem k internetu a webovému prohlížeči. Z hlediska požadavku na srozumitelnost informací uložených v archivu je cílem WebArchivu uchovávat a zpřístupňovat uchované informace v takové formě, aby byla co nejvíce podobná formě informací nacházejících se v daném okamžiku na tzv. živém webu.
Institucionálními uživateli se rozumí takové instituce, které potřebují a využívají data z webového archivu pro svou činnost. Takovými institucemi může být například policie, soudy, výzkumné ústavy atd. Specifikem těchto uživatelů je možnost získání dat z archivu na základě odůvodněného písemného požadavku. Mezi institucionální uživatele mohou také patřit provozovatelé počítačových či internetových služeb jako jsou internetové vyhledávače nebo momentálně vyvíjená aplikace Memento sloužící k přímému přístupu k archivovaným verzím webových stránek prostřednictvím webového prohlížeče.
Současným trendem v oblasti archivace webu je rostoucí význam a využití rozsáhlých souborů dat získaných z webových archivů. Tyto tzv. *big data* mohou sloužit pro zkoumání jazyka, technologie, historie nebo dalších oblastí. Pro výzkum těchto dat se požívá různých vizualizací, textových analýz, zkoumání trendů a dalších metod[21]. Z důvodu očekávaného nárůstu těchto aktivit byla vytvořena v rámci určené skupiny tato specifická kategorie.

[21] REYNOLDS, E. If We Capture, Will They Come? Researcher Uses for Web Archive Collections. In: *The Signal: Digital Preservation* [online]. Library of Congress, 2013 [cit. 2014-07-16]. Dostupný z: http://blogs.loc.gov/digitalpreservation/2013/03/if-we-capture-will-they-come-researcher-uses-for-web-archive-collections/

Požadavky kategorie výzkumníků zabývajících se těmito souhrnnými daty se budou odlišovat od požadavků individuálních uživatelů zaměřených na konkrétní informace z archivu. Do budoucna je tedy třeba zjišťovat a monitorovat znalostní základnu i požadavky této kategorie určené skupiny.

4. Závěr

Závěrem shrňme, že nutným předpokladem k dlouhodobé ochraně dat z WebArchivu je potřeba vytvořit tzv. *preservation policy*, která bude obsahovat pravidla, postupy a strategie, které zajistí dlouhodobou použitelnost a přístupnost archivovaných dat. V projektu Národní digitální knihovny jsou data zatím chráněna pouze na úrovni bitstreamové ochrany, která není dlouhodobě dostatečná.
Prvním krokem k vytvoření strategie dlouhodobé ochrany je vytvoření profilu kolekce, která je nezbytná k dalšímu plánování. Profil kolekce představuje znalost o datech na nejvyšší úrovni, která jsou uložena v archivu, a o uživatelích, kteří s archivem pracují. Druhým krokem pro plánování této strategie je stanovení určené skupiny WebArchivu Národní knihovny a identifikace jejích kategorií a znalostní báze.
Jakmile budeme znát naše data a uživatele, můžeme se začít připravovat na tvorbu *preservation policy*, při které může implementovat některých z plánovacích nástrojů např. PLATO. V současné době není možné implementovat žádný nástroj pro plánování dlouhodobé ochrany, protože jednoduše nemáme všechny potřebné informace k procesu rozhodování.
Andrea Fojtů ve své analýze píše, že pro evaluaci potenciálních ochranných řešení a strategií je nutné projít třemi fázemi: definicí požadavků; hodnocením alternativ a posouzením výsledků[22].

a. Definice požadavkuˇ – pozůstává ze specifikace a podrobnějšího popisu sbírky (obsahu dat), která je vybrána pro naplánování ochranných akcí.
b. Hodnocení alternativ – hodnotící kritéria z předchozí fáze jsou podkladem pro provádění experimentálních migračních aktivit, které přicházejí do úvahy.
c. Posouzení výsledkuˇ - experimentální migrace jsou výstupem pro následnou analýzu a vyhodnocení nejlepšího formátu dané vzorové sbírky digitálních dat.

Za současné situace nejsme schopni projít celým procesem evaluace. Teprve nyní, když jsme dokončili profil kolekce, můžeme začít definovat požadavky pro plánování dlouhodobé ochrany. Díky tomuto vytvoření profilu kolekce, nyní můžeme úspěšně vytvořit strategii dlouhodobé ochrany dat a implementovat plánovací nástroje v dalších krocích. Budeme také moci zaručit, že naše sklizená data budou v příštích letech pro kohokoliv použitelná. Vytvoření návrhu stanovení a rozlišení jednotlivých kategorií v rámci určené skupiny je krokem k identifikaci jejich znalostní základny, díky které můžeme lépe porozumět informačním potřebám této skupiny a na jejich základě tak zajišťovat srozumitelnost uchovávaných dat. Dalším úkolem v rámci vytvoření *preservation policy* pak bude průzkum a určení nástrojů k identifikaci a monitorování neustále se měnící znalostní základny určené skupiny.
Tvorba *preservation policy* je dlouhodobá záležitost a velmi užitečné může být navázání spolupráce se zahraničními institucemi, které se potýkají s podobným problém - jako tomu bylo v případě vytváření profilu kolekce.
Velkým benefitem vytvoření profilu kolekce se může stát i zpřístupnění metadat nebo jejich části našim uživatelům, kteří je mohou využít k dalšímu výzkumu. A tím získat další uživatele projektu českého WebArchivu.

[22] FOJTŮ, Andrea. Plánování dlouhodobé ochrany pomocí nástroje PLATO Preservation Planning Tool. *Národní knihovna České republiky* [online]. 2011 [cit. 2014-07-21]. Dostupný z: http://www.nkp.cz/soubory/ostatni/vav2011-plato-vav2-jh.pdf

Seznam literatury:

ALSUM, Ahmed, Michele C. WEIGLE, Michael L. NELSON a Herbert VAN DE SOMPEL. Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* [online]. 2014 [cit. 2014-07-15]. DOI: 10.1007/s00799-014-0118-y. Dostupný z: http://link.springer.com/10.1007/s00799-014-0118-y

ČSN ISO 14721. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model*. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, Srpen 2014.

ČSN ISO 16363. Systémy pro přenos dat a informací z kosmického prostoru - Audit a certifikace důvěryhodných digitálních úložišť. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, Říjen 2014.

DAY, M. The Long-Term Preservation of Web Content. In: MASANÈS, J., ed. *Web archiving*. New York: Springer, c2006, s. 177-199. ISBN 978-3-540-23338-1.
FOJTŮ, Andrea. Plánování dlouhodobé ochrany pomocí nástroje PLATO Preservation Planning Tool. *Národní knihovna České republiky* [online]. 2011 [cit. 2014-07-21]. Dostupný z: http://www.nkp.cz/soubory/ostatni/vav2011-plato-vav2-jh.pdf
LAKSHMI, V. a S. C. JINDAL. *Digital libraries*. Delhi: Isha Books, 2004. ISBN 81-820-5109-6.
LAVOIE, Brian F. *The Open Archival Information System Reference Model: Introductory Guide* [online]. OCLC: Dublin, Ohio, 2004 [cit. 2014-7-14]. 20 s. Dostupný z: http://www.dpconline.org/docs/lavoie_OAIS.pdf
KVASNICA, Jaroslav a Rudolf KREIBICH,. Formátová analýza sklizených dat v rámci projektu WebArchiv NK ČR. *ProInflow* [online]. 31.10.2013 [cit. 2014-07-21]. ISSN 1804–2406. Dostupný z: http://pro.inflow.cz/formatova-analyza-sklizenych-dat-v-ramci-projektu-webarchiv-nk-cr.
KVAŠOVÁ, Zuzana a Tomáš SVOBODA. Dlouhodobá ochrana elektronických publikací. *ProInflow* [online]. 31.10.2013 [cit. 2014-07-21]. ISSN 1804–2406. Dostupný z: http://pro.inflow.cz/dlouhodoba-ochrana-elektronickych-publikaci.
PETROV, Petar. To fits or not to fits. *Open Planets Foundation* [online]. 2012 [cit. 2014-07-21]. Dostupný z: http://openplanetsfoundation.org/blogs/2012-07-27-fits-or-not-fits
RESEARCH LIBRARIES GROUP (RLG). *Trusted Digital Repositories: Attributes and Responsibilities* [online]. 2012 [cit. 2014-7-15]. Dostupný z: http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

REYNOLDS, E. If We Capture, Will They Come? Researcher Uses for Web Archive Collections. In: *The Signal: Digital Preservation* [online]. Library of Congress, 2013 [cit. 2014-07-16]. Dostupný z: http://blogs.loc.gov/digitalpreservation/2013/03/if-we-capture-will-they-come-researcher-uses-for-web-archive-collections/