WEBARCHIV – OD VÝZKUMU K (TVRDÉ) REALITĚ

Ludmila Celbová – Markéta Simonová, Národní knihovna ČR
Petr Žabička, Moravská zemská knihovna

Při uplatnění výsledků výzkumu v oblasti získávání, archivace a zpřístupnění internetových zdrojů, tedy při postupném zavádění do praxe, se ukazuje jako těžko překonatelný problém legislativa, zejména autorské právo. V prezentaci výsledků za uplynulý rok, od minulé konference na Seči, bychom chtěli předvést možnosti přístupu ke zdrojům uloženým v českém webovém archivu v kontextu knihovnickém, technickém i legislativním.

Úvod

Internet plní dnes ve společnosti funkci masmédia, které je využíváno pro zábavu, pro komunikaci a nepřehlédnutelná je jeho funkce informační. Neustále přibývá významných webových stránek, které přinášejí různé služby, aktuální informace o denním dění, odborné informace z nejrůznějších oborů, portály zprostředkující elektronický obchod, zábavné hry či jiné produkty, vyskytují se zde též stránky přinášející společensky ne zcela vhodné dokumenty. Z počtu publikací na Internetu připadá více než 90 procent na dokumenty, které existují pouze v elektronické podobě. Přitom podle výsledků studií zmizí 40 procent publikací na síti během jednoho roku a dalších 40 procent se v průběhu této doby změní. Podle toho bude pouze 20 procent dokumentů na Internetu dostupných od nynějška za rok.

Úlohou shromažďování, ochrany a zpřístupnění online dostupných elektronických informačních zdrojů publikovaných výhradně na Internetu se zabývají moderní depozitní knihovny od poloviny 90. let minulého století. V souladu se svým posláním se touto cestou vydala i Národní knihovna ČR, která ve spolupráci s Ústavem výpočetní techniky MU a Moravskou zemskou knihovnou v Brně buduje archiv českého webu. Projekt ***WebArchiv*** vznikl jako pilotní projekt výzkumu a vývoje v roce 2000. O celkové situaci v oblasti získávání, archivace a zpřístupňování internetových informačních zdrojů ve světě a v návaznosti v České republice, a to z hlediska technického, knihovnického i legislativního, přinesli řešitelé informace na loňské konferenci Knihovny současnosti. V posledním roce se řešitelé snaží aplikovat výsledky pilotního projektu v praxi.

Za tuto dobu pokročilo významně řešení na poli technickém, jak pokud jde o *získávání*, tak i o *ukládání* informačních zdrojů a s tím související řešení knihovnických aspektů. *Zpřístupnění* dokumentů z digitálního archivu je technicky daleko náročnější, nicméně i v této oblasti vývoj pokračuje. Největším problémem a brzdou pro zpřístupnění dokumentů z digitálního

archivu se v současné době celosvětově jeví nedořešenost legislativních otázek, zejména pokud jde o autorská práva. V současné době se připravují doporučení k řešení této problematiky rovněž v rámci UNESCO.

V našem příspěvku je proto věnována hlavní pozornost problémům v oblasti legislativy; stěžejním bodem příspěvku je okruh problémů vyplývajících z autorského práva pro archivaci českých webových zdrojů a jejich zpřístupnění a informace o tom, jaký je přístup českých vydavatelů k této problematice a jak řeší tyto problémy ve vztahu k vydavatelům Národní knihovna ČR.

1 Problém je v legislativě

1.1 Zákon o povinném výtisku

Základním posláním národních knihoven je zachování kulturního dědictví dané země pro budoucí generace. Tato funkce je ve většině zemí podpořena zákonem o povinném výtisku.

V České republice existují 2 zákony, které musíme brát v úvahu. Je to zákon č. 37/1995 Sb., o neperiodických publikacích a zákon č. 46/2000 Sb. tzv. tiskový zákon. Odevzdávání elektronických zdrojů jakožto povinného výtisku u nás zákon jednoznačně neurčuje. Ze zkušenosti z jiných zemí jako např. Dánsko, Norsko se můžeme poučit o důležitosti a smyslu takovýchto právních zakotvení.

1.1.1 Zákon č. 37/1995 Sb.

Tento zákon lze při jeho volném výkladu aplikovat pro potřebu elektronických publikací včetně publikací přístupných online, jelikož dle jeho znění „zahrnuje rozmnoženiny literárních, vědeckých a uměleckých děl určené k veřejnému šíření“, nosič zde zmíněn není. K praktickému využití zákona v oblasti publikací na Internetu je však třeba jej upřesnit prováděcí vyhláškou, ve které by byl popsán proces odevzdávání.

Problematickým by ovšem zůstal fakt, že zákon se vztahuje pouze na monografické publikace (jednorázově vydané publikace), kterých v dnešní době na Internetu mnoho nenajdeme.

1.1.2 Zákon č. 46/2000 Sb.

Horší je situace v případě druhém, tzv. tiskovém zákoně. Jak už sám název napovídá, o tzv. netištěných publikacích, tedy i elektronických zdrojích zde nemůže být řeč, přestože by sem tyto zdroje z hlediska svých vlastností (zejména periodicity) nejlépe spadaly.

Tento zákon popisuje práva a povinnosti vydavatelů při vydávání periodického tisku, otázce odevzdávání povinného výtisku je věnován pouze jeden (§9) z celkových 19 paragrafů. Proto se domníváme, že lze tento paragraf vyjmout a zařadit ho do nového obecně formulovaného zákona o povinném výtisku.

1.1.3 Novela zákona/zákonů

Novela zákona o povinném výtisku je nutností, jelikož děl publikovaných elektronickou cestou denně přibývá. Pokud si chce Národní knihovna

udržet status knihovny, která má za úkol uchovávat národní kulturní dědictví, měla by se hlavně ona o tuto novelizaci zasadit. Příklady podobných opatření jsou patrné v mnoha ostatních zemích. Např. ve Velké Británii je již novela zákona o povinném výtisku připravena a je na dobré cestě projít třetím čtením dolní sněmovny britského parlamentu v červenci tohoto roku. Dále můžeme jmenovat Německo, Rakousko, Francii, Švédsko či Finsko, ve všech těchto zemích se o novelu zákona snaží již několik let a v nejbližší době by mělo proběhnout schvalovací řízení v parlamentech jednotlivých zemí.

Víme, že příprava kvalitního zákona bude náročná, avšak nevyhnutelná. Základním rozhodnutím nejspíš bude, zda sloučit oba stávající zákony do jednoho jako je tomu ve většině zemí, nebo ponechat oba dva a snažit se je doplnit vyhláškou či nařízením.

Formulace nového zákona musí být také co nejobecnější a zároveň musí přesně definovat to, co je předmětem odevzdávání (velmi problematická je správná definice co je dokument, co je médium nebo kdo je vydavatel). V několika málo zemích již mají v zákoně povinnost odevzdávání elektronických online zdrojů zahrnutu, ale ne všude byl účel zákona naplněn.

Existuje již několik modelů zákona o povinném výtisku, máme tedy z čeho čerpat a zároveň se můžeme poučit o výhodách či nevýhodách jednotlivých modelů.

1.2 Autorský zákon

Druhým důležitým předpisem, dle kterého se musíme při tvorbě archivu řídit, je autorský zákon (z. 121/2000 Sb.). Existuje několik výkladů tohoto zákona. My jsme vycházeli z výkladu Jana Kříže a kol. Autorský zákon a předpisy související.

Cílem WebArchivu je stahování elektronických online zdrojů, jejich archivace a následné zpřístupnění co nejširší veřejnosti.

Pokud budeme analyzovat zákon ve vztahu k elektronickým zdrojům a ve vztahu k cílům WebArchivu, potom zjistíme, že dle tohoto zákona jsou tyto cíle nereálné (zejména zpřístupnění zdrojů z archivu), jelikož dle jeho znění:

§4 Zveřejnění a vydání díla

odst.2 – *Zahájením oprávněného veřejného rozšiřování rozmnoženin je dílo vydáno.*

§14 Rozšiřování

odst.1 – *Rozšiřováním originálu nebo rozmnoženiny díla se rozumí* ***zpřístupňování díla v hmotné podobě*** *prodejem nebo jiným převodem vlastnického práva k originálu nebo k rozmnoženině díla, včetně jejich nabízení za tímto účelem.*

§37 Užití díla rozmnožováním a rozšiřováním rozmnoženin

odst.1 – *Do práva autorského nezasahuje knihovna, archiv a jiné nevýdělečné školské, vzdělávací a kulturní zařízení, zhotoví-li rozmnoženinu díla pro své archivní a konzervační účely.*

§38 Užití díla půjčováním a pronájmem originálu nebo rozmnoženiny
odst.1 – *Do práva autorského nezasahuje osoba uvedená v §37 odst.1, půjčuje-li originály či rozmnoženiny* ***vydaných děl.***
§90 Obsah zvláštního práva pořizovatele databáze
odst. 2 – *Vytěžováním databáze se rozumí trvalý nebo dočasný přepis celého obsahu databáze nebo jeho podstatné části na jiný podklad, a to jakýmikoli prostředky nebo jakýmkoli způsobem.*

Těchto pět paragrafů rozhoduje velkým dílem o naší práci. Vytváření archivu, tzn. stahování a ukládání elektronických online zdrojů je dle znění §37 legální. Musíme však dát pozor na databáze, na které se vztahují §88-§94. V současné době jednáme s právníky o tom, zda harvesting neboli plošné stahování online zdrojů lze považovat za vytěžování databází (§90), čímž bychom autorské právo porušovali. Kromě toho není jednoznačné, co se dá v souvislosti s prostředím Internetu považovat za databázi ve smyslu autorského zákona.
Problematické rovněž zůstává zpřístupnění uložených zdrojů z digitálního archivu, jelikož dle §38 smí knihovna půjčovat originály či rozmnoženiny vydaných děl. A právě slovo vydaných je pro nás překážkou. Pokud se podíváme na §4 a 14, potom zjistíme, že dílo je vydáno zahájením rozšiřování rozmnožením a zároveň rozšiřováním rozmnoženiny se rozumí zpřístupňování díla v hmotné podobě. Online zdroje však nelze považovat za díla v hmotné podobě.

Z těchto důvodů jsme byli nuceni přistoupit k variantě oslovování jednotlivých vydavatelů a uzavírání smluv o poskytování elektronických online zdrojů (viz odstavec 2). Touto metodou je možné oslovit několik desítek, možná stovek vydavatelů, není to však ideální řešení.

Dalším východiskem z naší složité situace by mohla být Směrnice o harmonizaci některých aspektů autorského práva a práv s ním souvisejících v informační společnosti (2001/29/ES), kterou vydaly Evropský parlament a Rada v roce 2001. Směrnice ovšem nemá bezprostřední použitelnost. Členské státy sice musí zahrnout její obsah do svého právního řádu, avšak v určité dané lhůtě. Podle právníků k tomu v České republice v nejbližších 5 letech nedojde, a tudíž se musíme řídit dle stávajícího autorského zákona. Je to škoda, jelikož směrnice dovoluje knihovně jednak zhotovování rozmnoženin nad rámec pouhé interní archivace či konzervace (čl. 5/2(c)), a zejména umožňuje sdělování nebo zpřístupňování autorských děl, která má knihovna ve svých sbírkách, na vyčleněných terminálech ve svých prostorách jednotlivým členům veřejnosti za účelem výzkumu nebo soukromého studia (čl. 5/3(n)).

1.3 Alternativní řešení

Alternativu legislativy povinného výtisku elektronických zdrojů nabízí *Mezinárodní deklarace k odevzdávání elektronických dokumentů do konzervačního fondu,* která byla připravena na základě rozsáhlé spolupráce CENL (Conference of European National Librarians) a Federace evropských vyda-

vatelů (FEP - Federation of European Publishers). Deklarace byla publikována v roce 2000.

Mnoho zemí, které získávají elektronické zdroje na základě dohody o dobrovolném odevzdávání, se při tvorbě takovýchto dohod opíralo právě o tuto mezinárodní deklaraci. V rámci deklarace byla stanovena pravidla pro dobrovolné poskytování kopie elektronických online dokumentů do elektronického archivu. Bylo doporučeno, aby se knihovny ve fázi pilotních projektů dohodly s vydavateli na otázkách definic pojmů „dokument" a „vydavatel". Implementace těchto pravidel by měla být průběžně monitorována a na základě zkušeností by se měla navrhnout účinná a oběma stranám vyhovující legislativa.

2 Současný WebArchiv

Novela zákona o povinném výtisku zahrnující elektronické online zdroje vyžaduje propracované podklady, proto jsme se rozhodli současnou situaci řešit vytvořením **Smlouvy o poskytování elektronických online zdrojů**, kterou jsme připravili ve spolupráci s právníky na základě výše zmiňovaného doporučení CENL/FEP.

Kritéria výběru zdrojů byla stanovena již v prvních letech existence WebArchivu. Jde nám především o to, archivovat ty publikace, u kterých je větší pravděpodobnost ztráty. Tzn. online publikace, které nevycházejí v jiné formě (papírové, fyzické médium). Dalším kritériem je obsah zdroje, který by měl být určen především pro informování, nikoliv pro zábavu. Třetím hlavním kritériem bylo stanovení formátu zdroje (preferují se všeobecně podporované formáty jako např. HTML, XML, JPG, RTF).

V první fázi jednání s potencionálními vydavateli a při přípravě zkušebního vzorku archivovaných zdrojů jsme vybrali 20 vydavatelů, které jsme prostřednictvím elektronické pošty oslovili. Z nich 4 spolupráci odmítli a se 4 vydavateli stále jednáme. V současné době má tedy Národní knihovna ČR v rámci projektu WebArchiv uzavřeno 12 Smluv o poskytování elektronických online zdrojů. Seznam vydavatelů a jejich zdrojů najdete na webových stránkách projektu.[1]

Pro letmé seznámení se smlouvou jsme vybrali její **nejdůležitější body**:

- Poskytovatel (vydavatel) souhlasí s tím, aby: „Národní knihovna vyhledávala, stahovala, ukládala, činila kopie a trvale uchovávala jím vydané elektronické online zdroje…".
- Poskytovatel souhlasí, aby se jím poskytnuté elektronické online zdroje staly součástí České národní bibliografie.
- Poskytovatel se dále zavazuje vytvářet resp. vkládat do zdrojů metadata dle standardu Dublin Core.
- Smlouva se nevztahuje na ty zdroje, které byly Národní knihovně odevzdány jako povinný výtisk v jiné formě, zdroje reklamní povahy, utajované skutečnosti, zdroje určené výhradně pro interní nebo soukromé

[1] http://www.webarchiv.cz/vydavatele.html

užití, materiály politických stran, občanských sdružení apod., jejichž cílem je rozšíření členské základny.

- Za poskytování zdrojů nepřísluší poskytovateli žádná odměna, veškeré náklady nese Národní knihovna.
- Elektronické online zdroje získané na základě této smlouvy od poskytovatele uloží /archivuje/ Národní knihovna na zvláštním počítačovém serveru spravovaném Národní knihovnou ve spolupráci s Ústavem výpočetní techniky Masarykovy univerzity v Brně a připojeném k akademické síti CESNET.
- Plný přístup k archivovaným zdrojům mají pouze oprávnění zaměstnanci.
- Národní knihovna je oprávněna zpřístupňovat rozmnoženinu elektronického online zdroje pouze na vymezených terminálech oprávněným uživatelům (rozuměj registrovaným uživatelům knihovny). Tito uživatelé mají ke zdrojům omezený přístup, tzn. mohou si zdroje pouze číst bez možnosti je kopírovat, měnit či rozmnožovat.

Důvodem pro odmítnutí smlouvy vydavatelem se stal právě poslední zmíněný bod. Vydavatelé nerozuměli tomu, proč by měl být jejich zdroj přístupný pouze z vymezeného terminálu a pouze oprávněným uživatelům, když na Internetu si ho může prohlédnout kdokoli a odkudkoli. Problém je v autorském zákoně, o kterém jsme se zmínili v předchozí kapitole a který sice umožňuje knihovně zhotovit rozmnoženinu díla pro své archivní a konzervační účely, ale již neumožňuje její veřejné zpřístupnění.

Zvažujeme proto přepracovat stávající smlouvu tak, aby byl umožněn přístup k uloženým zdrojům v digitálním archivu volně přes www rozhraní spravované Národní knihovnou ve spolupráci s ÚVT MU. Zatím čekáme na vyjádření právníka, zda je takovýto postup legální.

Všechny zdroje, s jejichž vydavateli máme podepsanou smlouvu, jsou katalogizovány v knihovnickém systému Aleph a měly by se stát jak součástí archivu elektronických online zdrojů (konzervační fond) tak součástí České národní bibliografie (bibliografické záznamy).

Při zpracování elektronických zdrojů hrají rovněž důležitou roli metadata, tj. strukturované, popisné údaje, které nesou informace o primárních datech a umožňují je správně interpretovat. Metadata plní vedle popisné funkce (reprezentují formální a obsahové znaky dokumentu) také funkci archivační (zachování integrity elektronického dokumentu) a vyhledávací.

Smlouvu vydavatelům předkládáme ve dvou verzích. Varianty smluv se liší v bodě II/5, a to v rozsahu práce, který pro vydavatele vyplývá z podepsání konkrétní smlouvy. V první verzi se vydavatel zavazuje sám vkládat do jím vydávaných elektronických online zdrojů (konkrétně jeho zdrojového kódu) metadatový záznam podle standardu Dublin Core. Ve druhé verzi smlouvy se vydavatel zavazuje vkládat metadata, která zpracuje Národní knihovna.

3 Potřeba spolupráce

3.1 Spolupráce s ISSN

Aktuální změnou při strategii oslovování potencionálních vydavatelů online zdrojů je spolupráce s Českým národním střediskem ISSN při Státní technické knihovně v Praze. Každý žadatel o přidělení čísla ISSN je povinen vyplnit základní údaje o sobě a o svém zdroji v elektronickém formuláři, který je umístěn na webových stránkách ČNS ISSN.

Naším požadavkem je, aby do formuláře byla nově zařazena otázka, zda vydavatel souhlasí se zařazením svého zdroje do WebArchivu. Rovněž jsou zde základní informace o projektu a jeho cílech. Chtěli bychom dosáhnout toho, aby se vydavatelé online zdrojů o takovéto možnosti archivace dozvěděli a začali s námi spolupracovat.

Kooperace s ČNS ISSN je v začátcích, proto si na konkrétní výsledky budeme muset ještě chvíli počkat.

3.2 Spolupráce s MI ČR

Kontakt byl taktéž navázán se zástupci Ministerstva informatiky, s nimiž proběhly již 2 schůzky. Cílem spolupráce z naší strany je archivace dokumentů veřejné správy, kterým již uběhla doba platnosti, a tudíž nejsou volně dostupné. Všechny dokumenty veřejné správy by měly být opatřeny metadaty dle standardu Dublin Core. Výhodou je také fakt, že se na tyto dokumenty nevztahuje autorské právo, tzn. můžeme je archivovat a následně zpřístupnit bez omezení.

Na druhé straně Ministerstvo informatiky, resp. firma, která vytváří elektronický katalog dokumentů veřejné správy, projevila zájem o naše softwarové nástroje – Nedlib harvester a generátor metadat, které by při tvorbě elektronického katalogu chtěla využít.

4 Digitální archiv a jeho zpřístupnění

Cílem projektu WebArchiv je *zajištění trvalého uchování domácích elektronických online publikovaných informačních zdrojů jako součásti národního kulturního dědictví*. Vzhledem k povaze, rozmanitosti a množství těchto zdrojů je zřejmé, že stanovení podmínek, které musí archivované elektronické zdroje splňovat, významně ovlivní budoucí hodnotu vytvořeného archivu.

4.1 Výběr zdrojů k archivaci

Již rozhodnutí zaměřit se primárně na „webové“ zdroje znamená, že se zaměřujeme jen na jistou *část množiny* všech online elektronických zdrojů. Jak ukazují dosavadní zkušenosti, z hlediska dlouhodobé konzervace je nejvýznamnější část dokumentů dostupná přes protokol http. Vedle protokolů lze jednotlivé dokumenty hodnotit podle použitého formátu (některé ze vzácněji se vyskytujících formátů nemá téměř smysl archivovat).

Předmětem zájmu projektu WebArchiv je *český web*. Ten můžeme zjednodušeně definovat jako *„všechny dokumenty publikované v domé-*

ně.cz,". Je ovšem zřejmé, že toto kritérium nemůže pokrýt celou českou online produkci. Proto by bylo vhodné tento rozsah rozšířit o mnoho dalších, vzájemně se doplňujících kategorií: dokumenty v doménách druhé úrovně registrovaných subjektem sídlícím v České republice; dokumenty publikované na serverech fyzicky umístěných v ČR; dokumenty v českém jazyce; dokumenty českých autorů; dokumenty se vztahem k Česku.

V ideálním případě by měl být výsledkem projektu archiv obsahující pokud možno vše, co kdy bylo v rámci českého webu publikováno. Na druhou stranu je zřejmé, že pokus o takový přístup by s sebou přinesl prohibitivně vysoké náklady a byl by i technicky v podstatě neproveditelný. Proto se provádí archivace dvěma cestami:

- *plošnou archivací*, kdy se s delším časovým odstupem vytvářejí co nejúplnější snímky celého českého webu (například jednou nebo dvakrát ročně); současná legislativa ale bohužel neumožňuje zpřístupnění takto získaného archivu (rovněž plošné),
- *výběrovou archivací*, kdy se naopak velmi často (v případě potřeby i každý den) doplňuje archiv zrcadlící jen omezenou vybranou skupinu nejvýznamnějších českých zdrojů. V tomto případě se jedná o skupinu zdrojů vybíraných na základě výše uvedených kritérií (těmi se také zabýval příspěvek ve sborníku Knihovny současnosti 2002), na jejichž využívání uzavírají řešitelé s vydavateli výše zmíněné smlouvy.

4.2 Dlouhodobé uchování a zpřístupnění zdrojů

Problematika archivace webu tak zahrnuje dvě oblasti: jednou z nich je automatizované (plošné či výběrové) sklízení informací nacházejících se na definovaném výseku webu a jejich uložení do archivu. Druhou oblast pak představuje zpřístupnění informací uložených v takto vytvořeném (a objemem dat velmi rozsáhlém) archivu.

4.2.1 Sklizeň českého webu

V loňském roce probíhala po několik měsíců již druhá testovací sklizeň celé domény .cz. Analýza jejího průběhu ukázala, jaké informační bohatství český web skrývá: mezi padesáti našimi objemem nebo počtem souborů největšími doménami druhé úrovně najdeme mimo jiné šest univerzit, jeden univerzitou provozovaný specializovaný server (linux.cz), Českou akademii věd a několik zpravodajských a vydavatelských serverů.

4.2.2 Provoz archivu

Velikost Harvesterem tvořeného archivu může snadno dosáhnout obrovských rozměrů: jedno kolo stahování představuje v našich podmínkách stovky GB. Archiv s tak velkým potenciálem růstu není samozřejmě snadné ani levné provozovat. Ačkoli v současné době již jsou na trhu levné pevné disky o kapacitách přes 200 GB, infrastruktura archivu se musí opírat o robustní a dlouhodobě perspektivní řešení. Toto řešení musí brát v potaz nejen aspekty technické, ale i finanční a personální a musí být z provozního hlediska dlouhodobě provozovatelné.

V pilotní fázi projektu bylo s výhodou využito stávajícího páskového robota Národní knihovny ČR, jehož nevýhodou ovšem je problematická dostupnost na něm uložených dat v okamžiku, kdy je nutné tato data zpřístupnit veřejnosti. Proto došlo v letošním roce ve spolupráci s Masarykovou univerzitou a Moravskou zemskou knihovnou k migraci celého archivu na diskové pole umístěné v Brně, kde se také nachází loni pořízený server pro archivaci. Ačkoli jsou stažené dokumenty společně s příslušnými metadaty ukládány v archivu jako tar+gzip komprimované soubory přímo do souborového systému, ukázalo se, že přenos celého archivu byl díky kontrolám konzistence přenášených dat časově velmi náročný.

Větším oříškem samozřejmě jsou samotné archivované soubory. Je sice pravděpodobné, že nejrozšířenější formáty zůstanou dlouhodobě interpretovatelné (html, txt, gif, jpg), lze ale mít oprávněné pochybnosti o všech proprietárních formátech, především těch, které nejsou tak rozšířeny jako například formáty firem Adobe nebo Microsoft. I u formátů Microsoftu je však zárukou jejich budoucí interpretovatelnosti spíše dostupnost alternativních programů s otevřeným kódem, které umějí s těmito formáty pracovat (OpenOffice), než podpora ze strany Microsoftu. Otázka, zda v budoucnosti takové formáty konvertovat, nebo zda jít cestou emulace, však zůstává stále otevřená.

Ať už bude v budoucnosti vývoj tohoto archivu jakýkoli, lze říci, že využitím NEDLIB Harvesteru získala Národní knihovna vhodný nástroj pro tvorbu konzervačního archivu českého webu. Vytvoření takového archivu je sice důležitým, ale zároveň jen prvním krokem na cestě k naplnění jeho smyslu, tedy ke zpřístupnění jeho obsahu.

4.2.3 Zpřístupnění informací v archivu

Pro zpřístupnění archivu se nabízejí technologie fulltextového indexování a automatizované extrakce autorem vytvořených metadat. Koncem roku 2001 byl na MFF UK vypsán ročníkový týmový vývojový projekt na vytvoření indexační a vyhledávací aplikace pro Webarchiv, koncem června 2003 byl projekt úspěšně obhájen a v současné době je vyvinutý produkt připravován pro ostré nasazení. Tato aplikace zpřístupňuje stažené dokumenty v jejich kontextu, tedy s vloženou grafikou ze stejné doby a s odkazy vedoucími primárně opět do archivu. Vyhledávání v archivu je umožněno nejen na základě URL nebo kontrolního součtu dokumentu, ale i na základě z dokumentu extrahovaných metadat nebo fulltextového vyhledávání. Celá aplikace je navržena tak, aby bylo možné k ní kdykoli připojit moduly pro indexování jiných, než textových typů souborů – jeden z takových nástrojů, Convera Retrievalware, je v NK již zkušebně provozován. Jedním z budoucích cílů projektu bude proto pokus o jeho využití pro indexování některých netradičních typů souborů obsažených v archivu. S využitím této indexační a vyhledávací aplikace se prozatím počítá pro vyhledávání v relativně malém souboru vybraných významných zdrojů, jejichž zpřístupnění umožnil vydavatel ve smlouvě uzavřené s Národní knihovnou ČR.

Nadějně se jevily kontakty s týmem Norské národní knihovny, která vyvinula a v letošním roce se chystá dát volně k dispozici vlastní systém pro indexaci a zpřístupnění webového archivu založený na indexovacím enginu Apache Jakarta Lucene. Jeho uvolnění je však po několika odkladech naplánováno až na říjen letošního roku.

4.2.4 Současné plány na zpřístupnění informací v archivu

Máme stanoven alespoň přibližně rozsah českého webu, v jeho rámci můžeme začít vyhledávat v podmnožině zdrojů archivovaných výběrově v co největší úplnosti, na něž získala Národní knihovna oprávnění pro zpřístupnění za dohodnutých podmínek. Vedle přímého vyhledávání v archivu, které je zatím pouze v testovacím režimu, se v současné době nabízí několik způsobů, jak tuto činnost zajišťovat. Nejperspektivnějším z nich by mohlo být využití potenciálu projektu Jednotné informační brány CASLIN (www.jib.cz). Jedním z jejích výstupů bude totiž průběžně aktualizovaný předmětově členěný informační portál online elektronických zdrojů. Správa jednotlivých oborů tohoto portálu bude svěřena vždy té knihovně, která má v daném oboru největší zkušenosti. Díky tomu lze očekávat, že každý obor bude v portálu reprezentován i nejvýznamnějšími národními informačními zdroji, které se tak stanou i předmětem zájmu projektu Webarchiv.

V rámci automatizovaného knihovního systému Aleph, používaného (nejen) v Národní knihovně ČR, se nabízí možnost ukládání podrobných bibliografických záznamů, z nichž lze prostřednictvím aktuální adresy URL zajistit přístup do zdroje na Internetu a prostřednictvím adresy archivovaného zdroje umožnit přístup do digitálního archivu. Záznamy v Alephu by tak mohly sloužit současně pro zpřístupňování v rámci tématických bran, pro konverzi do souborného katalogu, případně najít využití v dalších digitálních knihovnách. Popis zdrojů, ať už se jedná o aplikaci bibliografických formátů typu MARC nebo jednodušší a mezinárodně srozumitelný formát Dublin Core, používaný pro přímý popis zdroje, může ve vybraných případech jít až na úroveň jednotlivých dokumentů (článků v časopisech, příspěvků ve sbornících apod.). Dosažení této úrovně podrobnosti již ovšem bude vyžadovat nasazení prostředků pro kooperativní online budování této databáze.

V jakémkoliv systému je ovšem třeba zajistit přístup ke zdrojům v souladu s podmínkami dohodnutými s vydavatelem. Pokud tedy bude vydavatel, jenž publikoval své zdroje na Internetu volně, souhlasit se zpřístupňováním těchto zdrojů volně též z WebArchivu, může být zdroj přístupný v síti bez jakéhokoliv omezení a z jakéhokoliv připojeného počítače. V případě licencovaných zdrojů či zdrojů z jakéhokoliv důvodu chráněných, kde vydavatel nesouhlasí s volným přístupem ke zdroji uloženému v digitálním archivu, je třeba zabezpečit přístup ke zdroji tak, aby Národní knihovna jako poskytovatel chránila zpřístupňované zdroje proti nedovolenému kopírování. V tomto případě mohou být zdroje zpřístupněny pouze lokálně, na počítačích k tomu účelu vyhrazených a zabezpečených.

Závěr

Zda bude některá z dosud popisovaných technologií nasazena také v ostrém reálném provozu, bude samozřejmě záviset i na vyřešení autorskoprávní problematiky související s tvorbou a provozem takového archivu. Nedotaženost zákona o povinném výtisku u nás otevírá cestu různým výkladům omezení daných zákonem o autorském právu. Automatickou identifikaci a archivaci online publikovaných dokumentů lze srovnávat s běžně používanou technologií indexování webu, jak ji provádějí Internetové prohledavače. S tím, jak bude stoupat podíl čistě elektronické produkce, bude růst i význam její dlouhodobé archivace z hlediska ochrany národního kulturního dědictví. Přesto ale není jisté, zda bude bez opory v zákoně možné využívat stávající strategii plošné archivace.

Je možné prohlásit, že právo občana na informace by mělo být naplněno i existencí digitální knihovny obsahující elektronicky publikované dokumenty v nezměněné podobě. Zajištění integrity takové knihovny musí být proto jedním z prioritních úkolů jejího provozovatele.

Je patrné, že práce na poli zpřístupnění archivu budou dlouhodobou záležitostí, která si vyžádá nemalé prostředky. Jednou z cest, jak tyto prostředky účelně vynaložit, je spolupráce na meziresortní i mezinárodní úrovni, která se velmi osvědčila již během řešení pilotního projektu.

Záležitosti archivace digitálních dokumentů se však netýkají pouze knihoven v souvislosti se zajištěním trvalého přístupu k národnímu kulturnímu bohatství. Problematiku jinou obsahem, ale obdobnou technicky (příp. i legislativně) budou muset řešit např. archivy, muzea, ale i vládní a správní orgány. Zdá se, že je nejvyšší čas, aby se problematika archivace internetových zdrojů dostala k řešení na vyšší instanci, tedy na úroveň státních orgánů odpovídajících za informační politiku a za zajištění svobodného přístupu občanů k informacím.

Literatura:

1. *WebArchiv* [online]. Praha : Národní knihovna ČR, posl. aktual. 13. června 2003 [cit. 2003-07-06]. Dostupné na World Wide Web: <http://webarchiv.nkp.cz>.
2. CELBOVÁ, Ludmila; ŽABIČKA, Petr. WebArchiv – digitální knihovna českého webu. In *INFOS 2003 : zborník z 32. medzinárodného informatického sympózia, ktoré se konalo v dňoch 7. – 10. apríla 2003 v Starej Lesnej*. Bratislava : Spolok slovenských knihovníkov, 2003, s. 41-46. ISBN 80-86249-18-2. Dostupné též na World Wide Web: <http://webarchiv.nkp.cz/infos2003.pdf>.
3. CELBOVÁ, Ludmila. *WebArchiv – vytvoření podmínek pro zpřístupnění českých webových zdrojů (knihovnické, legislativní a technické aspekty) : zpráva o plnění cílů projektu VISK3* [online]. Praha : Národní knihovna ČR, leden 2003, [cit. 2003-07-06]. Dostupné na World Wide Web: <http://webarchiv.nkp.cz/zprava2002/zprava2002.pdf>.
4. ŽABIČKA, Petr. Konference ECDL 2002. *Ikaros* [online]. 2002, č. 10 [cit. 2003-07-06]. ISSN 1212-5075. Dostupné na World Wide Web: <http://www.ikaros.cz/ Clanek.asp?ID=200209068>.

5. ŽABIČKA, Petr. Webarchiv – digitální knihovna českého webu. In *RUFIS 2002*. Brno : ApS Brno, 2002, s. 121-129. ISBN 80-86510-40-9. Dostupné též na World Wide Web: <http://webarchiv.nkp.cz/rufis2002_pz.pdf>.
6. ŽABIČKA, Petr. Archiv českého webu v roce 3. *Národní knihovna*. 2002, roč. 13, č. 3, s. 168-176. ISSN 1214-0678. Dostupné též na World Wide Web: <http://webarchiv.nkp.cz/nk2002.pdf>.
7. CELBOVÁ, Ludmila; ŽABIČKA, Petr. Internetové zdroje jako součást digitálních knihoven i jako součást kulturního dědictví. In *Knihovny současnosti 2002*. Brno : Sdružení knihoven ČR, 2002, s. 294-308. ISBN 80-86249-18-2. Dostupné též na World Wide Web: <http://webarchiv.nkp.cz/sec2002_lc.doc>.
8. ŽABIČKA, Petr. Infrastruktura Webarchivu v roce 2002. In *Inforum 2002* [online]. Praha : Albertina icome Praha, 2002 [cit. 2003-07-06]. Dostupné na World Wide Web: <http://www.inforum.cz/inforum2002/prednaska8.htm>.
9. CELBOVÁ, Ludmila. *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet : závěrečná zpráva za léta 2000-2001* [online]. Praha : Národní knihovna ČR, leden 2002, [cit. 2003-07-06]. Dostupné na World Wide Web: <http://webarchiv.nkp.cz/zprava2001/zprava2001.pdf>.
10. CELBOVÁ, L., SIMONOVÁ, M., TATRANSKÁ, M. Zpřístupnění elektronických zdrojů z digitálního archivu: jak a pro koho. In *RAMAJZLOVÁ, Barbora (sest.). Automatizace knihovnických procesů – 9. : sborník z 9. ročníku semináře pořádaného ve dnech 15. – 16. května 2003 v Liberci.* Praha: ČVUT, 2003, s. 58-69. ISBN 80-0102-738-4. Dostupné též na World Wide Web: < http://knihovny.cvut.cz/akp2003/index.htm>.
11. zákon č. 37/1995 Sb., o neperiodických publikacích
12. zákon č. 46/2000 Sb., o právech a povinnostech při vydávání periodického tisku a o změně některých dalších zákonů (tiskový zákon)
13. Směrnice 2001/29/ES Evropského parlamentu a Rady z 22. května 2001 o harmonizaci některých aspektů autorského práva a práv s ním souvisejících v informační společnosti. [online]. Praha : Národní knihovna [cit. 2003-07-10]. Dostupné na World Wide Web: <http://www.nkp.cz/o_knihovnach/Dir01_29_ECcz.pdf>.
14. BOHÁČEK, Martin. Autorské právo a elektronický obchod po vstupu ČR do ES z hlediska knihoven. In *Inforum 2003* [online]. Praha : Albertina icome Praha, 2003 [cit. 2003-07-10]. Dostupné na World Wide Web: <http://www.inforum.cz/inforum 2003/prispevek.asp?CisloSekce=20&Kod=123>.
15. zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon)