

Místo internetových zdrojů v knihovních fondech

Ludmila Celbová, Petr Žabička

Elektronické informační zdroje a knihovny

Když se řekne „knihovna“, vybaví si většina lidí budovu plnou knih. Dodá-li se ještě „Internet“, ne všichni si už uvědomí všechny implikace tohoto spojení. Moderním knihovnám nejde totiž jen o to, poskytnout uživatelům počítače s přístupem na Internet, případně pasivně zpřístupnit touto cestou cizí informační zdroje. Snahou většiny knihoven je i nabídnutí vlastního obsahu, vlastních elektronických zdrojů, ať už formou jednoduchých webových stránek, nebo rozsáhlých digitálních knihoven. Obsahem digitálních knihoven se vedle digitalizovaných vlastních fondů knihovny mohou stát elektronické zdroje online, které knihovna získala do svého fondu. V tomto případě se nemusí nutně jednat pouze o zdroje pořízené nákupem, nejčastěji na základě licence, ale také o zdroje získané do digitálního archivu jako konzervačního fondu depozitní knihovny.

Úvod

Internet se stal symbolem moderní doby a bývá proto občas přirovnáván svým významem k vynálezu počátku novověku - knihtisku, kterým Gutenberg zahájil novou epochu komunikační techniky a který zásadně přispěl k šíření vzdělanosti v tehdejší Evropě. Zda je toto přirovnání na místě, ukáže nejspíše historie. Skutečností však je, že se Internet stal mocným komunikačním prostředkem umožňujícím celosvětové sdílení informací.

Internet je dnes informačním prostředím, kde průběžně narůstá obrovské množství webových stránek přinášejících informace menšího i značného rozsahu, významného i méně důležitého obsahu (včetně „závadného“ obsahu), charakterizované různými formami i formáty, kterým je společné jednak to, že jejich informační hodnota může být značně proměnlivá a jednak to, že i sama jejich existence je velmi pomíjivá. Vzhledem k tomu, že se jedná o dokumenty nehmotné, mohou být kdykoliv měněny či dokonce z Internetu smazány, musíme je považovat za velmi nestálé. 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. Z toho logicky vyplývá, že od nynějška za rok bude v nezměněné podobě na Internetu dostupných pouze 20 procent dnešních dokumentů.

S prudkým nárůstem objemu informací publikovaných výhradně na Internetu by se úkolem moderní depozitní knihovny mělo stát vedle péče o „tradiční“ dokumenty také shromažďování, ochrana a zpřístupnění online dostupných elektronických informačních zdrojů. Touto úlohou 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 Moravskou zemskou knihovnou v Brně a s významným přispěním Ústavu výpočetní techniky MU buduje archiv českého webu. Projekt **WebArchiv** vznikl jako pilotní projekt výzkumu a vývoje v roce 2000. V rámci menších grantových podpor Ministerstva kultury ČR se daří projekt i nadále rozvíjet a postupně připravovat podmínky pro zavádění výsledků výzkumu do praxe.

1. NĚCO MÁLO Z HISTORIE

Ve světě (zejména v USA, Kanadě, Austrálii a v evropských severních zemích) existuje již víceletá zkušenost s projekty zaměřenými na sbírky elektronických dokumentů publikovaných v síti Internet. V Evropě se za podpory Evropské komise zabývaly touto problematikou společné mezinárodní projekty (**CoBRA+**, **BIBLINK**, **NEDLIB**), jejichž cílem bylo stanovit budoucí úlohu evropských národních knihoven ve vztahu k elektronickým publikacím a vytvořit podmínky pro propojení národních bibliografických agentur a vydavatelů elektronických publikací, které by bylo užitečné pro obě strany.

K nejvýznamnějším národním projektům patří **PANDORA** (Austrálie), **EVA** (Finsko), **INDOREG** (Dánsko), **Kulturarw3** (Švédsko) a **DNEP** (Nizozemí).

Česká republika se jako první ze zemí bývalého východního bloku zařadila mezi nejvyspělejší země v oblasti řešení problémů trvalé ochrany a zpřístupnění internetových zdrojů v roce 2000, kdy Národní knihovna ČR získala grant Ministerstva kultury ČR pro **řešení dvouletého pilotního projektu v rámci programu výzkumu a vývoje. Na projektu Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet**, který dostal pracovní název **WebArchiv**, spolupracovala řešitelská instituce, Národní knihovna ČR, v oblasti informačních a komunikačních technologií s pracovníky Ústavu výpočetní techniky Masarykovy univerzity v Brně.

Během doby řešení pilotního projektu se tato problematika dostala mezi problémy řešené rovněž na půdě mezinárodních institucí. Nejvýznamnější v Evropě je společná iniciativa **CENL** (Conference of European National Librarians) a **FEP** (Federation of European Publishers), jejímž výsledkem je deklarace upravující vztahy mezi vydavateli elektronických zdrojů a depozitními institucemi *International declaration on the deposit of electronic publications*. Problém trvalého uchování národního bohatství v podobě elektronických publikací, zejména síťových, tedy už přestává být experimentem „pokročilejších“ zemí a stává se obecně naléhavou výzvou pro knihovny i nakladatele. Mnohé z elektronických zdrojů, které neexistují souběžně v tradiční (tištěné nebo analogové) formě (tzv. *digital born* dokumenty), byly již trvale ztraceny, neboť jejich tvůrci nebo vydavatelé odstranili své elektronické publikace z webu, aniž by zajistili jejich trvalou archivaci. Problematickou ochranu digitálního dědictví se proto v rámci programu Paměť světa (*Memory of the World*) začala zabývat i organizace **UNESCO**. Naléhavost řešení tohoto problému dokládá též vyjádření zástupců mezinárodní federace knihovnických asociací **IFLA**, která rovněž usiluje o dohody s vydavateli, resp. s **IPA** (International Publishers Association): „I když náklady na dlouhodobou archivaci jsou vysoké, náklady na nicnedělání v této oblasti by byly katastrofální.“

Jedním z průkopníků na poli archivace webu je americká nezisková organizace **Internet Archive** (www.archive.org), jejíž archiv sahá až do roku 1996. Tato organizace se ve spolupráci s dalšími institucemi snaží (vcelku úspěšně) vybudovat co nejrozsáhlejší archiv světového webu. Takový záměr je však finančně vysoce nákladný a proto zahájila spolupráci s největšími světovými národními knihovnami s cílem vyvinout novou generaci nástrojů pro archivaci a zpřístupnění webových informačních zdrojů. Předpokládá se, že softwarové nástroje vyvinuté v rámci tohoto projektu budou dány k dispozici dalším institucím, zabývajícím se touto problematikou.

Je zřejmé, že každá knihovna nemá prostředky na to, aby si vytvářela archiv celého světového webu pro vlastní potřebu, zároveň ale není možné spoléhat se výhradně na vydavatele elektronických informačních zdrojů, kteří mohou publikované dokumenty libovolně modifikovat nebo zcela odstranit. Je proto logické, že se každá vyspělá země snaží (většinou prostřednictvím národní knihovny daného státu) přednostně vybudovat národní archiv elektronických informačních zdrojů.

Přístup jednotlivých knihoven k řešení problému se ovšem velmi liší. Některé knihovny, jako například Australská národní knihovna, se snaží *archivovat výběrově* jen ty webové zdroje, jejichž kvalitu předem zhodnotí knihovník (pandora.nla.gov.au). Díky tomuto přístupu čítá sice nyní archiv australského webu po několika letech provozu jen 4723 webových sídel nebo jejich částí, nicméně jde o výběr toho „nejdůležitějšího“, co bylo v dané době na webu publikováno. Tento přístup je však velmi náročný na lidské kapacity a proto se většina knihoven vydala cestou automatizované *plošné archivace* všech dokumentů, které splňují automaticky vyhodnotitelná kritéria. K tomu využívají nejčastěji softwarových nástrojů vyvinutých v minulých letech v rámci evropských projektů nebo projektů evropských severovýchodních zemí. Vznikají však i další iniciativy, například v konsorciu s Internet Archive se po několikaletém zkoumání problematiky rozhodly spojit své síly americká **Kongresová knihovna**, **Britská knihovna**, **Francouzská národní knihovna** a některé **severovýchodní národní knihovny**. Pozadu nezůstává ani Japonská národní knihovna a zahájen byl i projekt na archivaci webových zdrojů v čínštině.

2. A TROCHU VÍCE ZE SOUČASNOSTI

Cílem projektu Webarchiv je, jak již jeho název napovídá, zajištění *trvalého uchování domácích elektronických, online publikovaných informačních zdrojů jako součásti národního kulturního dědictví*. Celou problematiku lze rozdělit rámcově do čtyř okruhů, které se ovšem vzájemně prolínají:

- *Legislativní problematika*
- *Kritéria výběru zdrojů a strategie jejich archivace*
- *Bibliografická správa a zpřístupnění zdrojů*
- *Návaznost na obdobné projekty na mezinárodní úrovni*

2.1 Legislativa

V současné době existuje několik kritických míst, která mohou, ať už v pozitivním nebo negativním smyslu, ovlivnit další řešení projektu. Pokud jde o získávání a ukládání webových informačních zdrojů, ale zejména o jejich zpřístupňování z digitálního archivu, je třeba řešit legislativní otázky týkající se zákonů o povinném výtisku a otázky autorského práva.

2.1.1 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 dva 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.

2.1.2 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 by dle našeho mínění mohlo stačit jej upřesnit prováděcí vyhláškou, ve které by byl popsán proces odevzdávání zdrojů deponitní instituci. Problematickým by ovšem zůstal fakt, že zákon se vztahuje pouze na monografické publikace (jednorázově vydané publikace), kterých v dnešní době na Internetu mnoho nenajdeme.

2.1.3 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.

2.1.4 Novelizace zákonů

Děl publikovaných elektronickou cestou denně přibývá a pro jejich trvalé uchování je novela zákona o povinném výtisku nutností. Národní knihovna, která má za úkol uchovávat národní kulturní dědictví, usiluje proto o novelizaci potřebné legislativy. Příklady podobných opatření jsou patrné v mnoha ostatních zemích. Např. ve Velké Británii je již novela zákona o povinném výtisku připravena a je řádně projednávána v britském parlamentu. 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.

2.1.5 Autorský zákon

Dalším důležitým předpisem, dle kterého se musíme při tvorbě archivu řídit, je autorský zákon (z. 121/2000 Sb.). V tomto zákoně jsou z hlediska cílů projektu WebArchiv důležité zejména následující paragrafy:

§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 jakýmkoli prostředky nebo jakýmkoli způsobem.***

Těchto pět paragrafů rozhoduje velkým dílem o oprávnění Národní knihovny ČR, případně dalších depozitních knihoven k provozování činností v rámci WebArchivu. Vytváření archivu, tzn. stahování a ukládání elektronických online zdrojů je dle znění §37 legální.

Zvláštní kategorií děl jsou 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 by byl autorský zákon porušován. 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, poněvadž dle §38 smí knihovna půjčovat originály či rozmnoženiny vydaných děl. A právě slovo „vydaných“ je pro nás překážkou. Podíváme-li se na §4 a §14 zjistíme, že dílo je vydáno zahájením rozšiřování rozmnoženin a zároveň rozšiřováním rozmnoženiny se rozumí zpřístupňování díla v hmotné podobě. Online zdroje však nelze považovat za díla v hmotné podobě.

Z těchto důvodů bylo třeba hledat náhradní řešení, které by nám umožnilo zpřístupňování archivovaných zdrojů. Přistoupili jsme k oslovování jednotlivých vydavatelů a uzavírání smluv o poskytování elektronických online zdrojů. Ve smlouvě vydavatel souhlasí mj. se zpřístupněním svých zdrojů uložených v archivu (varianta lokálního nebo „veřejného“ zpřístupnění) a současně s tím, že zajistí informování autorů o této skutečnosti. Není to však řešení ideální, poněvadž je časově velmi náročné. Takto je možné oslovit několik desítek, možná stovek vydavatelů.

Východiskem z této 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. V současné době se musíme řídit stávajícím autorským zákonem, takže bohužel nemůžeme uplatnit formulaci ze směrnice, která dovoluje knihovně jednak zhotovování rozmnoženin nad rámec pouhé interní archivace či konzervace (čl. 5/2(c)), ale 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)).

2.1.6 Potud paragrafy. Jak se můžeme se současnou situací vyrovnat v praxi?

Zkušenost ze Švédska ukazuje důležitost takového právního zakotvení: automatická archivace (sklizení) online publikovaných elektronických zdrojů národní knihovnou, která je jedním ze způsobů, jak získat povinný výtisk tohoto typu dokumentů, zde musela být na mnoho měsíců přerušena právě proto, že úřady se nedokázaly shodnout na tom, zda je taková činnost legální a patovou situaci vyřešilo až přijetí příslušného zákona. Naopak v Dánsku, kde je podobný zákon v platnosti již delší dobu, nemá většina vydavatelů o jeho existenci ani tušení a jediným efektivním řešením problému je tak opět jedine automatická archivace všech publikovaných dokumentů.

Automatickou identifikaci a archivaci online publikovaných dokumentů lze srovnávat s běžně používanou technologií indexování webu, jak ji provádějí Internetové prohlížeče. Rozdílem je ovšem to, že ve webovém archivu zůstávají díla přístupná i po jejich změně nebo stažení vydavatelem. Existující infrastruktura je proto nastavitelná tak, že bude možné zachovat alespoň omezený rozsah sklizení i v případě, že by bylo nutné se podřídit určitým zákonným omezením. Jediným důsledkem takových omezení by pak bylo velmi výrazné zmenšení rozsahu sbírky, tvořené pak víceméně na základě dobrovolně dodávaných dokumentů. Na druhou stranu by se díky takovému zásahu výrazně zmenšila i finanční náročnost hardwaru pro uložení vzniklého archivu.

Mnohem problematičtější je samozřejmě oblast zpřístupnění archivu. Dokud nebude jasné stanoveno kdy, komu, v jakém rozsahu a za jakých podmínek může být takový archiv zpřístupňován, není možné navrhnout optimální nástroj pro daný účel. Pokud bychom totiž zpřístupňovali jen archiv omezeného rozsahu, bylo by možné bez velkých investic využít stávající infrastruktury, která je v Národní knihovně k dispozici. Pokud by naopak bylo umožněno bez omezení zpřístupňovat archiv celého českého webu, vyžádá si vybudování a provoz potřebné infrastruktury poměrně vysoké náklady. Ty by byly dány jednak rozsahem samotného archivu a tedy rozsahem přístupových souborů a jednak tím, že by o tuto službu byl pravděpodobně mezi uživateli českého Internetu velký zájem, což by kladlo vysoké nároky na výkonnost hardwaru. Dalším kritickým momentem jsou případná omezení, daná nějakým budoucím zákonem nebo soudním rozhodnutím. Taková rozhodnutí nelze ale ani při nejlepší vůli předjímat a je pravděpodobné, že každé takové rozhodnutí bude znamenat buď další finanční zátěž, nebo naopak zmaření části již investovaných prostředků.

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. V případě Webarchivu je tato kontrola zajištěna použitým systémem jednoznačných identifikátorů dokumentů na bázi kontrolního součtu. Ten však musí být podpořen dalšími opatřeními jak na úrovni technické, tak organizační. V případě podcenění tohoto aspektu se totiž můžeme dostat do téměř orwellovské situace, ve které nebude možné nijak ověřit autenticitu v minulosti publikovaných informací a dokumentů – ať už proto, že z veřejného prostoru zcela zmizí, nebo proto, že budou z různých důvodů pozměněna nebo zcela nahrazena jiným obsahem.

Že nejde o plané hrozby dosvědčuje i loňský případ serveru underground.cz, který stáhnul ze svých stránek všechny texty, zabývající se drogovou problematikou „v souladu s paragrafem 188a trestního zákona“. Ve svém důsledku to znamená, že veřejnost je možná dočasně, možná trvale, připravena o dříve publikované informace. Pokud by ale na druhé straně byly stejné informace zpřístupněny prostřednictvím archivu Národní knihovny, nebyla by to Národní knihovna, kdo by podstupoval riziko trestního stíhání?

2.2 Kritéria výběru zdrojů a strategie jejich archivace

Stanovení podmínek, které musí splňovat elektronické zdroje kandidující na včlenění do budovaného digitálního archivu, je jedním z nejkritičtějších okamžiků každého podobného

projektu. Při stanovování těchto podmínek je nutné brát v úvahu jak objem finančních prostředků, které jsou pro tuto činnost k dispozici, tak i aktuální stav rozvoje celé oblasti informačních a komunikačních technologií.

Vzhledem k překážkám souvisejícím s legislativou je v současné době projekt zaměřen tak, aby zajistil zpřístupnění alespoň vybraného souboru nejvýznamnějších zdrojů, ale současně aby zajistil uchování „českého webu“. To představuje dvě souběžné cesty:

- a) intelektuální činnost - shromažďování, registraci a archivaci vybraných domácích elektronických online dostupných zdrojů jako legitimní součásti národní publikační produkce podle stanovených kritérií výběru pro účely České národní bibliografie,
- b) automatizovaný proces shromažďování a archivace domácích zdrojů z Internetu v relativní úplnosti (servery v národní doméně .cz a další omezení) pomocí speciálních programů.

2.2.1 Výběr nejvýznamnějších zdrojů jako intelektuální činnost

V prvním případě se jedná o výhradně intelektuální činnost, při níž zpracovatelé vyhledávají na Internetu významné informační zdroje. Cílem této činnosti je zajistit bibliografickou kontrolu těchto zdrojů (bibliografický popis zdrojů a národní registrující bibliografie) a současně zajistit přístup k těmto zdrojům – plným textům dostupným na webu, ale též umožnit využívání archivovaných zdrojů v případě, kdy již původní dokument není na webu dostupný. Vzhledem k tomu, že se jedná o zpřístupňování zdrojů uložených v digitálním archivu, je třeba zajistit k nim i legální přístup, tedy uzavřít na všechny tyto zdroje výše zmíněné smlouvy s vydavateli.

Výběr zdrojů je do značné míry závislý na „vkusu“ zpracovatele. Individuální přístup by měla co nejvíce minimalizovat *kritéria výběru*. Při formulaci kritérií, podle nichž budou vybírány zdroje, které budou zařazovány do České národní bibliografie, se vycházelo ze strategií archivace webových zdrojů přijatých v rámci obdobných zahraničních projektů (zejména projektu National Library of Australia PANDORA), ovšem s přihlédnutím ke specifické situaci v České republice. Byla stanovena následující kritéria:

▪ podle místa uložení zdroje

Primárně jsou brány v úvahu zdroje přístupné na serverech s doménou prvního stupně .cz. V této souvislosti však vyvstává problém, jak správně vymezit tzv. národní web (tj. zda uplatňovat pouze teritoriální hledisko nebo také jazykové hledisko podobně, jako je tomu u tradičních bohemikálních dokumentů). Faktem zůstává, že není možné výše uvedenou podmínku za všech okolností striktně dodržet, protože v některých případech čeští vydavatelé záměrně nebo nuceně (obvykle z důvodu předchozí registrace žádané domény ze strany spekulantů) využívají servery s doménami .com, .net, .org a další. V těchto případech je třeba identifikovat vlastníka domény druhého stupně pomocí specializovaných služeb. Stejně zkušenosti byly získány při automatickém sběru švédských webových zdrojů v rámci projektu Kulturarw³ – bylo zjištěno, že až 40 % zdrojů je uloženo na serverech mimo národní doménu .se.

Podle obsahu zdroje

Jsou brány v úvahu zdroje odborného nebo uměleckého charakteru, u nichž se předpokládá, že mají informační hodnotu pro větší okruh budoucích uživatelů. Záměrně jsou pominuty zdroje, které jsou výsledkem soukromých, firemních nebo ryze reklamních publikačních aktivit, i když s vědomím, že i v této oblasti se mohou vyskytovat zdroje, které mohou být pro některé uživatele zajímavé, resp. zdroje, které nejsou jiným způsobem zveřejněny.

Podle typu zdroje

Repertoár typů zdrojů je poměrně pestrý a je do jisté míry ovlivněn předchozím kritériem. Při jejich výběru se vychází z běžných klasifikací dokumentů. Jde především o seriály, konferenční příspěvky, výzkumné a jiné zprávy, studie vzniklé např. jako výstupy vědeckých

a výzkumných projektů, akademické práce, dokumenty veřejné správy. Je příznačné, že tyto zdroje spadají do kategorie tzv. šedé literatury.

Podle formy

Jsou brány v úvahu ty zdroje, které jsou publikovány pouze v elektronické formě (online), aby se zabránilo duplicitě zpracování webových zdrojů a tradičních (tištěných) dokumentů s identickým obsahem.

Podle přístupu

Jsou brány v úvahu pouze ty zdroje, které jsou volně přístupné, to znamená, že nejsou k dispozici v rámci placených informačních služeb.

Podle formátu

Z pragmatických důvodů jsou preferovány formáty, které jsou všeobecně podporovány producenty aplikačního softwaru (zejména webových prohlížečů), nikoliv tedy proprietární formáty, pro jejichž korektní zobrazení je třeba zvláštní aplikační software. Některé z těchto formátů se ovšem staly – díky dominantnímu postavení producenta na trhu – standardy elektronického publikování de facto (např. Adobe – *pdf*, Microsoft – *doc*). Archivaci webových zdrojů usnadňuje empiricky dokázaný fakt (harvesting /stahování/ českého, švédského, nizozemského a finského webu), že navzdory velkému množství formátů, se kterými se na webu můžeme setkat, je většina webových zdrojů (85 až 90 procent) uložena v malém počtu formátů (resp. MIME podtypů) – *html/htm* (k tomu připojme *asp* a *php* v případě dynamických webových informačních systémů), *jpeg*, *gif* (pro statickou grafiku) a *txt*. Zastoupení zdrojů ve formátech *pdf*, *doc*, *rtf* a *ps* (PostScript) na českém webu není výrazné, ale jejich informační hodnota je obvykle vyšší než u zdrojů ve formátu *html*.

Sběr vybraných informačních zdrojů z webu a jejich archivace se provádí v rámci dále popsaneho automatizovaného procesu.

Ad b) Automatizovaný proces shromažďování a archivace

V případě automatizovaného sběru webových zdrojů se často hovoří o „plošném sběru“ českého webu. Ani v tomto případě se neshromažďuje absolutně vše, co si lze představit pod českou produkcí webových zdrojů, ale i zde jsou četné aspekty ovlivňující sběr a archivaci, tedy opět jakási kritéria výběru.

Protokoly, formáty

Pokud padla v úvodu tohoto článku zmínka o „online“ publikovaných zdrojích, je nutné upozornit na to, že tento pojem je hyperonymem pojmu „na Internetu“ (ačkoli se s rostoucí dominancí Internetu stávají tyto pojmy postupně synonymickými podobně jako pojem „web“ začíná splývat s pojmem „Internet“). Z toho mimo jiné vyplývá, že již rozhodnutím zaměřit se na webové zdroje pomíjíme jistou část elektronických zdrojů. Z ne-internetových zdrojů lze zmínit například počátkem 90. let i u nás poměrně rozšířený a dnes již téměř zapomenutý FidoNet, ze zdrojů internetových pak mezi jinými například streamované audio a video, obsah různých sítí peer-to-peer a mnohé další zdroje, dostupné výhradně přes některý z méně rozšířených komunikačních protokolů.

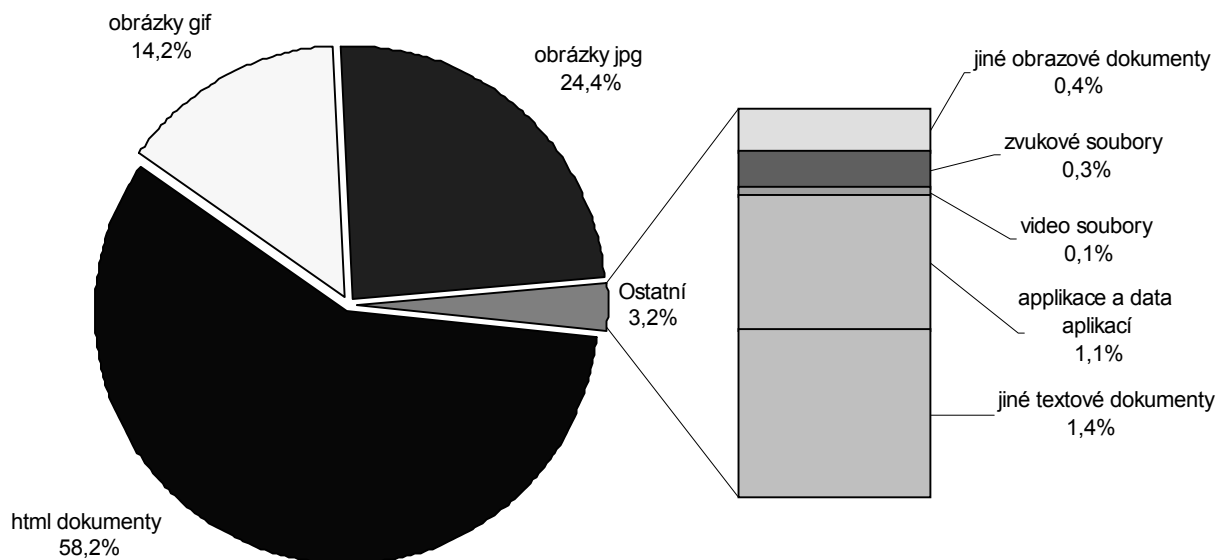
Je zřejmé, že pokus archivovat online elektronické zdroje, dostupné jinak, než prostřednictvím Internetu, by byl velmi nákladný a jeho přínos by byl mizivý. Takové jednoznačné tvrzení již však nelze pronést, máme-li na mysli newebové internetové zdroje. Většinou totiž dopředu nelze určit, která technologie začne mít v budoucnosti význam a která je jen drobnou epizodou v dějinách Internetu (tak skončily v propadlišti dějin technologie typu „push“, kterým byla kdysi prorokována velká budoucnost, tak se naopak objevuje přes aktivní odpor zábavního průmyslu stále větší množství různých typů sítí peer-to-peer). Přesto lze zatím stále obhájit názor, že většině populace je reálně přístupná jen ta část zdrojů, ke kterým se dostanou prostřednictvím běžného prohlížeče. Pokud tedy pomineme relativně velkou množinu mailových a newsových diskusních skupin, zůstává před námi dvojice

protokolů http a ftp (protokol gopher lze již považovat za mrtvý, protokol https pak díky tomu, že je určen pro šifrovaný přenos dat, za protokol určený k přenosu neveřejných a důvěrných informací, tedy informací, dostupných sice elektronicky, ale ne nutně veřejně).

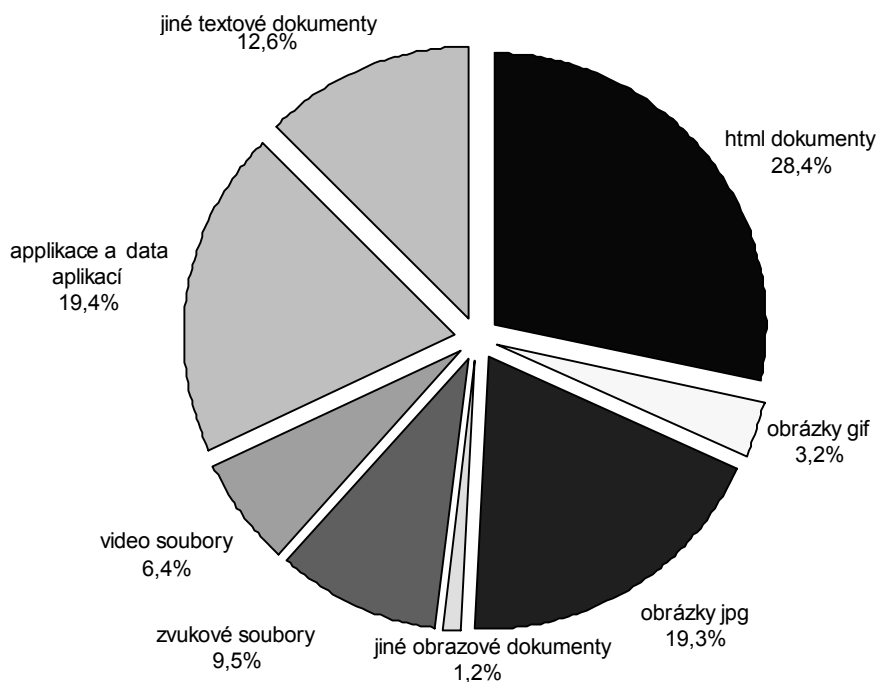
Pominuli-li jsme v předchozím odstavci diskusní skupiny, bylo to především proto, že archivy mnoha z nich jsou zároveň přístupné na webu. Pokud by se ukázalo, že je důležité vytvářet jejich archiv, nabízí se k tomu standardní prostředek – instalace news serveru, který bude zrcadlit české diskusní skupiny a bude si udržovat celou jejich historii.

Podobně jako v případě protokolů bychom mohli jednotlivé dokumenty hodnotit i co do použitého formátu. Grafy 1 a 2 ukazují, jak jsou v současnosti v archivu zastoupeny jednotlivé formáty souborů. Je vidět, že trojice formátů html, jpg a gif tvoří dohromady 96,8% počtu všech archivovaných souborů, ačkoli co do velikosti zauímají jen polovinu celkového objemu uložených dat. Pokud tedy dokážeme odpovědně určit, které z vzácně se vyskytujících formátů nemá smysl z různých důvodů archivovat, můžeme snadno ušetřit až třetinu objemu ukládacího prostoru, což může snadno představovat úsporu statisícových částek.

Graf 1: Relativní četnost souborů v archivu podle typů



Graf 2: Zastoupení hlavních typů souborů v archivu podle velikosti



Prostorový a časový rozsah

Jak již bylo uvedeno, je předmětem zájmu projektu Webarchiv archivace online publikované části národního kulturního bohatství, tedy, zjednodušeně řečeno, český web (ať už je definován jakkoli). V ideálním případě by měl být výsledkem projektu archiv, obsahující vše, co kdy bylo v rámci českého webu publikováno. Je ale zřejmé, že takový archiv by byl prakticky nerealizovatelný. Je proto nutné stanovit taková kritéria, která by umožnila zachytit v daném časovém úseku to nejvýznamnější, co český web nabízí.

Na jedné straně je možné pokusit se s delším časovým odstupem vytvářet co nejúplnější a rozsahem co nejpodrobnější časové snímky celého českého webu, na straně druhé pak budovat pravidelně (v případě potřeby i každý den) doplňovaný archiv zrcadlící vybranou skupinu zdrojů. Místo hledání kompromisu mezi těmito dvěma přístupy jde řešitelský tým zároveň jak cestou extenzivní, tak i intenzivní.

Aby bylo možné tyto postupy realizovat, je nutné nejprve stanovit, jaký je vlastně rozsah českého webu. Ačkoli jej můžeme zjednodušeně definovat jako

1. všechny dokumenty publikované v doméně .cz

Je zřejmé, že toto kritérium nemůže pokrýt celou českou online produkci. Proto by bylo vhodné tento rozsah rozšířit o mnoho dalších, vzájemně se prolínajících kategorií:

2. dokumenty v doménách druhé úrovně, registrovaných na subjekt se sídlem v České republice
3. dokumenty publikované na serverech fyzicky umístěných v ČR
4. dokumenty v českém jazyce
5. dokumenty českých autorů
6. dokumenty se vztahem k Česku

Z uvedeného seznamu je patrné, že již na počátku velká zájmová oblast by se tímto způsobem dala zvětšovat téměř neomezeně. Je také vidět, že se stoupajícím pořadím podmínek stoupá jak náročnost nalezení všech dokumentů, podmínku splňujících, tak i náročnost prokázání, že nalezený dokument danou podmínku splňuje.

Už získání údajů o rozsahu domény .cz (1) není triviální. Správce domény nejvyšší úrovně .cz, sdružení CZNIC (www.nic.cz), sice na svých stránkách zveřejňoval přehledové statistiky (z těch plyne, že celkový počet registrovaných domén druhé úrovně se nyní pohybuje okolo 140.000), ale kompletní seznam domén druhé úrovně nezveřejňoval a nezveřejňuje. Naštěstí je zatím možné tento seznam standardní cestou získávat z jednoho ze zahraničních sekundárních jmenných serverů pro doménu .cz a je tak možné zkoumat, nakolik účinná je druhá cesta, vedoucí k získání těchto údajů, používaná všemi webovými roboty, tedy i Nedlib Harvesterem. Tato metoda je založena na extrakci odkazů vedoucích na další dokumenty během postupného procházení všech webových stránek, splňujících zadaná kritéria.

V případě (2) je situace ještě složitější: v ideálním případě by bylo nutné získat a analyzovat kompletní seznamy domén nejvyšší úrovně a pak postupovat stejným způsobem jako finský tým, který analyzoval adresy a telefonní čísla vlastníků jednotlivých domén a automaticky rozšířil databázi adres pro sklizení o adresy, patřící finským vlastníkům.

Obdobně v případě (3) je možné se pokusit zjistit co nejpřesněji rozsahy IP adres, používaných našimi primárními poskytovateli připojení (tj. členy sdružení Neutral Internet eXchange, www.nix.cz). O tyto adresy by pak mohla být obohacena databáze povolených adres. Tím zajistíme, že při sklizení nebudou vynechány ty servery, na které není odkazováno jménem, ale jen IP adresou. Teoreticky bychom sice mohli i aktivně skenovat tyto rozsahy IP adres a hledat tak spuštěné www servery, není ale pravděpodobné, že by takto nalezené neregistrované servery obsahovaly hodnotné informace. Tento přístup se však s rozšiřováním globálních virtuálních sítí stává stále méně spolehlivým.

V případě (4) je situace složitější: procházení celosvětového webu s cílem najít stránky v českém jazyce je sice technicky realizovatelné, zároveň však neefektivní. Je ale možné, že v budoucnu půjde tento problém alespoň částečně vyřešit ve spolupráci s dalšími institucemi zabývajícími se archivací webu tak, že všechny dokumenty stažené danou národní institucí budou podrobeny automatické analýze pro rozpoznání jazyka a odkazy na nalezené stránky v cizím jazyce by byly předány příslušné národní instituci. Výhodou pro nás může být to, že neexistuje druhá země, jejímž národním jazykem by byla čeština – nemusíme tedy řešit problémy které mají anglofonní země nebo například portugalská národní knihovna ve vztahu k brazilským webům.

Na rozdíl od výše uvedených bodů by v případech (5) a (6) bylo velmi obtížné, ne-li nemožné automaticky rozhodnout, zda daný dokument spadá do zájmové oblasti. Zde bude záležet především na knihovnicích nebo na vydavatelích samotných, zda takový server nebo dokument zaregistrují.

Předpokládáme proto, že rozšiřování oblasti zájmu mimo doménu .cz bude probíhat pomalu, spíše po jednotlivých serverech, nebo pouze jednotlivých dokumentech. Určitým usnadněním a urychlením tohoto procesu by snad mohla být analýza těch dokumentů, uložených mimo doménu .cz, na které vedou odkazy z této domény. Zde by se mohlo efektivně uplatnit i automatické rozpoznání jazyka dokumentu.

Stanovili-li jsme si tedy alespoň přibližně rozsah českého webu, můžeme v jeho rámci začít hledat takovou podmnožinu zdrojů, kterou by bylo vhodné archivovat výběrově v co největší úplnosti. V současné době se 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 budou totiž průběžně aktualizované oborové brány do světa elektronických zdrojů. Správa jednotlivých oborů tohoto portálu bude svěřena knihovně, která má v daném oboru největší zkušenosti. Díky tomu lze očekávat, že každý obor bude reprezentován i nejvýznamnějšími národními informačními zdroji, které se následně stanou i předmětem zájmu projektu WebArchiv.

Je zřejmé, že takto pojatý systém může mnoho serverů neoprávněně vyloučit, na druhé straně je nutno mít na zřeteli to, že každý zdroj zahrnutý do skupiny pro intenzivní výběrové sklizení s sebou nese nemalý díl kvalifikované lidské práce spojené s jeho knihovnickým popisem, který může ve vybraných případech jít až na úroveň jednotlivých dokumentů (článků v časopisech, příspěvků ve sbornících apod.). Finanční náročnost může být v takovém případě samozřejmě snížena, dojde-li k nějaké formě dohody o spolupráci s příslušným vydavatelem.

Vliv technického řešení na rozsah a průběh sklizně

Volbou nejvhodnějšího nástroje pro plošnou archivaci webu se v současné době zabývá několik projektů v různých evropských zemích, za všechny lze zmínit testování v Rakousku nebo v Dánsku. U nás používaný produkt, NEDLIB Harvester, vyvinutý Helsinskou národní knihovnou, v těchto srovnávacích testech rozhodně nezaostává. Díky tomu, že byl navržen pro potřeby archivace webu národními knihovnami, vyhovuje nejlépe i našim požadavkům. Mezi možnosti jeho nastavení patří volba seznamu výchozích webových stránek, omezení rozsahu sklizně pomocí URL nebo jejich částí, povolení nebo zakázání podpory protokolu ftp, logování zamítnutých URL, akceptování omezení pro roboty na jednotlivých serverech (robots.txt), podpora URL s parametrem nebo maximální hloubka zanoření v rámci jednoho serveru. Zvláště poslední dva parametry pak mohou velmi významně ovlivnit rozsah a kvalitu sklizně.

Podpora URL s parametry umožňuje omezit sklizení jen na ta URL, která neobsahují znak '?' uvozující seznam parametrů. Díky tomu lze sice do značné míry zabránit problémům spojeným s nekonečnými smyčkami při procházení serverů, na druhou stranu se tak nepříjemně omezuje rozsah sklizně. Jako typický příklad lze uvést server *lupa.cz*, jehož jedinou stránkou, na kterou se dá dostat pomocí URL bez parametru, je jeho hlavní stránka. Protože podobně funguje většina elektronických periodik, vyřadili bychom ignorováním URL s parametry právě ty zdroje, které jsou z hlediska našeho kulturního dědictví nejcennější.

Je samozřejmě pravděpodobné, že mnohé dynamicky generované stránky se v archivu vyskytnou několikrát jen proto, že se navzájem nepatrně liší. Může se tak stát, že se opakovaně archivují již navštívené stránky jen proto, že součástí URL je například identifikátor sezení, nebo aktuální čas. Takový cyklus se pak opakuje tak dlouho, dokud není vyčerpán povolený počet zanoření v rámci jednoho serveru (nyní se operuje s hodnotou 50, která by měla zajistit stažení všech stránek z většiny serverů). Je však nutno poznamenat, že k podobným problémům dochází pouze v případě, kdy správce daného serveru ve vlastním zájmu v souboru robots.txt nezakáže všem robotům přístup na problematická URL.

Je zřejmé, že ať už je pro archivaci webu zvolen jakýkoli produkt, bude jím vytvořený archiv poplatný jeho limitům. Ani NEDLIB Harvester není v tomto směru samozřejmě výjimkou a tak existuje několik prozatím nepřekročitelných omezení. Jeho nejbolestivějším omezením je absence podpory javascriptu. V důsledku toho v archivu zcela chybí stránky, na něž vedou jen odkazy generované javascriptem až v prohlížeči (typickým příkladem takových odkazů jsou odkazy do archivu Neviditelného psa). Zatím méně palčivým nedostatkem stejného charakteru je absence podpory odkazů z prezentací ve formátu flash.

Přestože je NEDLIB Harvester nyní používán pro archivaci webu v mnoha zemích, ukazuje se, že jeho éra se postupně chýlí ke konci. Již v únoru příštího roku bude totiž pravděpodobně zveřejněna první verze harvesteru vzešlého z vývoje konsorcia okolo organizace Internet Archive a pokud to licenční podmínky dovolí, brzy tento nástroj asi převáží.

Sklizeň českého webu

V loňském roce probíhala po několik měsíců již druhá testovací sklizeň celé domény .cz. Tato sklizeň měla ukázat mimo jiné i to, jaký je skutečný rozsah českého viditelného webu. Výchozími body pro tuto sklizeň byly především hlavní stránky internetových portálů seznam.cz a quick.cz. Přes různé problémy se podařilo stáhnout z 10.490.000 URL celkem

10.090.000 souborů o celkové velikosti přes 240 GB. Alespoň jednou přitom bylo navštíveno přibližně 30.000 domén 2. úrovně (tj. čtvrtina domén v doméně .cz).

Analýza této sklizně ukázala, jaké informační bohatství český web skrývá: mezi padesáti našimi objemem nebo počtem souborů největšími doménami druhé úrovně bylo mimo jiné šest univerzit, jeden univerzitou provozovaný specializovaný server (linux.cz), Česká akademie věd a několik zpravodajských a vydavatelských serverů. Dále jsou na předních místech zastoupeny především webhostingové farmy, které sice přinášejí jen minimum vlastního obsahu, ale o to větší rozmanitost.

V letošním roce byla tato sklizeň doplněna o plošnou sklizeň domény .cz, zaměřenou přednostně na pokrytí co největšího počtu serverů. Ze 140.000 registrovaných domén druhé úrovně byl tak nalezen běžící www server na 120.000 serverech „www.doména.cz“ a na 76.000 serverech „doména.cz“. Takto zjištěné adresy byly doplněny analýzou registrovaných adres třetí úrovně na celkem 393.000 adres potenciálních webových serverů v doméně .cz. Takto získaný seznam byl poté zadán harvesteru pro stažení dokumentů z nejvyšší úrovně každého serveru a po jejich analýze bude pokračovat jejich plošnou sklizní.

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. Díky tomu, že v současné době již jsou na trhu levné pevné disky o kapacitách téměř 300 GB, je možné pořídit i cenově dostupná disková pole s kapacitou v řádu jednotek TB.

V pilotní fázi projektu bylo s výhodou využito stávajícího páskového robota Národní knihovny ČR, jehož nevýhodou ovšem byla problematická dostupnost na něm uložených dat. Proto byl celý webový archiv přesunut na diskové pole, které má dostatečnou kapacitu pro jednu nebo dvě další celoplošné sklizně. Lze předpokládat, že v budoucnu bude podobná migrace na nové úložiště již rutinní záležitostí.

Větším oříškem samozřejmě z dlouhodobého hlediska budou 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 zatím zůstává 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.

Bibliografická správa a zpřístupnění zdrojů

Prvním krokem při zpřístupňování archivu musí samozřejmě být alespoň rámcové stanovení použitých postupů, s nimi spojených pracovních procesů a jejich rozsahu. Je zřejmé, že ani společným úsilím všech českých knihoven nebude nikdy možné zkatalogizovat celý archiv českého webu – tento úkol bude nutné přenechat „strojům“. Přes značný pokrok v oblasti počítačového porozumění přirozenému jazyku v posledních letech bude pravděpodobně ještě řadu let trvat, než bude možné začít uvažovat o nasazení plně automatizovaného nástroje pro bibliografický popis archivovaných dokumentů.

Fulltext

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.

V červnu letošního roku, po 15 měsících práce, byl program dokončen a nyní je zprovozněn na serveru projektu WebArchiv. Tento program sice indexuje pouze html soubory, ale jeho API je otevřen dalším formátovým modulům. Program zohledňuje českou gramatiku i diakritiku. Podporuje 9 vyhledávacích kategorií včetně fulltextu, zvýrazněného textu a metadat. Bohužel tento software nepodporuje žádný z hlavních knihovnických standardů, které se při vývoji zdály jeho autorům nevýznamné. Z tohoto důvodu budeme nyní nuceni sami doplnit podporu pro Z39.50, OAI-PMH nebo jiný podobný standard, protože není pravděpodobné, že by se autoři zajímali o další podporu a vývoj tohoto produktu.

Nadějně se proto jeví i kontakty s týmem Norské národní knihovny, která v rámci projektu Nordic Web Archive 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.

Požadavek na zpřístupnění archivu s sebou přináší nutnost investovat každým rokem částku v řádu nejméně statisíců korun do hardwarového vybavení a další velké částky do softwaru a/nebo lidských zdrojů (vývoj, správa, ale i popis zdrojů apod.). Do doby, než budou takové finanční částky dostupné, bude nutné se snažit najít méně nákladná řešení, která by zpřístupnila alespoň to nejdůležitější, co archiv nabízí.

Metadata

Jedním z takových řešení je využití faktu, že někteří autoři a vydavatelé mají zájem nebo jsou ochotni vkládat do publikovaných dokumentů metadata, daný dokument popisující. Nejrozšířenějším standardem na tomto poli, pomineme-li obecná klíčová slova, jsou metadata standardu Dublin Core. Proto byla již v rámci pilotního projektu vybudována infrastruktura, zaměřená na podporu využívání metadat DC u nás. Tato infrastruktura by měla usnadnit zapojení autorů a vydavatelů do procesu tvorby a zveřejňování metadat již v okamžiku publikování dokumentu.

Nejdůležitější částí této infrastruktury je Dublin Core Metadata Generator. Tento nástroj, veřejně přístupný na serveru projektu, umožňuje autorům webových stránek poloautomaticky nebo ručně vytvořit, editovat, konvertovat a ve zvolené syntaxi uložit metadata respektující pravidla kvalifikovaného Dublin Core (ta byla v rámci pilotního projektu přeložena do češtiny a zveřejněna na českých stránkách iniciativy Dublin Core – http://www.ics.muni.cz/dublin_core/).

Dublin Core Metadata Generator byl původně společně s dalšími nástroji převzat s minimálními úpravami od Helsinské univerzitní knihovny, která jej vyvinula v rámci projektů Nordic Metadata I a II (<http://www.lib.helsinki.fi/meta/>). Na základě výsledků zkušebního provozu byl program postupně upravován až do dnešní podoby.

Nyní je možné volit kvalifikátory prvku Subject tak, aby odpovídaly u nás používaným systémům věcného třídění. Byla také doplněna funkce automatického vložení jedinečného čísla národní bibliografie ve formátu URN přímo do pole Identifier, pokud bylo toto pole předtím prázdné, což zajišťuje uživateli větší pohodlí a výrazně zmenšuje riziko chyb, hrozících jinak při kopírování nebo přepisu identifikátoru.

Zmíněné přidělení jednoznačného identifikátoru je umožněno propojením Dublin Core generátoru s generátorem URN. Již nyní je sice dostupné rozhraní systému přidělování URN tak, že program přidělující URN funguje jako samostatný URN server, což umožňuje jeho snadnou integraci přímo do publikačních systémů vydavatelů online zdrojů. Dalším krokem pak bude delegování přidělování URN jednotlivým dokumentům na vydavatele, který bude muset garantovat jejich jednoznačnost a trvalost v rámci svého publikačního systému. Díky tomu by se přidělování URN mělo stát zcela automatickým procesem.

Řadu pomůcek dostupných na serveru webarchivu doplnil i kalkulátor MD5. Ten umožňuje spočítat kontrolní součet MD5 zadaného textového řetězce. Pokud je tímto řetězcem platné URL nějakého dokumentu, může kalkulátor tento dokument stáhnout a spočítat jeho kontrolní součet. Protože jsou tyto kontrolní součty používány pro identifikaci dokumentů,

archivovaných Harvesterem, je jedna z možností využití Kalkulátoru zřejmá: může sloužit jako pomůcka při analýze práce Harvesteru i při zkoumání archivu samotného.

Národní bibliografie

V červnu letošního roku byla v elektronickém katalogu Národní knihovny ČR zpřístupněna báze WEB. Jedná se o *bázi bibliografickou*, obsahující záznamy elektronických zdrojů vybraných na základě výše zmíněných selekčních kritérií (obdobných kritériím výběru tradičních druhů dokumentů pro konzervační fond). Výběr zdrojů je zaměřen na jejich trvalé uchování jako kulturního dědictví a současně k účelu jejich registrace v České národní bibliografii. V rámci silně limitovaných pracovních kapacit na tuto činnost je proto snahou řešitelů zařadit do souboru vybraných zdrojů pokud možno ty, které si z hlediska svého obsahu a/nebo formy zařazení mezi „kulturní dědictví národa“ skutečně zaslouží. Záznamy jsou samozřejmě ve formátu MARC (nyní UNIMARC, v budoucnu MARC21).

Vedle funkce bibliografické slouží báze také *zpřístupňování zdrojů* popsaných v databázi bibliografickým záznamem. Prostřednictvím zapsané URL adresy je umožněn přístup ze záznamu přímo do zdroje přístupného na internetu. Jakmile bude k dispozici funkční vyhledávací nástroj, měl by být ze záznamů v databázi umožňován také přístup do digitálního archivu (WebArchiv). Tento přístup bude umožňován v souladu s platnou legislativou, což znamená v současné době uzavírání smluv s autory/vydavateli internetových zdrojů.

Současně bude báze WEB využívána při tvorbě oborových bran provozovaných pod hlavičkou Jednotné informační brány CASLIN. Na tyto zdroje, kterých budou řádově desítky, nejvýše pak stovky, by případně mohlo navázat kooperativní analytické zpracování vybraných článků, opět standardním způsobem ve formátu MARC. Pomůckou od vydavatele by zde samozřejmě mohlo být vložení metadat přímo do zdrojového textu článku. Metadata Dublin Core jsou zde předpokladem pro vzájemnou komunikaci se zahraničními oborovými branami.

Mezinárodní spolupráce

Je patrné, že práce na poli zpřístupnění archivu jsou dlouhodobou záležitostí, která vyžaduje nemalé prostředky a nasazení. Jednou z cest, jak tyto prostředky získat, je spolupráce na mezinárodní úrovni, která se velmi osvědčila již během řešení pilotního projektu.

Spolupráce s Helsinskou univerzitní knihovnou, která započala převzetím u nich vyvinutých nástrojů (NEDLIB Harvester, Dublin Core Metadata Generator, URN Generator), pokračovala dále spoluprací na jejich dalším vývoji – všechny opravy a úpravy, které byly v průběhu řešení projektu na převzatých programech provedeny, byly poskytnuty i finskému týmu, který se zaměřil především na další vývoj Harvesteru.

Dalším souvisejícím krokem na poli mezinárodní spolupráce bylo pak navázání kontaktů s týmem Technické univerzity ve Vídni, řešícím problematiku archivace rakouského webu ve spolupráci s Rakouskou národní knihovnou, a úspěšná prezentace dosavadních výsledků projektu na mezinárodním fóru.

Národní knihovna ČR a Moravská zemská knihovna v Brně se zapojily do iniciativ národních knihoven a dalších organizací z evropských zemí, usilujících o získání podpory Evropské unie na sjednocení roztržité národní iniciativy jednotlivých evropských zemí a vytvoření distribuovaného archivu evropského webu, založeného na síti národních archivů jednotlivých zemí. Bohužel tato iniciativa nebyla prozatím úspěšná, přesto doufáme, že se v budoucnosti podaří získat podporu i na této úrovni.

Závěr

Jaká je perspektiva projektu?

Zda bude některá z dosud popisovaných technologií nasazena také v ostrém reálném provozu, bude záviset nejen na řešení technickém, ale i na vyřešení autorskoprávní

problematiky související s tvorbou a provozem archivu. Nedotaženost zákona o povinném výtisku u nás otevírá cestu různým výkladům omezení daných zákonem o autorském právu. Proto je jednou z priorit řešitelů projektu sledování a aktivní připomínkování připravovaných legislativních dokumentů.

Ačkoli je díky vytvořené infrastruktuře již nyní možné udělat mnohé pro zachování dnešních informačních zdrojů pro budoucí generace, další rozvoj této infrastruktury, stejně jako vývoj v podstatě všech softwarových produktů, nemůže být nikdy zcela ukončen. Zde nejde jen o hledisko potřeb uživatele nebo provozovatele, ale i o hledisko technického vývoje, mezinárodní spolupráce nebo problematiku legislativní. S tím, jak bude stoupat podíl čistě elektronické produkce, bude růst i význam její dlouhodobé archivace z hlediska ochrany národního kulturního dědictví.

Je patrné, že práce na poli zpřístupnění archivu budou dlouhodobou záležitostí, která si vyžádá nemalé prostředky. Jednou z cest, jak tyto prostředky získat, je spolupráce na 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.

Další informace k tématu, publikace řešitelů i odkazy na zahraniční informační prameny najdete na webových stránkách na serveru WebArchiv: <<http://webarchiv.nkp.cz>>.