Univerzita Karlova v Praze
Filozofická fakulta
Ústav informačních studií a knihovnictví

Diplomová práce

2005 **Markéta Škodová**

Univerzita Karlova v Praze

Filozofická fakulta
Ústav informačních studií a knihovnictví

Studijní program: informační studia a knihovnictví
Studijní obor: informační studia a knihovnictví

Markéta Škodová

Strategie archivace elektronických online zdrojů a politika jejich výběru do digitálního archivu (se zaměřením na český systém WebArchiv)

Diplomová práce

Praha 2005

Vedoucí diplomové práce: PhDr. Eva Bratková

Oponent diplomové práce: Mgr. Ludmila Celbová

Datum obhajoby:

Hodnocení:

Prohlášení:

Prohlašuji, že jsem diplomovou práci zpracovala samostatně a že jsem uvedla všechny použité informační zdroje.

V Praze, 8.září 2005

……..

podpis diplomanta

Identifikační záznam

ŠKODOVÁ, Markéta. *Strategie archivace elektronických online zdrojů a politika jejich výběru do digitálního archivu (se zaměřením na český systém WebArchiv) [Web archiving strategy and selection policy of electronic online resources for digital archive (focus on Czech system WebArchiv)],* Praha, 2005. 86 s., 16 s. příl. Diplomová práce. Univerzita Karlova v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví 2005. Vedoucí práce PhDr. Eva Bratková.

Abstrakt

Tématem diplomové práce je archivace elektronických online zdrojů na národní úrovni. Od 90. let minulého století se touto činností začalo zabývat několik málo organizací, mezi které řadíme především národní knihovny. Ty mají za cíl zejména ochránit národní kulturní dědictví v digitální podobě.

V úvodu práce je objasněn celý proces archivace a jeho dílčí operace (registrace a výběr zdrojů, vlastní archivace a zpřístupnění). Další část je věnována analýze dostupných metod archivace online zdrojů a na jejím základě formulován návrh optimální metody pro český projekt WebArchiv. Samostatná část práce je zaměřena na složitou otázku definování kritérií výběru zdrojů do digitálního archivu. Následuje nástin východisek pro politiku výběru zdrojů WebArchivu. V závěru je pozornost soustředěna na současný stav archivace webu na národní úrovni a její možné perspektivy.

Klíčová slova

archivace webu, elektronické online zdroje, webové dokumenty, digitální archivy, harvesting, kritéria výběru, WebArchiv

OBSAH

PŘEDMLUVA

Značná část informací, se kterými se dnes setkáváme, existuje pouze v digitální formě. Pro kulturní instituce, tradičně zodpovědné za shromažďování a archivaci kulturního dědictví, se tak otázka uchování elektronických online zdrojů stává velice aktuální.

Cílem této diplomové práce je především seznámení s celou komplexní problematikou archivace online zdrojů na národní úrovni, analýza známých metod archivace, zmapování politiky výběru zdrojů do digitálního archivu a následné vyhodnocení nastínění postupů v rámci českého projektu Webarchiv.

Výběr tématu diplomové práce se odvíjel z mé dosavadní praxe. Jelikož třetím rokem pracuji na projektu WebArchiv v Národní knihovně ČR, je mi zvolené téma velmi blízké.

Při psaní diplomové práce jsem vycházela z volně dostupných, vesměs internetových zdrojů věnovaných této problematice. Jednalo se zejména o články z časopisů, konferenční příspěvky a webové stránky jednotlivých zahraničních projektů či iniciativ. Neméně důležitým okruhem zdrojů byly mé vlastní zkušenosti z praxe a možnost konzultace s odborníky z oboru.

Diplomová práce je rozčleněna do sedmi kapitol. Po úvodu následuje druhá kapitola, která pohlíží na proces archivace a její dílčí operace (registrace a výběr zdrojů, příprava dokumentů na archivaci, vlastní archivace a zpřístupnění) nejen z hlediska knihovnického, ale i technického i legislativního. Třetí a čtvrtá kapitola jsou věnovány metodám archivace a jejich komparaci, a to nejprve z celosvětového hlediska, následně v kontextu českého projektu WebArchiv. Pátá kapitola analyzuje dosavadní kritéria výběru zdrojů do digitálního archivu a v šesté kapitole je snaha nastíněnou politiku výběru zdrojů aplikovat v českém prostředí WebArchivu. V závěru je pozornost věnována současnému stavu archivace webu na národní úrovni a jejím možným perspektivám.

Použitá literatura je citována v souladu s normou ISO 690 a ISO 690-2. Citované zdroje jsou uvedeny v abecedním pořadí.

Na tomto místě bych ráda poděkovala vedoucí mé diplomové práce PhDr. Evě Bratkové a řediteli ÚISK PhDr. Richardu Papíkovi, Ph.D. za vstřícný přístup při jejím

zpracování. Za cenné rady a připomínky děkuji také Mgr. Ludmile Celbové, vedoucí oddělení elektronických online zdrojů Národní knihovny ČR.

ÚVOD

Internet je jedním z nejmladších a nejrychleji rostoucích médií ve světě. Jeho růst je stále velice dynamický, hlavně díky službě World Wide Web. Web je informačním zdrojem číslo jedna pro miliony uživatelů po celém světě. Denně vzniká více než 7 miliónů nových stránek, které však ve stejnou dobu i nenávratně mizí. Průměrná životnost webové stránky se odhaduje na 3 až 4 měsíce [62]. jiná studie uvádí pouhých 44 dní. Také studie švédské národní knihovny z roku 2000 dokazuje, že pouze 20% dokumentů nalezených na webu zůstává po uplynutí jednoho roku nezměněna. Studie uvádí, že více než třetina stránek zmizí, jejich URL adresy již neexistují.

V důsledku absence jediné instituce zodpovědné za uchovávání webu v celém jeho rozsahu se tohoto úkolu zhostily organizace různého typu, jež se zaměřily na jednotlivé podskupiny webu. Dnes již existuje mnoho iniciativ, které se archivací webu zabývají. Mezi tyto organizace řadíme archivy, národní knihovny, dokonce i samy producenty. Asi nejznámější a nejambicióznější iniciativou současnosti je americký projekt Internet Archive. Tato nezisková organizace archivuje webové stránky již od roku 1996 a dnes má již obrovskou sbírku. Národní knihovny jsou nositeli dalších úspěšných iniciativ v této oblasti. První projekty vznikly ve Švédsku (Kulturarw[3]) a Austrálii (PANDORA), další země je brzy následovaly – Dánsko, Finsko, Francie, Norsko, Nový Zéland, Rakousko, Velká Británie. Jako první z bývalého východního bloku se v roce 2000 připojila i Česká republika se svým projektem WebArchiv, následovala Litva a Slovinsko.

I další organizace začaly experimentovat s archivací webu - národní archivy. Archivy byly vlastně donuceny se této problematice začít věnovat, jelikož mezi jejich hlavní funkce spadá archivace dokumentů veřejné správy. A právě v oblasti orgánů veřejné správy je cítit velká snaha přesunout značnou část své agendy do elektronické podoby. Některé národní archivy začaly webové dokumenty zařazovat do svých sbírek, např. americký národní archiv (US National Archives and Records Administration), jiné archivy – Národní archiv Austrálie (NationalArchives of Australia) nebo britský archiv (UK´s Public Record Office) spolupracují při archivaci webových stránek veřejné správy s místními národními knihovnami.

Dokonce i univerzity a vědecké organizace se v několika málo zemích zapojily do snahy archivovat obsah zveřejněný na webu a začaly budovat vlastní malé projekty. Příkladem je projekt holandské univerzity v Groningenu - Archipol[1], který sbírá holandské webové stránky zaměřené na politiku. Nebo DACHS[2] – projekt Institutu čínských studií na univerzitě v Heidelbergu, který se soustřeďuje na archivaci a zpřístupnění internetových zdrojů se vztahem ke studiu čínštiny.

V neposlední řadě to jsou samy vydavatelé webových stránek, kteří se snaží o dlouhodobou archivaci vlastních dokumentů, jmenovat můžeme například British Broadcasting Corporation (BBC).

Malé či soukromé archivy ale většinou nejsou vystaveny žádné kontrole. Nelze s jistotou tvrdit, že informace v nich obsažené jsou úplné a pravidelně aktualizované. Proto je třeba, aby se uchovávání národního dědictví v digitální podobě věnovaly zejména instituce, které jsou za to již tradičně zodpovědné.

V minulosti bylo mnoho důležitých součástí našeho kulturního dědictví ztraceno, jelikož nebylo archivováno – částečně proto, že minulé generace nerozpoznaly nebo nemohly rozpoznat jejich historickou hodnotu. Dnes už takovou situaci, v literatuře se setkáme s výrazem - *Digital Dark Age* (digitální doba temna), nemůžeme dopustit. Stále více organizací i jednotlivců si uvědomuje nutnost začít efektivně sbírat a dlouhodobě uchovávat kulturní dědictví v elektronické podobě. Za uchování kulturního dědictví tištěných dokumentů jsou všude na světě zodpovědné zejména národní knihovny. To samé se od nich očekává i v dnešní době, kdy se obsah přenesl z papíru do digitální formy. Knihovny však nikdy a snad nikde na světě nebyly institucemi oplývajícími přebytkem finančních prostředků, které jsou k naplnění tohoto ambiciózního cíle zapotřebí. Problémem zůstává i personální zabezpečení, jelikož pro uchování digitálních dat je třeba odborníků spíše z oblasti počítačové vědy než knihovníků. Knihovny se však snaží všechny tyto překážky překonat a nelze jim upřít, že již sklízejí i první úspěchy.

[1] http://www.archipol.nl

[2] http://www.sino.uni-heidelberg.de/dachs/

0.1 Vymezení pojmů

Odborná literatura nabízí širokou škálu pojmů vztahujících se k elektronickým zdrojům. Existuje mnoho synonymních výrazů, pojmů širšího či užšího rozsahu. Za zastřešující pojem této kategorie lze považovat termín:

Elektronické informační zdroje – zdroje určené pro práci s počítačem, včetně dokumentů, které vyžadují periferní jednotky. Za jeho synonyma lze považovat: digitální informační zdroje, elektronické (digitální) materiály, elektronické (digitální) dokumenty, elektronické (digitální) publikace.

Elektronické informační zdroje je možné rozdělit dle hlediska přístupu na:

- Offline – zdroje distribuované na fyzickém, přenosném nosiči (např. CD-ROM) a čitelné prostřednictvím počítače
- Online – zdroje uložené pomocí digitální technologie a šířené prostřednictvím digitální sítě

Tato práce je zaměřena na druhou skupinu zdrojů s online přístupem. Pojem jako „zdroj“, „materiál“, „dokument“ či „publikace“ mají v této práci naprosto stejný význam.

Také následující pojmy (s vědomím určitých rozdílů) jsou v rámci této práce používány jako synonyma: online zdroje, internetové zdroje, webové zdroje.

Další důležité pojmy:

Archivace webu (web archiving) – soubor procesů – výběr, akvizice, dlouhodobá archivace a zpřístupnění zdrojů publikovaných na webu.

Deep web (hluboký / neviditelný web) – oblast Internetu tvořená dokumenty většinou vysoké kvality, které jsou uloženy ve formě databází a jsou tak nedostupné pro vyhledávací nástroje.

Dlouhodobá archivace (preservation) – dílčí proces archivace s cílem minimalizovat důsledky rychlého morálního stárnutí digitálních technologií, a tak zajistit dlouhodobé uchování těchto dokumentů.

Flash – software, který umožňuje vytvoření webových stránek obsahujících zvuk, grafiku a animaci.

Harvester - speciální softwarové nástroje využívané ke sběru webových zdrojů, dokážou dokument z webu stáhnout a uložit tak, aby byl připraven na opětovné zpřístupnění.

Harvesting – metoda stahování dokumentů z webu za použití harvesteru.

JavaScript – programovací jazyk pro www stránky používaný pro vyšší míru interaktivity a dynamičnosti HTML dokumentů.

Metoda Konspektu - vznikla v USA v 70. letech, dnes je celosvětově rozšířena, slouží primárně k věcnému popisu knihovních fondů s cílem dosažení mezinárodní srozumitelnosti a hodnocení sbírek dle jednotného principu. Dnes je používána v rámci věcného popisu i v Národní knihovně ČR.

1 PROCES ARCHIVACE ELEKTRONICKÝCH ONLINE ZDROJŮ

V současné době existuje v celosvětovém měřítku velké množství výrazů k označování systémů, které částečně nebo zcela automaticky získávají, zpracovávají, ukládají či archivují, rozšiřují a zpřístupňují informace přes Internet [13]. Takovéto systémy jsou často nazývány digitálními knihovnami.

Pro označení systému archivace elektronických online zdrojů na národní úrovni za účelem uchování národního kulturního dědictví se používá termínu „webový“ nebo „digitální archiv“.

Z funkčního hlediska jsou pojmy digitální knihovna a digitální archiv mírně odlišné.

Digitální knihovna (Digital Library)

Z hlediska počítačové vědy - je to spravovaná sbírka informací spolu s odpovídajícími službami, přičemž informace jsou uloženy v digitální podobě a jsou dostupné prostřednictvím sítě [9].

Z knihovnického hlediska – jsou digitální knihovny organizace, které poskytují zdroje (včetně specializovaného personálu) umožňující provádět výběr, strukturování a zpřístupnění sbírek digitálních prací, tyto práce dále distribuovat, udržovat jejich integritu a dlouhodobě uchovávat. To vše s ohledem na snadné a ekonomické využití určitou komunitou nebo množinou komunit uživatelů [9].

Digitální archiv (Digital Archive)

Označení digitální archiv je výraz pro systém, jenž zajišťuje prioritně archivní funkci [13].

Digitální archivy x digitální knihovny

Digitální knihovny jsou pro digitální archivy nejdůležitějším příbuzným oborem. Řada přístupů oblasti digitálních knihoven byla převzata i tvůrci standardů

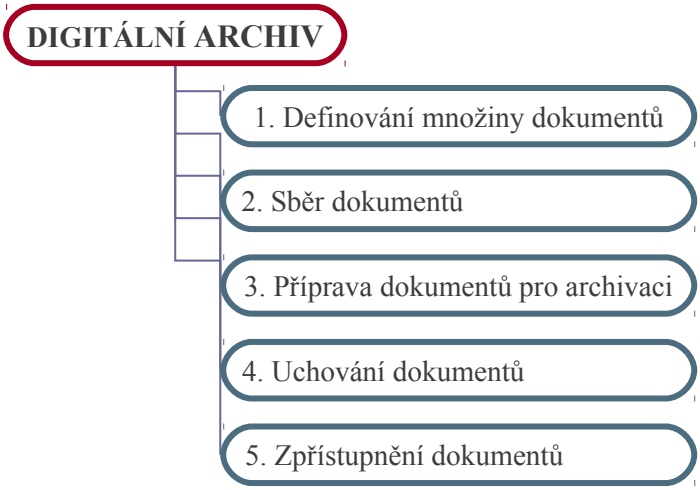
pro archivy. Na druhé straně je digitální archiv od digitální knihovny podstatně odlišný v tom, že zde nedochází k rovnoměrnému studiu všech uložených dokumentů a zejména v tom, že v archivu je třeba zajistit autenticitu dokumentu. Digitální archiv je definován jako organizace zajišťující ukládání a zabezpečování digitální informace. Jedná se především o zajištění dlouhodobého přístupu k národnímu, ekonomickému, kulturnímu a intelektuálnímu dědictví existujícímu v digitální podobě. Rozdíl mezi digitálním archivem a digitální knihovnou pak spočívá především v požadavku na dlouhodobý přístup k digitálním informacím. Většina jejich ostatních funkcí se překrývá [54].

1.1 Proces archivace

Proces archivace digitálních dokumentů představuje komplexní problém. Při jeho řešení je nutné se zabývat několika samostatnými otázkami technické, knihovnické a právní povahy, které se navzájem prolínají.

V oblasti knihovnické jde zejména o potřebu vytvořit metodiku pro výběr dokumentů a jejich zpracování s aplikací národních a mezinárodních standardů. Legislativní otázky se týkají povinného výtisku, resp. prozatímního řešení otázky oprávnění k získávání, archivaci a zpřístupňování elektronických zdrojů, a dále řešení autorskoprávní problematiky, tedy využívání archivovaných a zpřístupňovaných elektronických zdrojů. V oblasti informačních technologií bylo třeba vyvinout softwarové nástroje umožňující získávání, archivaci, zpracování a zpřístupňování elektronických zdrojů při dodržení legislativních podmínek [19].

Obr.1: Schéma procesů při tvorbě digitálního archivu



1.1.1 Definování množiny dokumentů

První pokusy o archivaci webových dokumentů se začaly objevovat zhruba v roce 1996. Vycházely z uvědomění si dynamické povahy digitálních zdrojů, znamenající v mnoha případech jejich nenávratnou ztrátu. Zvolené postupy archivace se však významně lišily.

Jedni chtěli archivovat vše, jelikož nedokázali odhadnout, co bude v budoucnosti pro uživatele zajímavé a podstatné. Druzí oblast svého zájmu vymezili národní doménou (oba tyto postupy odpovídají metodě plošné archivace). Jiní se snažili vytvořit skupinu kritérií, podle kterých vybírají jen ty nejvýznamnější zdroje (takováto metoda je nazývána výběrová metoda archivace).

Ambice archivovat absolutně vše záhy narazila na limity, s nimiž se potýká většina depozitních knihoven – zejména omezené rozpočty a z toho vyplývající omezené kapacity jak personální, tak technické. To vše vedlo k závěru, že je třeba si vymezit rozsah zájmové oblasti a v jeho rámci potom využít jedné z metod archivace webu. Jedním z řešení je spolupráce Národních knihoven s ostatními významnými národními institucemi, kdy by každá z nich zaměřila pozornost na svou prioritní oblast.

Vedle možností materiálně-technického a personálního zabezpečení či finančních možností realizace vědeckovýzkumného bádání v této oblasti, ovlivňuje metodu výběru zdrojů zejména existence zákona o povinném výtisku, který by se vztahoval i na dokumenty v elektronické podobě a zajistil tak knihovnám bezproblémové získávání všech publikovaných dokumentů.

Zákon o povinném výtisku

Povinný výtisk je definován jako: „zákonná povinnost, která vyžaduje od jakékoli komerční nebo veřejné organizace i jakéhokoli jednotlivce, vytvářejících jakýkoli typ dokumentů v různých kopiích, povinnost ukládat jednu nebo více kopií oprávněné depozitní instituci“ [84].

Ve většině zemí je institut povinného výtisku základním či výlučným zdrojem pro budování národní bibliografie. Díky národní legislativě tak mohou depozitní knihovny zajistit trvalý přístup k publikovanému národnímu kulturnímu dědictví. Zákony se vesměs vztahují na tradiční typy dokumentů (knihy, seriály, tištěné hudebniny, mapy), v některých zemích registrují i speciální typy dokumentů (zvukové a vizuální dokumenty).

Forma a nosič publikace se však v posledních desetiletích změnily, velká část informací, se kterými se dnes setkáváme, existuje pouze v digitální formě. Instituce zodpovědné za uchování národního kulturního dědictví a zároveň tradiční příjemci povinného výtisku si začali uvědomovat, že vzniká nevyplněná mezera v archivaci publikovaných dokumentů a nedochází tak k naplnění jejich depozitní funkce. Systém zákonů o povinném výtisku však na tuto situaci nebyl připraven.

Situace je poněkud komplikovaná zejména povahou internetových zdrojů. V mnoha případech je velice těžké vymezit jejich hranice, často nelze zjistit jejich vydavatele, informace v nich obsažené jsou těžko ověřitelné, jejich kvalita bývá různá. Hlavním problémem je ale forma online zdrojů, jsou přístupné po síti, jejich povaha je nehmotná a není tudíž možné je fyzicky odevzdat.

Legislativa povinného výtisku je ve většině zemí zastaralá a je nutné ji nahradit novým, moderním zákonem vztahujícím se i na online zdroje a zahrnující otázky autorských práv, veřejného přístupu, metod akvizice, sankcí apod.

Již na Mezinárodní konferenci o národních bibliografických službách v Kodani konané v roce 1998, byl zdůrazněn význam povinného výtisku a nutnost přehodnocení a aktualizace stávajících legislativ doplněním o odevzdání povinných výtisků elektronických zdrojů včetně síťových. Praktický dopad však toto prohlášení nepřineslo.

Do řešení problematiky aktualizace zákonů o povinném výtisku se vložila i mezinárodní organizace UNESCO. V roce 2000 byla publikována nová forma modelu právní úpravy povinného výtisku - *Guidelines for Legal Deposit Legislation* (Směrnice pro legislativu povinného výtisku), vytvořená Kanaďanem Julesem Larivièrem.

Některé země se pustily do přípravy novely vlastního zákona o povinném výtisku a některé byly úspěšné. Je to např. Kanada, Dánsko, Nový Zéland, Norsko, Island, Jižní Afrika či Velká Británie. Další země se o totéž pokoušejí s většími či menšími úspěchy. Projednávání novely zákona probíhá například v zemích jako Francie, Německo či Finsko. Třetí skupinu tvoří státy, mezi něž patří i Česká republika, kde se o nutnosti změny již dlouho hovoří, konkrétní kroky však ještě nebyly podniknuty.

Podrobné informace o stávajícím stavu legislativy povinného výtisku jsou obsahem diplomové práce Zuzany Volmuthové - *Povinný výtisk online publikací a jeho legislativní zajištění*, obhájené v roce 2003 na ÚISK, FF UK [84].

Jelikož je příprava novely zákona velmi zdlouhavým procesem, přistoupilo mnoho zemí na metodu dobrovolného odevzdávání povinného výtisku online publikací. Modelovým příkladem takovéto dohody mezi vydavateli a depozitní knihovnou byla *Mezinárodní deklarace k odevzdávání elektronických dokumentů do konzervačního fondu* připravená na základě rozsáhlé spolupráce CENL (Conference of European National Librarians) a Federace evropských vydavatelů (FEP - Federation of European Publishers). Deklarace, jejímž cílem bylo nabídnout jakousi alternativu legislativy povinného výtisku elektronických zdrojů, byla publikována v roce 2000. V rámci deklarace byla stanovena pravidla pro dobrovolné poskytování kopie elektronických online dokumentů do elektronického archivu.

Na základě tohoto dokumentu vytvořili systém dobrovolného odevzdávání online publikací např. ve Velké Británii, v České republice byla podle deklarace

vytvořena vzorová smlouva mezi vydavatelem a Národní knihovnou pro archivaci a zpřístupnění domácích online zdrojů.

1.1.2 Sběr dokumentů

Ať už si instituce vybere pro budování vlastního digitálního archivu jakoukoli z výše zmíněných metod (tj. plošná nebo výběrová), musí zajistit sběr relevantních dokumentů. Ke sběru se využívají speciální softwarové nástroje nazývané harvestery, které dokáží dokument z webu stáhnout a uložit tak, aby byl připraven na opětovné zpřístupnění . Instituce tedy sama aktivně stahuje a ukládá data. Tento přístup, který dnes využívá většina zemí a kterému je věnována tato práce, nese označení jako „pull model".

V několika málo zemích by vyzkoušen odlišný, dnes již téměř nevyužívaný přístup pro oblast online dokumentů, tzv. „push model". V tomto případě vydavatelé sami posílají své dokumenty depozitní instituci na fyzických nosičích buď klasickou poštou nebo prostřednictvím e-mailu, případně přenosem přes FTP přímo na archivační server instituce. Touto metodou lze sice získat i omezeně přístupné dokumenty, ovšem na druhé straně nechává příliš velkou zodpovědnost na samotném vydavateli a klade i vysoké nároky na jeho technické a technologické znalosti a vybavení.

1.1.3 Příprava dokumentů na archivaci

Digitální dokumenty, které jsou vybrány k dlouhodobému uložení do archivu, musí být na archivaci připraveny. Tento proces zahrnuje kontrolu kvality stažených dokumentů, popis a katalogizaci objektů včetně jejich původu a kontextu a jejich doplnění metadaty.

Metadata

Při zpracování elektronických dokumentů mají zásadní význam metadata - strukturovaná data o jiných, primárních datech, která je umožňují správně interpretovat. Obvykle popisují obsah, fyzický popis, lokaci, typ a formu informací,

údaje nezbytné pro další nakládání s daty jako zabezpečení, autenticitu, použité formáty či vztah k dalším entitám.

Metadata jsou také nezbytná pro zachování integrity elektronického zdroje. K tomu je třeba splnit důležitou podmínku - metadata musí být co nejméně svázána s konkrétním digitálním prostředím. Použití dobře propracovaného a standardizovaného metadatového schématu, který je již v širší míře používaný, je zárukou snadného využití a dalšího nakládání s digitálními informacemi. Rozeznáváme několik základních **kategorií metadat**:

- Popisná (*descriptive metadata*) – slouží k identifikaci objektu (např.MARC, Dublin Core)
- Administrativní (*administrative metadata*) – reprezentují informace jako datum vytvoření zdroje, informace o formátech, autorská práva apod. (např. elektronický podpis)
- Technická (*technical metadata*) - definují atributy, které popisují fyzické vlastnosti objektů (např. hlavičky protokolu HTTP)
- Archivační (*preservation metadata*) – administrativní + technická metadata

V procesu archivace hrají nejdůležitější roli právě archivační metadata, která zajistí dlouhodobý přístup k archivovaným datům. Jsou určena pro zaznamenávání technických detailů týkajících se formátu, struktury a použití digitálního obsahu, evidují historii všech procesů, kterými zdroj prošel [67].

Ucelený soubor prvků publikovala pracovní skupina pro archivační metadata OCLC/RLG ve své práci *Preservation Metadata for Digital Objects* [72]. V současné době intenzivně pokračují další výzkumy v této oblasti.

1.1.4 Uchování dokumentů

Informace v digitální podobě jsou nestálé a mají krátkou životnost. Na rozdíl od dokumentů v papírové podobě, které přežívají a zůstávají čitelné po staletí, digitální dokumenty nepřetrvají tak dlouho bez lidského zásahu a zůstanou přístupné jen po dobu několika málo měsíců, výjimečně roků. Příkladem mohou být 5 1/4 palcové diskety, na které sice může být informace stále uložena, neexistuje však již

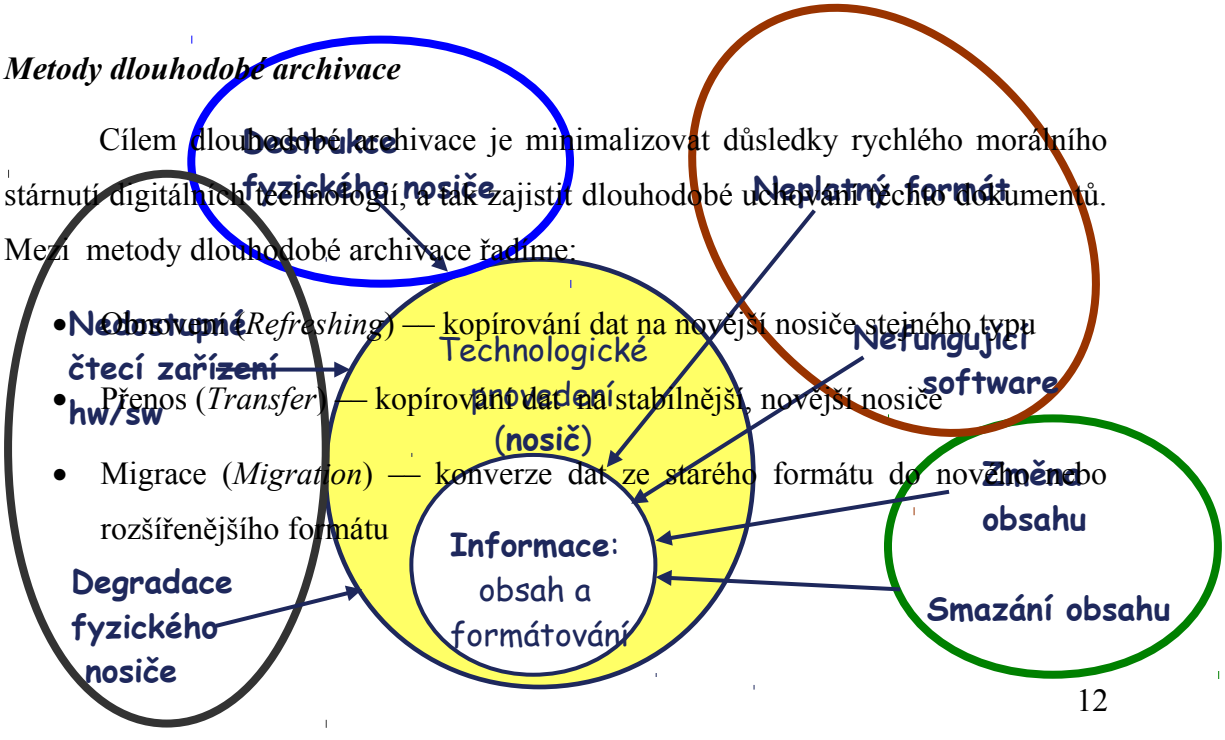
hardwarové ani softwarové vybavení, které by informace na ní uložené dokázalo zpřístupnit. Jiným případem je digitální informace uložená na optických discích CD nebo DVD. U těchto nosičů se doba použitelnosti odhaduje na desítky let, poté musí být obsah přemístěn na nové médium.

Mnoho současných projektů archivace webu se více soustředí na výběr a sběr zdrojů než na jejich dlouhodobé uchovávání a techniky s tímto spojenými. Takovýto postoj je možný v krátkodobé perspektivě, z dlouhodobého hlediska je však potřeba se zaměřit i na tuto problematiku. Mezi hlavní faktory ovlivňující životnost informace v digitální podobě řadíme:

- Degradaci média
- Ztrátu funkčnosti přístroje pro zpřístupnění
- Ztrátu možnosti manipulace
- Ztrátu možnosti prezentace
- Ztrátu kontextuální informace [42]

Miroslav Bartošek na konferenci Infos 2003 představil hlavní faktory ovlivňující životnost digitální informace ve výstižné grafice [9].

Obr. 2: Hrozby pro digitální informaci



Metody dlouhodobé archivace

Cílem dlouhodobé archivace je minimalizovat důsledky rychlého morálního stárnutí digitálních technologií, a tak zajistit dlouhodobé uchování těchto dokumentů. Mezi metody dlouhodobé archivace řadíme:

- Obnovení (*Refreshing*) — kopírování dat na novější nosiče stejného typu
- Přenos (*Transfer*) — kopírování dat na stabilnější, novější nosiče
- Migrace (*Migration*) — konverze dat ze starého formátu do nového nebo rozšířenějšího formátu

- Emulace (*Emulation*) — zachování digitálních objektů v původní formě použitím softwarů, které umí napodobit na nové platformě původní, zastaralý software nebo operační systém
- Enkapsulace (*Encapsulation*) — seskupení digitálního objektu a dalších součástí nezbytných ke zpřístupnění tohoto objektu jako např. metadata [42]

Filip Vojtášek ve svém příspěvku na Inforu 2001 zmínil ještě další 2 možné metody:

- Technologické muzeum (*technology museum*) – uložení digitálních dokumentů, aplikačního softwaru a operačního systému v originální formě a rovněž (na rozdíl od emulace) udržování platformy včetně příslušných čtecích zařízení ve funkčním stavu. Jde v podstatě o konzervaci digitálního prostředí. V praxi ale tato metoda přináší závažné technické a organizační problémy, proto se tato metoda pro dlouhodobou archivaci nepoužívá
- Převod do analogové podoby (*analogue form conversion*) – jedná se vlastně o verzi migrace, kdy se data z digitální podoby přenesou buď do tištěné formy nebo na mikrofilm, případně na niklový disk (tzv. optická nano-litografie). Pokud vezmeme v úvahu množství dat, které by mělo být takto převedeno, je jasné, že tato metoda rozhodně využívána nebude, nehledě na fakt, že např. multimediální soubory nelze do takovéto podoby přenést [83].

Pro účely dlouhodobé archivace na národní úrovni přicházejí v úvahu dvě ze zmíněných strategií – migrace a emulace. Tyto metody jsou založeny na naprosto protichůdných principech.

Metoda migrace je v současnosti považována za hlavní strategii dlouhodobé archivace digitálních dokumentů. Migrace je proces opakující se konverze dat ze staršího digitálního prostředí do nového (pozornost je soustředěna na obsah), zpravidla z jednoho formátu do druhého. Tuto činnost provádějí podniky, úřady veřejné správy a další instituce na různé úrovni, které manipulují s datovými soubory, v rámci svého běžného provozu. Knihovny s migrací týkající se digitálních dokumentů nemají zkušenosti, ovšem migrací ve své podstatě je rovněž převod

klasických fondů na nová média (především digitalizace), jehož cílem je usnadnit přístup k uživatelsky atraktivním dokumentům, které jsou z různých příčin ohroženy, formou jejich kopie a současně přispět k jejich uchování [83].

Nespornou výhodou migrace jsou praktické zkušenosti, jde o osvědčenou metodu (ve srovnání s emulací), nevýhodou je riziko postupné ztráty integrity digitálního dokumentu jako celku či jeho jednotlivých objektů.

Zvláště problematický je stav v oblasti formátů digitálních dokumentů. V současnosti existuje celá řada formátů, mnoho z nich je však vzájemně nekompatibilních. Kromě jiného je totiž nekompatibilita nástrojem konkurenčního soupeření producentů aplikačního softwaru. Vedle poměrně nepočetné skupiny formátů, které lze považovat za obecné a široce podporované (např. RTF, TXT, JPG, GIF, TIFF, MP3, HTML, XML), se používá množství dalších proprietárních formátů, k jejichž interpretaci potřebujeme konkrétní software. V opačném případě se vystavujeme riziku, že dojde k narušení integrity daného digitálního dokumentu. Do jisté míry se můžeme spolehnout na zpětnou kompatibilitu u aplikačního softwaru od téhož producenta a zejména hardwarových zařízení, danou zejména respektováním mezinárodní standardizace [83].

Velmi limitovanou životnost mají digitální dokumenty, k jejichž spuštění je třeba speciální aplikační software. Obdobně riskantní je spoléhat se při archivaci na proprietární formáty, které jsou vyvíjeny a podporovány menšími producenty, jakkoliv se jeví ve srovnání se zavedenými formáty jako momentálně výhodnější.

Metodou emulace se rozumí proces, jehož smyslem je co možná nejvěrněji modelovat funkční vlastnosti morálně zastaralého digitálního prostředí či jeho komponentů na jiném počítači, než pro který bylo určeno. Jinými slovy jde o konzervaci dokumentu spolu s příslušným aplikačním softwarem a operačním systémem v původní podobě a specifikace hardwarové platformy pro jejich budoucí oživení. Smyslem je tak uchování nejen obsahu, ale i funkčních vlastností. K tomu je třeba přiřadit množinu technických metadat, která specifikují příslušnou hardwarovou platformu s cílem zajistit, aby kdykoliv v budoucnu mohl být vyvinut program-emulátor, který v rámci pozdějšího digitálního prostředí, jehož architektura je v současnosti neznámá, umožní (jako další vrstva v tomto prostředí) “oživit” digitální

dokument podle potřeby v jeho, tj. v té době již virtuálním prostředí, aniž by byl vystaven riziku narušení integrity jako důsledku opakované migrace.

I když i kritici připouštějí teoretickou opodstatněnost této strategie a výsledky jejího testování se ukazují jako slibné, její nedostatek spočívá zejména v tom, že nebyla v reálném provozu ověřena [83].

Migrace x emulace

V řadě materiálů jsou diskutovány charakteristiky obou metod a jejich potencionální vhodnost pro účely dlouhodobé archivace. Na základě výpočtů cenových nákladů bylo dokázáno, že využití metody migrace by bylo výhodné pro dlouhodobou archivaci malých sbírek. Migrace má nižší počáteční náklady, ale musí být prováděna opakovaně (průměrně po 3-5 letech). Emulace je oproti tomu výhodná pro velké sbírky (cca od 500 000 objektů). Čím více objektů pro emulaci, tím levnější náklady. Požadavky na počáteční investice jsou sice vyšší, ale emulace nemusí být prováděna opakovaně v intervalech. Výhodou emulace je fakt, že její využití by mohlo být celosvětové, zatímco migrace je vhodná pouze pro lokální použití.

I přes elegantnost metody emulace však zřejmě bude v nejbližším období schůdnější metoda migrace. Za hlavní důvod lze považovat skutečnost, že migrace se (i když v relativně omezeném měřítku) již řadu let používá na řadě pracovišť v rutinním provozu a je tak dobře propracována metodika jejího používání (periodicita kontrol čitelnosti, organizace záložních kopií atd.). Z toho důvodu bude zřejmě vhodné řešení problému archivace digitálních dokumentů založit na metodě migrace s tím, že se nedá vyloučit aplikace metody emulace někdy ve vzdálenější budoucnosti [54].

1.1.5 Zpřístupnění dokumentů

Aby digitální archiv splnil svůj účel, je třeba archivované dokumenty zpřístupnit veřejnosti. Problémem souvisejícím s jejich zpřístupněním je odpovídající legislativa. Jedná se o vyhovující autorský zákon, či zákon o povinném výtisku, který by zajistil nejen možnost stahování a archivace elektronických online zdrojů, ale také jejich zpřístupnění. Stav těchto legislativních norem je však v mnoha zemích pro tyto účely zcela nevyhovující.

Jedním z východisek je pak vyjednávání s vydavateli online zdrojů. V rámci tohoto zdlouhavého procesu je třeba nejprve zjistit identitu vydavatele, objasnit mu účel archivace a obvykle v rámci diskuse vyjasnit sporné otázky. Výsledkem je uzavření smlouvy či potvrzení souhlasu s archivací a zpřístupněním zdroje z digitálního archivu. Smlouva je většinou formulována tak, aby bylo možné zpřístupnit zdroj online přes webové rozhraní příslušného projektu.

Další možností je zdroj z archivu zpřístupnit omezeně – pouze z terminálů umístěných v budově dané instituce. Tento způsob využívá pro svůj digitální archiv Švédská národní knihovna (díky existenci speciálního vládního nařízení).

Evropský parlament a Rada vydaly v roce 2001 směrnici týkající se harmonizace některých aspektů autorského práva[3], která umožňuje knihovnám zpřístupnit online zdroje uložené v jejich sbírkách na vymezených terminálech v rámci budovy knihovny. Některé evropské země již tuto směrnici zapracovaly do svého autorského zákona a digitální zdroje z archivu tímto omezeným způsobem již zpřístupňují.

Mnoho zemí, které budují digitální archivy, však stále nemá možnost archivované zdroje jakýmkoli způsobem zveřejnit.

Pro zpřístupnění archivovaných zdrojů je třeba speciálních softwarových nástrojů, které simulují běžný webový prohlížeč. Severské země vyvíjejí ve spolupráci s mezinárodním konsorciem IIPC (*viz 3.1.1*) nástroj na prohledávání webového archivu - NWA Toolset. Ten se v současné době jeví jako nejvhodnější volně dostupný nástroj pro fulltextové vyhledávání.

[3] Směrnice o harmonizaci některých aspektů autorského práva a práv s ním souvisejících v informační společnosti (2001/29/ES)

2 METODY ARCHIVACE WEBU NA PŘÍKLADECH NEJVÝZNAMNĚJŠÍCH PROJEKTŮ SOUČASNOSTI

Povinností každé země by mělo být uchování národního kulturního dědictví a to včetně jeho digitální podoby. Od 90. let minulého století se touto činností začalo zabývat několik málo organizací, které nastínily první přístupy k archivaci webových zdrojů. Mezi organizace, jež vybudovaly první projekty, řadíme národní knihovny a archivy, vědecké spolky a univerzity. Za nejambicióznějšími iniciativami stojí národní knihovny, které pocítily svou odpovědnost za uchování národního kulturního dědictví, které již nemá pouze hmotnou podobu, mnoho děl již najdeme pouze v elektronické podobě – na Internetu.

Zpočátku se iniciativy chopilo několik málo knihoven z vyspělých států, zejména z USA, Kanady, Austrálie nebo Švédska. Během posledních let se jejich počet rapidně zvýšil a neustále roste. Již od začátku se kromě jiného diskutovalo i o tom, jakou zvolit nejlepší strategii pro archivaci webových zdrojů. Část zemí a jejich institucí volila cestou výběrové archivace (zejména Austrálie a Kanada), kdy z celého spektra webových zdrojů vybírají jen ty nejkvalitnější a nejdůležitější. Vytváření takového archivu je ale velmi náročné na čas a intelektuální práci. Proto se mnoho zemí vydalo cestou plošné archivace, která probíhá pomocí automatizovaného sběru (tzv. harvestingu) v předem specifikovaném prostoru (např. národní doméně) a ve stanovených intervalech. Ani v současnosti nelze hovořit o zřejmé převaze jednoho či druhého přístupu, mnoho zemí se přiklání ke kombinaci těchto dvou základních strategií.

Při volbě strategie hrají roli zejména finanční prostředky knihoven, jejich personální zabezpečení, tj. počet pracovníků a jejich profesní zaměření, stav národní legislativy (zejména povinný výtisk a autorské právo) nebo možnosti spolupráce s dalšími organizacemi zabývajícími se archivací webu na národní či mezinárodní úrovni.

2.1 Mezinárodní iniciativy

2.1.1 The International Internet Preservation Consortium (IIPC)

V rámci mezinárodní spolupráce dnes hraje zásadní roli Mezinárodní konsorcium pro archivaci Internetu (IIPC), které bylo založeno 24. července 2003 v Paříži. V současné době má 12 členů:

- Národní knihovna Austrálie (National Library of Australia)
- Národní knihovna Kanady (Library and Archives od Canada)
- Královská knihovna Dánska (Det Kongelige Bibliotek)
- Helsinská univerzitní knihovna - Národní knihovna Finska (Helsingin yliopiston kirjasto – Suomen Kansalliskirjasto)
- Národní knihovna Francie (Bibliothèque nationale de France)
- Národní a univerzitní knihovna Islandu (Landsbokasafn Islands – Haskolabokasafn)
- Národní knihovna Itálie (Biblioteca Nazionale Centrale di Firenze)
- Národní knihovna Norska (Nasjonalbiblioteket)
- Královská knihovna – Národní knihovna Švédska (Kungliga biblioteket Sveriges nationalbibliotek)
- Britská národní knihovna (The British Library)
- Knihovna kongresu (Library of Congress, USA)
- Projekt Internet Archive (USA)

Jeho cílem je umožnit sběr a archivaci bohatého obsahu internetu jako celosvětového souboru informací a zajistit k němu trvalý přístup, dále napomáhat vývoji a využití společných nástrojů, technologií a standardů, které umožní tvorbu mezinárodního archivu a v neposlední řadě podporovat národní knihovny v jejich úsilí zaměřeném na archivaci a ochranu internetových zdrojů.

Základní smlouva, která ustanovuje 12 členů konsorcia, je platná po tři roky, tj. do července 2006. Členové IIPC se dohodli, že budou společně financovat a

podílet se na projektech a pracovních skupinách, aby dosáhli cílů, které si stanovili. V průběhu roku 2006 by mělo být rozhodnuto, zda a kdy konsorcium přijme nové členy, kteří o spolupráci projevili enormní zájem.

Konsorcium již publikovalo na svých stránkách[4] první výsledky své práce. Je hlavním iniciátorem při vývoji softwarového nástroje typu harvester s názvem „Heritrix“, dále se snaží vyvinout nástroj pro tzv.“smart crawling“ (inteligentní sběr), což by měl být nástroj pro automatické sklízení zdrojů dle předem stanovených kritérií. Podílí se také na vývoji fulltextového nástroje pro vyhledávání zdrojů z archivu, či na nástroji umožňujícím získávat data z oblasti Deep webu. Celkově je možné aktivity konsorcia hodnotit jako velmi úspěšné.

2.1.2 Nordic Web Archive (NWA)

Mezi významné mezinárodní iniciativy řadíme i projekt Severského webového archivu[5]. Jedná se o fórum národních knihoven Dánska, Finska, Islandu, Norska a Švédska pro koordinaci a výměnu zkušeností v oblasti harvestingu a archivace webových zdrojů. Fórum vzniklo jako reakce na zjištění, že zásadním problémem při sklízení a archivaci webových dokumentů je zabezpečení softwaru a nástrojů, které by umožnily řešitelům, vědcům i všeobecné veřejnosti přístup do archivu a jeho využívání. Zároveň by umožnily řešitelům ověřit, že harvester sklidil požadovaný objem dokumentů z Internetu [64].

Proto se spolupráce členů NWA od listopadu 2000 soustředí na vývoj souboru softwarových nástrojů pro zpřístupnění archivovaných dokumentů souhrnně nazvaných „NWA Toolset“. Projekt je financován severskými národními knihovnami a dvěma granty. S využitím fulltextového vyhledávacího systému NWA Toolset se počítá v rámci konsorcia IIPC i v dalších zemích včetně České republiky.

[4] http://www.netpreserve.org

[5] http://nwa.nb.no/

2.2 Metody archivace webových zdrojů

2.2.1 Plošná metoda archivace

Pomocí plošné metody archivace je vytvářen tzv. kompletní archiv. Pro jeho tvorbu je charakteristický automatizovaný sběr tzv. harvesting (*viz. 3.3.1.1*), v předem specifikovaném prostoru - např. národní doméně (*viz. 3.3.1.2*) a ve stanovených intervalech (*viz. 3.3.1.3*).

Tento archiv obsahuje relativně všechny veřejně přístupné online zdroje z vymezeného prostoru. Vlastní sběr nenárokuje intelektuální práci, ale o to nižší je kvalita zařazených zdrojů. Jelikož většinou neprobíhá žádné vyjednávání mezi archivující organizací a vydavateli archivovaných zdrojů, nese takto tvořený archiv větší právní rizika.

Tento způsob je uplatňován např. ve Švédsku (projekt Kulturarw[3]), Rakousku (projekt AOLA) nebo Finsku (projekt EVA). Výjimečné postavení v této kategorii zastává americký projekt Internet Archive.

Internet Archive[6]

Internet Archive začal se sbíráním a archivací webu jako jeden z prvních již v roce 1996. Archiv sám byl založen jako nezisková organizace, ale vlastní sběr webových stránek byl a stále je prováděn komerční organizací s názvem Alexa Internet[7]. Tato organizace také stojí za rozvojem mnoha významných softwarových nástrojů.

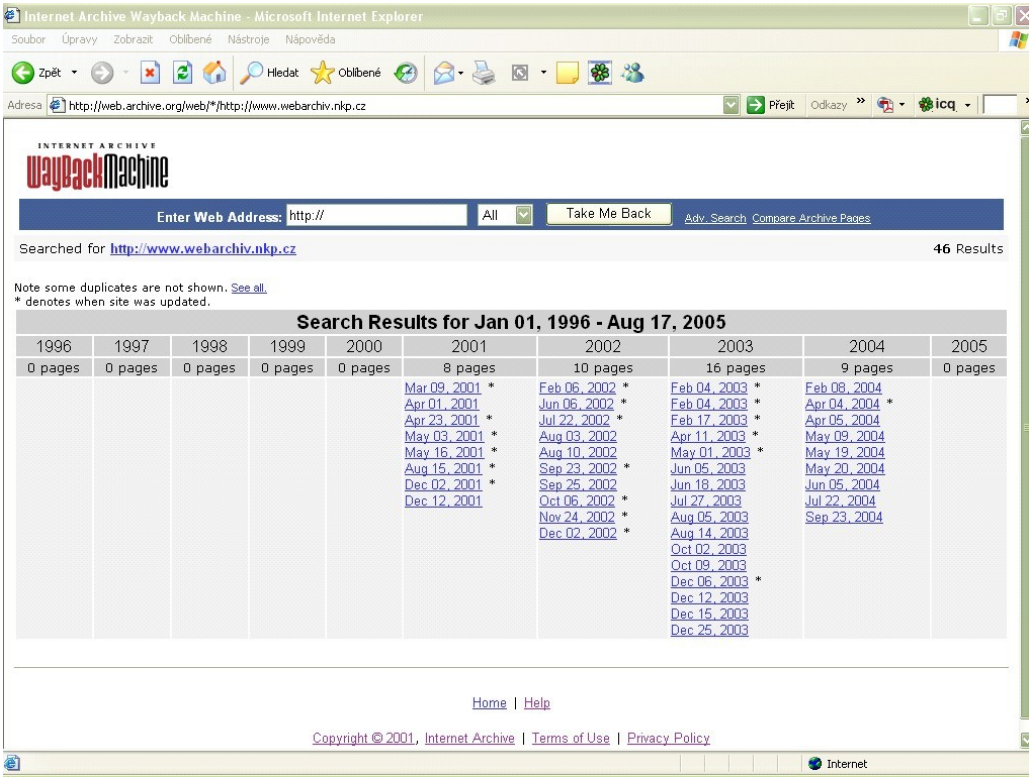
Velikost Internet Archivu je obrovská, sklizeň celosvětového webu probíhá každé 2 měsíce a sbírka dnes (údaj k 17.8.2005) obsahuje kolem 40 miliard stránek včetně jejich časových verzí, což v současnosti představuje zhruba 1 petabyte dat s měsíčním nárůstem kolem 20 terabytů. Pro lepší představu bychom tento objem dat mohli srovnat například s největší knihovnou na světě – Library of Congress. Bylo odhadnuto, že její tištěné fondy čítají kolem 20 terabytů textu [31].

[6] http://www.archive.org

[7] http://www.alexa.com

V roce 2001 zpřístupnil Internet Archive svou sbírku online pomocí vyhledávače WayBack Machine, který po zadání URL vyhledá požadovanou stránku s jejími časovými verzemi.

Obr. 3:Ukázka vyhledaných archivovaných verzí webových stránek českého projektu WebArchiv



Internet Archive také vytvořil speciální sbírky na specifická témata. Stojí například za malou kolekcí nazvanou „Průkopníci webu“, která shromažďuje klíčové stránky z doby počátků webu. Jiné sbírky byly vytvářeny společně s dalšími organizacemi. Například společně s Library of Congress se podílel na sbírce stránek týkajících se amerických voleb v roce 2000. V nedávné době pomáhal Internet Archive vytvářet „Web Archiv 11. září 2001“ ve spolupráci s Library of Congress, WebArchivist.org a Pew Internet & American Life Project. Internet Archive je aktivní i na poli mezinárodní spolupráce, je zakládajícím členem mezinárodního konsorcia IIPC.

Kulturarw[3] [8]

Švédský projekt Kulturarw[3] je průkopníkem metody plošné archivace na národní úrovni a jedním z nejúspěšnějších projektů tohoto typu v současnosti. Byl spuštěn již v roce 1996 a jeho iniciátorem byla švédská Královská knihovna ve Stockholmu, která plní funkce národní knihovny. Podobně jako v případě Internet Archive byla zvolena metoda plošné archivace. Hlavním důvodem pro volbu této metody byla absence zákona o povinném výtisku, který by se vztahoval i na online publikace.

Projekt Kulturarw[3] se postupně stal velmi významným a ostatními zeměmi je ceněn a považován za vzorový. Problémem však zůstávala otázka zpřístupnění archivovaných zdrojů. Průlom se řešitelům projektu podařil až v roce 2002, kdy se po jejich stížnosti zaštítěné Královskou knihovnou začala problémem zabývat švédská vláda. V květnu 2002 nabylo účinnosti vládní nařízení[9], které opravňuje Královskou knihovnu ke stahování veškerých švédských internetových zdrojů, k jejich archivaci a zpřístupnění, ovšem pouze z vymezených terminálů v budově knihovny. Archivované dokumenty nejsou indexovány, proto uživatelé, aby požadovaný zdroj mohli v archivu vyhledat, musí znát přesné URL. I přes toto omezení je Kulturarw[3] zatím nejlépe fungujícím projektem na evropském kontinentu [7].

Softwarové nástroje pro automatizovaný sběr dat

Pro sběr webových stránek při plošné metodě archivace je využíván softwarový nástroj označovaný jako „harvester“, do češtiny přeloženo jako robot pro sklízení, případně kombajn (z anglického slova „to harvest“ - sklízet).

Web harvester je aplikace, která stahuje a ukládá webové stránky dle parametrů definovaných uživatelem. Jeho funkce je poměrně jednoduchá - na základě předem daného seznamu URL adres stáhne první dávku dokumentů. Ty jsou podrobeny analýze s cílem najít v nich obsažené hypertextové odkazy na další webové stránky. Ty jsou následně také staženy a celý proces je opakován tak dlouho, dokud nejsou staženy a analyzovány všechny stránky z předem definované oblasti.

[8] Cultural Heritage Cubed - http://www.kb.se/kw3/ENG/

[9] Decree (2002:287)

Vzhledem k vysoce propojené struktuře webových dokumentů je harvester schopen podchytit značnou část celého Internetu [90].

Vyloučení duplicit je zajištěno porovnáním kontrolního součtu MD5 (hashovací algoritmus) archivovaného a nově staženého dokumentu. Algoritmus MD5 složený z 128 bitového čísla slouží jako jednoznačný identifikátor dokumentu.

NEDLIB Harvester

První harvestery byly vyvinuty v polovině 90. let minulého století pro tvorbu rešeršních systémů webu jako např. Alta Vista. Tyto běžné harvestery obvykle stažené dokumenty nearchivují a po zaindexování je zahazují. Myšlenka použít specializovaný harvester k permanentní archivaci stažených dokumentů vznikla v rámci projektu Kulturarw[3]. Právě na výsledky tohoto projektu navázal o rok později projekt evropské komise – NEDLIB (Networked European Deposit Library, 1998-2000) a projekt severských zemí NWA (Nordic Web Archive). V obou naposledy zmíněných projektech hrála významnou roli Helsinská univerzitní knihovna a finské Centrum pro počítačovou vědu (Center for Scientific Computing), které vyvinulo na základě specifikací čtrnácti partnerských organizací program NEDLIB Harvester [39].

NEDLIB Harvester se skládá z mnoha vzájemně propojených modulů a několika pomocných programů, jádrem celého systému je databázový systém MySQL. Původně byl napsán v jazyce Perl, ale na zakázku Helsinské univerzitní knihovny byl přepsán do jazyka C, čímž došlo k jeho výraznému zrychlení a zároveň byla zlepšena i jeho funkčnost [19].

Software má kontrolu pro vyloučení duplicit, nástroj pro kompresi dat a stahuje pouze dokumenty, které byly změněny nebo nově vytvořeny. Dokumenty jsou stahovány do pracovního adresáře, tam jsou analyzovány, zpracovány a poté archivovány ve formátu TAR.GZ.

První testovací sklizně byly provedeny v prostoru islandského webu v roce 2001, o rok později ve finské doméně .fi. Testy prokázaly mnoho problémů – např. slabě vyvinuté HTTP server aplikace, specifické problémy s některými CGI skripty, což způsobovalo nekonečné cykly stahování a další potíže [39].

Kvůli poměrně velkým nedostatkům probíhaly úpravy až do roku 2002, kdy byla spuštěna poměrně uspokojující verze 1.2.2. NEDLIB Harvester je freeware stále ještě dostupný na http://www.csc.fi/sovellus/nedlib/index.phtml.en. Finální verze aplikace je poměrně robustní, schopná sklidit desítky miliónů webových dokumentů. Vývoj a podpora NEDLIB Harvesteru byla ukončena a v roce 2004 většina zemí, které ho využívaly, přešla na nový typ harvesteru s názvem Heritrix.

HERITRIX

Moderní a dobře propracovaný nástroj pro procházení a archivaci webu nesoucí název Heritrix[10] vznikl na půdě Internet Archive. Jelikož posláním Heritrixu je sběr a archivace digitálních artefaktů naší kultury pro užitek budoucích generací, zdá se jeho název být vypovídající.

Internet Archive vytvořil v jazyku Java napsaný open-source software, aby se na jeho vývoji mohly podílet i další instituce, jež se angažují v oblasti archivace webu. Práce na harvesteru započaly začátkem roku 2003, kdy byl vytvořen prototyp a jeho výkon byl testován a porovnáván s již existujícími harvestery. Od října 2003 do dubna 2004 byli do projektu přizváni i programátoři z NWA, zejména za účelem vytvoření uživatelského rozhraní. V lednu 2004 byla publikována první beta verze (0.2.0) a v dubnu 2004 byla spuštěna verze 1.0.0. [61].

Heritrix je možné použít pro několik různých typů sklizní:

- pro širokou sklizeň – kdy jde o sběr co největšího počtu dokumentů s cílem co největší kompletnosti
- pro úzce zaměřenou sklizeň – kdy jde o sběr malý až středně velký (obvykle méně než 10 miliónů unikátních dokumentů), sbíráme stránky dle kvalitativních kritérií do hloubky
- pro pokračující sklizeň – kdy jsou sbírány stránky opakovaně, pokaždé když byly změněny

[10] Anglické slovo *Heritrix* je archaickým výrazem pro dnešní výraz *heiress* – dědička

- pro experimentální sklizeň – kdy sběr probíhá neobvyklými technikami – např. je změněn postup sklízení, sklízení probíhá pomocí jiných protokolů, analyzují se a archivují výsledky sklizně.

Sklizeň probíhá v rámci několika procesů, které jsou společné většině harvesterů. Stažené dokumenty jsou ukládány ve formátu ARC.

Harvester během jednoho cyklu provede následující operace:

1. vybere URI z předem zadaných
2. sklidí tuto URI
3. analyzuje nebo archivuje výsledky
4. vybere nově objevené URI z oblasti zájmu a přidá je do seznamu mezi zadané
5. potvrdí, že URI byla sklizena a opakuje proces [61]

Systém Heritrix[11] je nyní dostupný ve verzi 1.4.0. Přestože jeho vývoj ještě zdaleka neskončil, nabízí tento nástroj již nyní sofistikované webové ovládání včetně propracovaného systému monitorování probíhající sklizně a dynamických úprav jejích parametrů.

Předem specifikovaný prostor

Cílem projektů, jež se soustřeďují na archivaci webu na národní úrovni, je snaha pokrýt plošným sběrem co nejširší oblast národního prostoru. Vymezení takového prostoru je však velice náročné. Jako základní parametr slouží národní doména. Pro označení národních domén se používají dvoumístné kódy podle ISO 3166. Pro Českou republiku je tak používáno označení .cz, pro Francii .fr apod.

Vydavatelé webových zdrojů však často využívají i jiné domény než národní. Často je k tomu vedou finanční důvody. Oblíbené domény jako .org, .com nebo .net

[11] Veškeré informace o Heritrixu, včetně možnosti jeho stažení, lze najít na jeho stránkách – http://crawler.archive.org.

jsou totiž často levnější variantou než doména národní. Dalším důvodem může být skutečnost, že jimi požadovaná doména je již registrována jiným subjektem.

Jako jedni z prvních se o výrazném využívání jiné domény než národní přesvědčili řešitelé švédského projektu Kulturarw[3]. Podle jejich statistik se ve švédské doméně .se nachází pouze asi 60% archivovaných zdrojů.

Zaměření se jen na prostor vymezený národní doménou tudíž rozhodně nedostačuje a je potřeba oblast národního webu vymezit pomocí dalších principů. Většinou se vychází z tradičního vymezení používaného v rámci národních bibliografií.

- Teritoriální princip (*publikace vydané na území státu, národa*) - je již splněn vymezením národní domény.
- Jazykový princip (*publikace vydané v národním jazyce v zahraničí*) – pro vymezení takovýchto publikací se využívají softwarové nástroje pro automatické rozpoznání jazyka. Velice obtížná situace pro anglosaské země, francouzsky a španělsky mluvící státy.
- Autorský princip (*publikace vytvořené národními autory v zahraničí*) – automatickým způsobem nezjistitelné, nutný intelektuální výběr.
- Obsahový princip (*publikace vydané v zahraničí, jejichž obsah má vztah ke státu, národu apod*.) – automatickým způsobem nezjistitelné, nutný intelektuální výběr.

Interval sběru

Z důvodu značně rozsáhlé oblasti sběru dokumentů, kterou představuje národní doména, může stahování dat trvat poměrně dlouhou dobu, někdy i několik měsíců. Vzhledem k časté aktualizaci určitých typů internetových zdrojů (např. zpravodajské servery, elektronické časopisy) by mohlo docházet k tomu, že některá čísla či vydání nebudou archivována. Proto je třeba vytipovat skupinu zdrojů s častější aktualizací a nastavit harvester na paralelní sběr dokumentů, kdy jedna akce je věnována sklízení celé domény a druhá sklízí pravidelně jen vybranou skupinu zdrojů.

Zhodnocení metody plošného sběru

Plošná metoda archivace je velice efektivní způsob, jak se pokusit archivovat oblast národního webu. Velkou předností jsou relativně nízké finanční náklady na tvorbu takového archivu a poměrně nízké nároky na lidskou práci. Většina z dnes využívaných softwarových nástrojů je volně dostupná a ceny paměťových médií neustále klesají.

Aby však tyto nástroje mohly být účelně využívány, musí být přizpůsobeny národnímu prostředí. Proto je nutné mít k dispozici tým technicky zdatných pracovníků, kteří budou schopni proces sklízení a archivace spustit, řídit a průběžně kontrolovat, případně opravovat chyby. Většina národních knihoven tyto problémy řeší ve spolupráci s externími odborníky z univerzitních či vědeckých pracovišť z oblasti výpočetní techniky.

Aby byl archiv využitelný, vzhledem k obrovskému množství dat staženému plošnou metodou archivace, je nezbytné mít kvalitní vyhledávač, který by umožnil efektivní vyhledávání v celém archivu. Ideálním řešením je vyhledávací systém, jehož funkce jsou podobné běžnému prohlížeči.

Tvorbou kompletního archivu je třeba sledovat i v kontextu autorského práva. Pokud země, které provádí harvesting a zpřístupňují archivované dokumenty veřejnosti, nemají odpovídající legislativní zajištění těchto procesů, vystavují se poměrně vysokému nebezpečí zejména ze strany vydavatelů.

Velkou nevýhodou plošné metody archivace je, že pomocí automatického sklízení není zatím možné sklidit dokumenty z oblasti Deep webu. Zejména z důvodů této, ale i dalších negativních stránek, přistoupilo několik zemí k opačnému řešení.

2.2.2 Výběrová metoda archivace

Na rozdíl od plošné archivace, kde je cílem pokrytí co nejširšího spektra zdrojů, se důraz při tvorbě výběrového archivu klade na kvalitu zdrojů, na jejich obsahovou, informační hodnotu. Vytváření takového archivu je náročnější na čas a intelektuální práci. Je založeno na individuálním jednání s vlastníky práv nebo vydavateli, jsou zde předem stanovena obsahová a formální kritéria, dle kterých jsou zdroje vybírány. Takto vytvořený archiv zahrnuje sice jen malé procento z toho, co je

na Internetu přístupné, ale nese menší právní rizika, protože má na základě jednání s vydavateli jasně definovaná přístupová práva.

Touto metodou tvoří národní archivy elektronických online zdrojů např. v Kanadě (projekt E-Collection) či v Japonsku (projekt WARP). Za nejvýznamnější iniciativu archivace webu výběrovou metodou je však právem považován australský projekt PANDORA.

PANDORA[12]

Tento projekt byl vůbec první iniciativou archivace webu na národní úrovni. Práce na projektu započaly na půdě Národní knihovny Austrálie (National Library of Australia, dále jen jako NLA) počátkem roku 1996. NLA se vydala cestou výběrové archivace, což znamená, že si vytyčila prioritní kategorie dokumentů, mezi které patří např. akademické publikace, vládní dokumenty, elektronické časopisy, materiály z konferencí, zdroje vztahující se k Austrálii aj. PANDORA se soustřeďují na takové dokumenty, které existují pouze v elektronické podobě, tzn. nemají tištěnou verzi.

NLA má poměrně detailně propracovaný postup spolupráce s ostatními institucemi, které na projektu PANDORA participují. Je to 9 dalších organizací - australských knihoven a organizací zabývajících se uchováním kulturního dědictví. Všechny spolupracující organizace mají podobná kritéria pro výběr zdrojů, jednotný postup při oslovování vybraných vydavatelů.

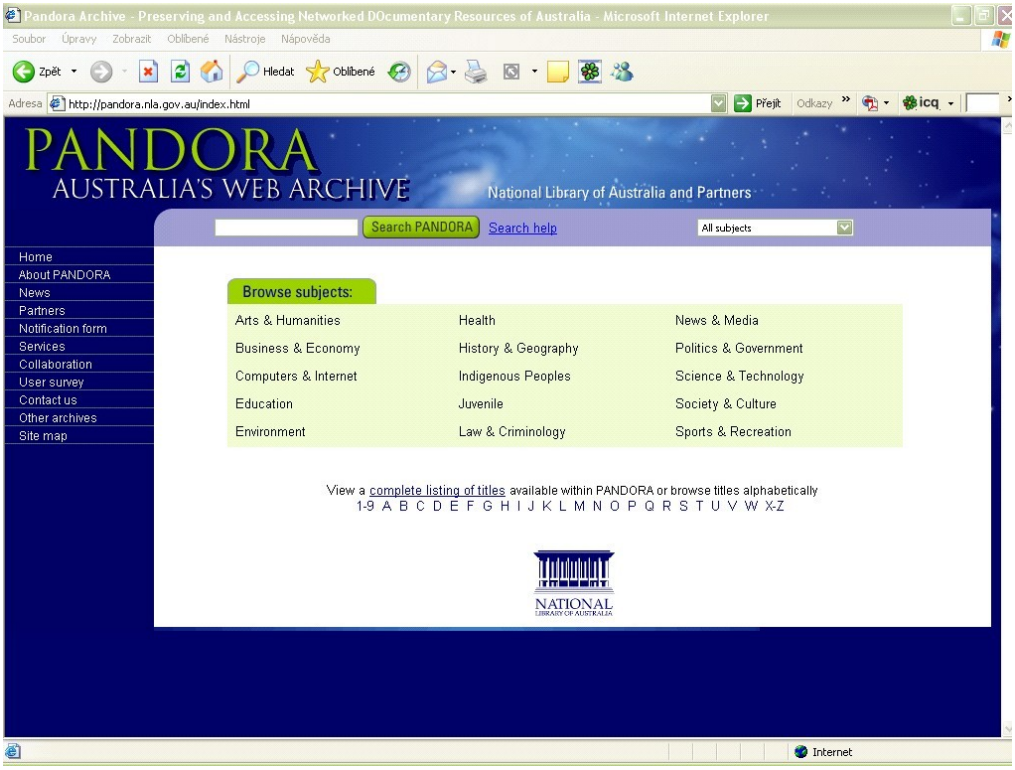
Pro správu celého archivačního procesu slouží systém PANDAS (PANdora Digital Archiving System). Tento systém kontroluje celý proces od stažení stránky přes záznam všech údajů (administrativní data), přiděluje dokumentům jednoznačný identifikátor (PURL), připravuje dokumenty pro zpřístupnění, poskytuje zprávy o provozu. Kontroluje také omezení přístupu – pokud stránky z komerčních či jiných důvodů nejsou určeny pro veřejný přístup, zobrazí příslušné přístupové omezení (např. - zdroj je možné si prohlédnout pouze v budově knihovny). PANDAS nemá funkci harvesteru, poskytuje pouze interface k software pro stahování dat (nyní HTTrack, v budoucnu se počítá s využitím programu Heritrix) [17].

[12] Preserving and Accessing Networked Documentary Resources of Australia - http://pandora.nla.gov.au/index.html

Všechny archivované zdroje jsou zkatalogizovány ve formátu MARC 21 a tyto katalogizační záznamy jsou součástí Národní bibliografie a také katalogů spolupracujících organizací.

Přístup k archivovaným zdrojům je přes webové rozhraní projektu. Vyhledávání pomocí browsingu je umožněno díky rozdělení archivovaných zdrojů do patnácti tematických kategorií, v rámci kterých jsou pak zdroje řazeny abecedně. Další možností vyhledávání je zadání přesné URL adresy požadovaného zdroje.

Obr.4: PANDORA – úvodní stránka



Většina zdrojů je volně přístupná, několik málo jich je zpřístupněno pouze v omezeném režimu v budově knihovny, jedná se o komerční zdroje, s jejichž vydavateli knihovna vyjednala alespoň takto omezený přístup.

Australský zákon o povinném výtisku je z roku 1968 a nevztahuje se na elektronické zdroje. Proto pracovníci NLA a spolupracujících organizací musí žádat každého vydavatele o povolení ke stahování a následnému zpřístupnění zdroje z archivu.

V současnosti Pandora obsahuje více než 9.000 titulů, celkový objem dat je 906 gigabytů. V loňském roce proběhl průzkum archivu, který odhalil, že zatím pouze 7,4% dokumentů uložených v archivu již nelze najít na webu, zajímavé je i využití archivu - 53% přístupů ze zámoří, 27% z Austrálie, zbytek nelze zjistit.

Zásadním zjištěním je, že se i NLA pokusila sebrat dokumenty z celé australské národní domény. Sklizeň provedl na žádost knihovny Internet Archive během června a července letošního roku. Bylo sklizeno přibližně 185 miliónů unikátních dokumentů z 811 000 URL. Celkový objem takto sklizených dat byl 6,69 terabytů [69].

UKWAC[13]

Konsorcium vzniklo v červnu 2004 jako dvouletý projekt šesti britských institucí pod vedení Britské národní knihovny, které se spojily, aby sdílely náklady, zkušenosti, techniku a technologie nutné k vytvoření archivu britských webových stránek. Britové si pro svůj projekt také zvolili výběrovou metodu archivace a s každým vydavatelem vyjednávají o archivaci a zpřístupnění jeho zdroje z archivu.

Každá z partnerských institucí vybírá a archivuje stránky z oblasti svého zájmu:

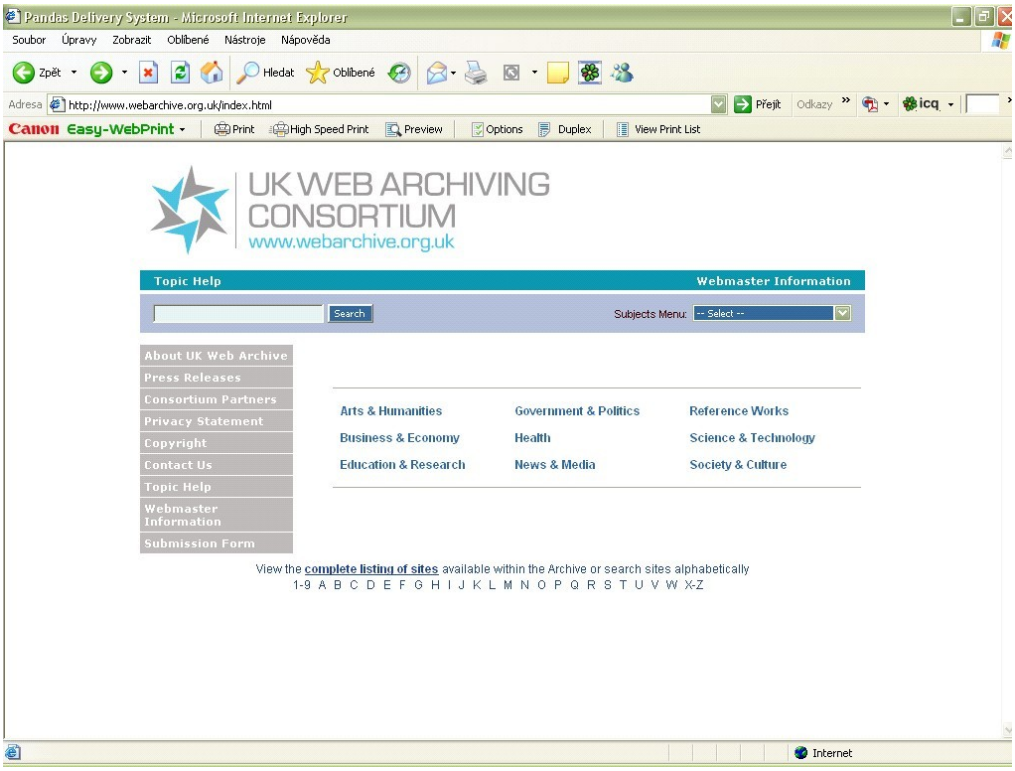
- British Library - stránky s významným kulturním, historickým a politickým obsahem
- The National Archives – vládní publikace
- Wellcome Library - stránky s lékařskou tematikou
- The National Library of Walles - stránky věnované kultuře a historii Walesu
- The National Library of Scotland – stránky věnované kultuře a historii Skotska
- JISC (Joint Information Systems Commitee of the Higher and Further Education Councils) – stránky s vysokoškolskou tematikou [78].

Poté co jsou zdroje archivovány, zkatalogizovány a je provedena kontrola kompletnosti, mohou být zpřístupněny veřejnosti. V archivu je uloženo kolem 6 000

[13] UK Webarchiving Consortium - http://www.webarchive.org.uk/

stránek. Přístup je umožněn přes webové rozhraní projektu a díky tomu, že pro správu archivu je využíván australský software PANDAS má UKWAC s australským projektem PANDORA velmi podobnou grafickou úpravu.

Obr.5: UKWAC – úvodní stránka



Zhodnocení metody výběrové archivace

Výhody výběrové archivace

- Každý zdroj v archivu je kvalitativně posouzen, je zaručena maximální funkčnost
- Každý zdroj může být plně zkatalogizován, a tak se stát součástí národní bibliografie
- Každý zdroj v archivu může být zpřístupněn, závisí to na výsledku jednávání s vydavateli
- Mohou být archivovány i stránky, které jsou nedostupné harvesterům – komerční zdroje, zdroje přístupné přes heslo, databáze.

Nevýhody

- Subjektivní posouzení hodnoty zdroje
- Významné zdroje mohou být opomenuty – nearchivovány
- Vysoké nároky na lidskou práci a finanční prostředky
- Zdroj bývá vytržen z kontextu a odloučen od ostatních zdrojů, se kterými byl provázán odkazy
- Rychlé změny informací a vědeckého prostředí – je těžké odhadnout jak budou vědci využívat informace v budoucnosti

2.2.3 Kombinace plošné a výběrové metody

Jelikož každá z nastíněných metod má svá pro a proti, mnoho zemí se snaží kombinovat obě metody najednou k dosažení optimálního pokrytí online národního kulturního dědictví. V některých případech je takto vytvořený digitální archiv ještě doplněn o tematicky zaměřenou kolekci, kdy jsou vybírány zdroje týkající se jednoho významného tématu. Touto cestou se rozhodly jít země jako například Dánsko či Česká republika.

Největšího pokroku ve výzkumu a praktické aplikaci této metody však dosáhla **Francie.**

Od roku 2001 podniká Francouzská národní knihovna (Bibliothèque nationale de France, dále jako BnF) ve spolupráci s Francouzským národním institutem pro výzkum v oblasti počítačové vědy a automatické kontroly (INRIA) experimenty s archivací národního webu.

Ze zkušeností zahraničních projektů a podle velikosti a povahy webu pracovníci BnF usoudili, že aplikace automatického sběru online dokumentů je nezbytná. Zároveň si ale byli vědomi, že je velice obtížné prostor francouzského webu vymezit. Kritéria jako doména .fr, francouzština nebo fyzická lokace zdroje ve Francii nebyly zcela uspokojivé. V rámci experimentů byly hranice zájmové oblasti vymezeny doménou .fr a vybírány stránky vytipované pracovníky BnF.

Pro harvesting byly využívány technologie vyvinuté v rámci INRIA. Při prováděných experimentech bylo zjištěno, že velký problém představuje doba, po kterou jsou dokumenty stahovány. Stává se, že než celý proces stahování vytipovaného dokumentu proběhne, je již jeho obsah neaktuální. Obsah mnoha webových stránek jako např. elektronické noviny, je měněn i několikrát denně. Proto bylo přistoupeno k automatickému zaznamenávání údajů o aktualizaci stránek, které jsou pak využívány při stanovování frekvence sběru [1].

Dalším důležitým prvkem je zjišťování míry významnosti stránek (Page importance). To je měřeno podobnou metodou, kterou užívají vyhledávací systémy jako např. Google. Původní algoritmus „PageRank“ používaný Googlem vychází z předpokladu, že stránka je významná, pokud na ní odkazuje mnoho dalších stránek.

BnF pomocí automatického výpočtu významnosti stránky určuje skupinu webových zdrojů vhodných k archivaci. První zhodnocení porovnávající vzorek vybraný pomocí automatického výběru s webovými stránkami vybranými pracovníky BnF dle jejich relevance vykazuje dobrý stupeň korelace [1].

I přes nutnost automatického procesu archivace webu klade BnF stále důraz i na lidskou práci. Lidský vklad je nutný zejména při identifikaci a výběru zdrojů z oblasti Deep webu. Pracovníci BnF vyberou relevantní zdroje z této oblasti webu a poté se snaží vyjednávat s jejich vydavateli o možnosti archivace. To je však organizačně i technicky velmi náročný proces. V roce 2002 spustila BnF pilotní projekt, kdy oslovila 100 vydavatelů, s 50 z nich podepsala smlouvu o archivaci, ale pouze 34 jich nakonec své zdroje BnF předalo. Celý proces archivace online zdrojů probíhá v BnF na několika úrovních:

- Automatický harvesting celé národní domény několikrát ročně
- Opakovaný sběr automaticky vybraných stránek – dle jejich významnosti (cca 10% z celkového počtu)
- Sběr dokumentů z oblasti Deep webu
- Tematicky zaměřený sběr stránek na důležité téma

Výsledky těchto experimentálních pokusů zatím BnF bohužel veřejně nezpřístupnila [1].

Zhodnocení kombinované metody

Kombinovaná metoda se zatím jeví jako ideální postup při archivaci online zdrojů na národní úrovni. Na jedné straně se díky harvestingu snaží o co nejúplnější a rozsahem co nejpodrobnější časové snímky celého národního webu, na straně druhé pak pravidelně doplňuje archiv zrcadlící vybranou skupinu významných zdrojů.

2.3 Komparace archivačních metod

Srovnání popsaných metod je velmi obtížné. Zastánci automatického sběru zdůrazňují, že je to relativně levný způsob archivace ve srovnání s výběrovou archivací, zejména co se týká počtu zaměstnanců, tedy personálních nákladů. Na druhé straně existující technologie harvestingu si neumí poradit s mnoha stránkami vytvořenými na základě databáze, mají problémy se stránkami vyžadujícími plug-in nebo obsahujícími skriptovací jazyky. Selektivní přístup tyto problémy umí částečně vyřešit, ale velice omezuje rozsah zdrojů, které budou archivovány.

Automatický sběr by mohl být užitečný pokud je rozsah národní domény relativně malý a jednoduše identifikovatelný. Ideálně pokud by se jednalo o statické stránky propojené standardními HTML odkazy. Ale vzhledem k měnící se povaze webu – s přibývajícím množstvím databází, stránek využívajících Flash či skriptovací jazyky se tento přístup stane čím dál tím méně efektivní. Automatický přístup si zatím umí poradit pouze s veřejně dostupnými zdroji. Pro dokumenty uložené v oblasti Deep webu je zatím reálně možné využít pouze výběrovou metodu archivace.

Důležitým faktorem je také právní otázka, kdy při výběrové archivaci dochází k vyjednávání s vydavateli. Je to sice velice zdlouhavý a nákladný proces, avšak při snaze vyloučení právního rizika jediný možný přístup.

Z těchto důvodů zvažují mnohé země využití obou metod současně – tzn. dát se na cestu třetího způsobu – kombinace obou metod a využít výhody obou přístupů a limitovat jejich omezení.

2.4 Další metody archivace

Existují i jiné metody archivace online zdrojů, které využívají spíše menší organizace k vytvoření menších sbírek. Tyto metody nejsou vhodné pro archivaci webových dokumentů na národní úrovni.

2.4.1 Tematická

Jedná se o druh výběrové archivace, kdy se archivované stránky vztahují k určitému tématu, události či vědní disciplíně. Takový je například projekt Library of Congress s názvem Minerva a projekt DACHS Univerzity v Heidelbergu.

MINERVA[14]

Library of Congress začala na projektu pracovat v roce 2000. Minerva se soustřeďuje na zhodnocení, výběr, sklizeň, katalogizaci, archivaci a zpřístupnění volně dostupných webových zdrojů z určité tematické oblasti. Sklizeň pro Library of Congress provádí Internet Archive. Přes webové rozhraní projektu jsou volně přístupné následující kolekce:

- Volby 2000 (*800 archivovaných stránek*)
- 11. září 2001 (*30 000 archivovaných stránek*)
- Volby 2002 (*4 000 archivovaných stránek*) - pouze část dostupná volně, většina pouze z budovy Library of Congress

Další připravované sbírky: Zimní olympijské hry 2002, Památka na 11.září 2001, 107. Kongres, Válka v Iráku, 108. Kongres, Volby 2004 [60].

DACHS[15]

Projekt Institutu čínských studií na Univerzitě v Heidelbergu, který se soustřeďuje na archivaci a zpřístupnění internetových zdrojů se vztahem ke studiu

[14] Mapping the Internet Electronic Resources Virtual Archive) - http://www.loc.gov/minerva/

[15] Digital Archives for Chinese Studies - http://www.sino.uni-heidelberg.de/dachs/

čínštiny. Důraz je kladen na volně dostupné stránky zaměřené na sociální a politická témata, tak jak je předkládá čínský internet.

Přístup do archivu je umožněn v budově heidelberské univerzity, online přístup je zabezpečen heslem, které mohou získat pouze vědci zabývající se danou problematikou.

Po pěti letech trvání sbírka obsahuje přibližně 2 milióny dokumentů, což odpovídá velikosti zhruba 36 GB [30].

2.4.2 „Deposit" – na základě zákona o povinném výtisku nebo dobrovolného odevzdávání

V některých zemích vydavatelé odevzdávají online publikace na základě zákona nebo dobrovolně. V současnosti neexistuje země, kde by efektivně fungovalo odevzdávání online zdrojů na základě zákona o povinném výtisku, i když některé země jako Velká Británie, Island či Nový Zéland, již takový zákon mají.

Úspěšný model dobrovolného odevzdávání zejména časopisů v elektronické podobě funguje v **Nizozemí**, kde již tradičně nemají (a nikdy neměli) žádný zákon o povinném výtisku. Od roku 2002 uzavírají pracovníci holandského projektu **e-Depot**[16] smlouvy s velkými vydavatelstvími o možnosti dlouhodobé archivace elektronických článků z jejich časopisů. Mezi vydavateli jsou např. Kluwer Academic, Elsevier, BioMed Central, Blackwell Publishing, Oxford University Press, Taylor and Francis, Sage Publications či Springer [33].

2.4.3 VRC (Virtual Remote Control) [17]

VRC představuje inovační metodu, jejímž cílem je za pomocí softwarových nástrojů monitorovat a identifikovat změny na webových stránkách v průběhu času a snažit se archivovat ty, které mají nejvyšší pravděpodobnost zániku. Proces, kterým každá vytipovaná stránka prochází, začíná *identifikací* – podle URL adresy získají informace o struktuře stránky, jméno organizace, instituce či jednotlivce zodpovědného za zveřejnění stránky. Následuje *zhodnocení* – hodnotí se rozsah,

[16] http://www.kb.nl/dnp/e-depot/dm/dm-en.html

[17] http://irisresearch.library.cornell.edu/VRC/index.html

struktura, status stránky, na které navazuje *odhad rizika ztráty* posuzovaný dle určitých kritérií. Na základě tohoto odhadu se rozhodne o *strategii* – buď pasivní monitoring nebo aktivní archivace zdroje [81].

Metoda VRC je vyvíjena na katedře počítačové vědy (Computer Science Department) na Cornell University v New Yorku.

3 METODY ARCHIVACE V RÁMCI ČESKÉHO PROJEKTU WEBARCHIV A NÁVRH OPTIMÁLNÍHO ŘEŠENÍ

Mezi země, které se intenzivně zabývají archivací webu na národní úrovni, se zařadila i Česká republika. **Projekt WebArchiv**[18] vznikl v rámci programového projektu výzkumu a vývoje „Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet“. Je řešen od roku 2000 v Národní knihovně v těsné spolupráci s Moravskou zemskou knihovnou v Brně a Ústavem výpočetní techniky Masarykovy univerzity v Brně.

Hlavním cílem je vyvinout a zajistit vhodné postupy při výběru, sklízení, popisu, uchování a zpřístupnění všech typů elektronických online dokumentů.

3.1 Plošná metoda archivace českých webových zdrojů

Od počátku spuštění projektu WebArchiv se jako jednoznačná možnost archivace jevila plošná metoda při použití v té době rozšířeného a volně dostupného nástroje pro automatické sklízení dat - NEDLIB Harvesteru. Moravská zemská knihovna (dále jako MZK) společně s Ústavem výpočetní techniky Masarykovy univerzity (dále jako ÚVT MU) připravily konfiguraci tohoto nástroje pro prostředí českého webu. Prvním úkolem bylo vymezení prostoru, který bude harvesterem sklízen.

3.1.1 Vymezení oblasti českého webu

Velikost českého webu je velice těžké kvantifikovat, stejně jako web samotný. Zjednodušeně bychom mohli prostor českého webu vymezit jako:

[18] http://www.webarchiv.cz

1) Dokumenty publikované v doméně .cz

Získání údajů o rozsahu domény .cz není jednoduché. Statistiky o celkovém počtu registrovaných domén[19] včetně detailů o každé jednotlivé doméně sice na svých stránkách poskytuje správce domény .cz - sdružení CZ.NIC[20]. Sdružení však bohužel nezveřejňuje kompletní seznam domén druhé úrovně, který je důležitý pro fungování harvesterů.

2) Dokumenty publikované na serverech fyzicky umístěných v ČR

Abychom byli schopni využít tuto možnost, potřebovali bychom získat co nejpřesněji rozsahy IP adres používaných našimi primárními poskytovateli připojení. Takovéto adresy by poté rozšířily databázi adres pro sklízení. Tím by se zajistilo, že při sklizni nebudou vynechány ty servery, na které není odkazováno jménem, ale jen IP adresou.

3) Dokumenty publikované v doménách druhé úrovně, které má zaregistrované organizace se sídlem v ČR

V tomto případě by bylo nutné získat a analyzovat kompletní seznamy domén nejvyšší úrovně, poté analyzovat adresy a telefonní čísla vlastníků jednotlivých domén, a tak rozšířit databázi adres pro sklízení o adresy patřící českým vlastníkům.

4) Dokumenty v českém jazyce

Procházení celosvětového webu s cílem najít stránky v češtině je sice technicky realizovatelné, avšak velice neefektivní. V současnosti je možné tento problém řešit tak, že pokud z dokumentu, který je zařazen v databázi pro sklízení, vedou odkazy na jiný dokument, v jiné doméně než .cz, pak je i tento dokument sklizen a později podroben automatické analýze pro rozpoznání jazyka, resp. češtiny.

5) Dokumenty českých autorů

Nelze rozpoznat automatickými metodami, je nutný intelektuální výběr takovýchto dokumentů.

[19] Aktuální stav ke dni 23. 8. 2005 je 210 767 registrovaných domén v rámci domény .cz

[20] http://www.nic.cz

6) Dokumenty se vztahem k České republice

Nelze rozpoznat automatickými metodami, je nutný intelektuální výběr takovýchto dokumentů [90].

Dnes je již velmi obvyklé, že vydavatelé využívají pro zveřejnění svých dokumentů jinou než národní doménu .cz. Registrátoři v ČR již dnes tuto službu bez problémů nabízejí. Mezi nejoblíbenější patří nadnárodní domény jako: .com, .org, .net, .biz, .info, .name. Od dubna byla spuštěna možnost registrace v rámci domény .eu, ale oficiální provoz je naplánován až na podzim letošního roku.

Hlavním důvodem pro výběr jiné domény je fakt, že jsou často cenově výhodnější než registrace v rámci národní domény.

Tab.1: Srovnání průměrných ročních nákladů registrace některých domén

Doména	Cena (v Kč)
Národní doména - .cz	900–1000
Nadnárodní domény - .com, .org, .net, aj.	200–300
Evropská doména - .eu	1200–3000 (odhad)

Zdroj: webové stránky registrátorů domén v ČR

3.1.2 Softwarové nástroje pro automatizovaný sběr dat

NEDLIB Harvester byl používán od úplných začátků projektu až do roku 2004, kdy byl nahrazen výkonnějším Heritrixem. Pomocí NEDLIB Harvesteru proběhly celkem čtyři celoplošné sklizně domény .cz, žádnou z nich se však nepodařilo zcela dokončit. Důvody přerušení sklizně byly různé. Častou příčinou byly problémy harvesteru, jeho nutné úpravy a vyladění, malá kapacita diskového prostoru. V srpnu roku 2002 sběr přerušil dlouhodobý funkční výpadek síťové infrastruktury NK ČR způsobený tehdejšími povodněmi.

Kromě celoplošné sklizně domény .cz bylo provedeno několik dílčích sklizní. První byl sběr několika tisíc dokumentů ze zpravodajských serverů zaměřených na

povodně 2002, další dílčí sklizně se věnovaly serverům, se kterými měla v té době NK ČR uzavřenou smlouvu na uchování a zpřístupnění zdroje z archivu. Hledala se při tom taková nastavení harvesteru, která by umožnila efektivnější periodické sklízení a archivaci zadané množiny serverů [10].

Důvodem ukončení využívání služeb NEDLIB Harvesteru byla jednak řada neřešitelných problémů, jednak skončily vývojové práce na tomto nástroji a také se objevil nový volně dostupný harvester – Heritrix. Ten prokázal značnou stabilitu a má mnoho vlastností, které scházely NEDLIB Harvesteru. Proto bylo rozhodnuto, že všechny následující sklizně budou probíhat již výhradně pomocí Heritrixu [10].

3.1.3 Uložení stažených dat

Velikost harvesterem tvořeného archivu dosahuje obrovských rozměrů: jedno kolo stahování představuje v našich podmínkách stovky gigabytů (GB) a celkový objem stažených dat již překročil hranici 1 terabytu (TB). Archiv s tak velkým potenciálem růstu není snadné ani levné provozovat. Ačkoli v současné době jsou již na trhu dostupné pevné disky s velkou kapacitou za nízkou cenu, infrastruktura archivu se musí opírat o robustní a dlouhodobě perspektivní řešení. Toto řešení musí brát v potaz nejen problémy technické, ale i finanční a personální a musí být z provozního hlediska i dlouhodobě únosné. Nejde tedy jen o to uložit někam jednorázově 1 TB dat, ale o to, aby byla tato data trvale online přístupná, aby byla zajištěna průběžná rozšiřitelnost archivu, zálohování dat a v neposlední řadě i jeho správa a údržba [90].

V pilotní fázi projektu byl pro ukládání dat využíván páskový robot. Jeho výhodou byla bezpečnost na něm uložených dat a především jeho snadná rozšiřitelnost. V průběhu sklizně 2002 se prokázalo i množství jeho nevýhod – nízká rychlost, nedostatečná úložná kapacita a vysoká cena jejího rozšiřování, časté výpadky jeho provozu. Jako cenově dostupná a technicky schůdná strategie pro řešení problému uložení velkého objemu dat bylo nakonec zvoleno pořízení hardwarového diskového pole a jeho umístění v MZK. Přesunutím dat na diskové pole došlo ke značnému urychlení všech procesů souvisejících s archivací a správou digitálního archivu. Stávající kapacita diskového pole použitelná pro uložená data je 2,5 TB,

nákupem nových 7 disků o kapacitě 400 GB v roce 2004 se použitelná kapacita rozšířila cca na 5 TB.

V posledních měsících se úvahy o budoucnosti uložení dat WebArchivu ubírají jiným směrem. Národní knihovna ČR se rozhodla využít služeb centrálního diskového úložiště, které by bylo schopné zajistit požadavky velkých objemů dat, průtočnosti a zabezpečení. Bylo vypsáno výběrové řízení na poskytovatele služby. V současné době jsou zvažovány nabídky, rozhodnutí by mělo padnout v horizontu měsíců. Tím by se mohly vyřešit každoroční velké problémy s ukládáním dat WebArchivu a řešitelský tým by se mohl soustředit na další důležité otázky archivace.

3.1.4 Využití služeb Internet Archive

Teoreticky vzniká otázka, proč se vůbec Národní knihovna složitě snaží o archivaci českých online zdrojů, pokud existuje projekt Internet Archive, jehož archiv lze neomezeně prohledávat a české zdroje tam najít.

Národní knihovna má depozitní funkci, má tudíž povinnost archivovat veškerou českou produkci. Internet Archive sice archivuje celosvětový web, a tak v jeho archivu najdeme i české zdroje, ovšem NK ČR se nemůže spolehnout pouze na zahraniční projekt a doufat, že sklidí všechny zdroje, které by řešitelé projektu WebArchiv považovali za významné. Dále neexistuje záruka, že Internet Archive bude stále existovat a vykonávat svou činnost. Pokud by se jeho tvůrci rozhodli aktivitu ukončit, jistě by stáhli i celý archiv a část českého kulturního dědictví by byla nenávratně ztracena. I přes tyto jasné argumenty jsem se pokusila udělat malý test a vyhodnotit, jaké nedostatky má Internet Archive z hlediska WebArchivu.

Dokumenty uložené v archivu:

- Problémy se zobrazením složitějších stránek
- Přesměrovací problémy - archivované verze jsou přesměrovávány na aktuální stránku
- Chybějící obrázky – zejména u starších verzí
- Nefunkční odkazy- mnoho z testovaných odkazů nevrátilo žádnou odpověď

- Nefunkční multimediální obsah
- Nepravidelné intervaly sklízení dokumentů
- Absence některých českých zdrojů

Prohledávání archivu (softwarový nástroj WayBack Machine):

- Vyhledávání možné pouze přes URL
- Zobrazování výsledků dotazu pouze v angličtině

3.2 Výběrová metoda archivace českých webových zdrojů

Z prvních výsledků plošné metody za použití NEDLIB Harvestru bylo zřejmé, že tento způsob archivace nebude stačit k pokrytí všech dokumentů národního významu. Problematická byla zejména vysoká časová náročnost jedné celoplošné sklizně.

Dalším důvodem pro rozhodnutí o aplikaci výběrové metody archivace byl stávající stav české legislativy.

3.2.1 Povinný výtisk

V České republice existují dva zákony týkající se povinného výtisku, 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 online zdrojů jakožto povinného výtisku u nás zákon jednoznačně neurčuje.

Zákon č. 37/1995 Sb. o neperiodických publikacích

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 ovšem zůstává fakt, že zákon se vztahuje pouze na monografické publikace (jednorázově vydané publikace), kterých v dnešní době na Internetu mnoho nenajdeme [24].

Zákon č. 46/2000 Sb. tzv. Tiskový zákon

Horší je situace v případě druhém, tzv. tiskovém zákoně. Jak už 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 z celkových 19 paragrafů (par.9). Proto se domnívám, že systematickým řešením by bylo tento paragraf vyjmout a zařadit ho do nového obecně formulovaného zákona o povinném výtisku [24].

Novela zákona/zákonů

Novela zákonů týkajících se povinného 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, je nutné, aby se právě ona o tuto novelizaci zasadila. Příklady podobných opatření existují v mnoha ostatních zemích.

Příprava kvalitního zákona však bude velice náročná. Základním rozhodnutím zřejmě 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 zahrnutou, ale ne všude byl účel zákona naplněn [24].

Existuje již několik modelů zákona o povinném výtisku. Ty se mohou stát inspirací, lze se z nich poučit o výhodách či nevýhodách jednotlivých modelů.

3.2.2 Autorské právo

Jelikož cílem WebArchivu je nejen elektronické online zdroje sbírat a uchovávat, ale také je zpřístupňovat veřejnosti, musíme se zaměřit i na další důležitý právní předpis, kterým je zákon č. 121/2000 Sb. tzv. Autorský zákon.

Jeho stávající verze knihovně sice dovoluje digitální archiv vytvářet, ale neumožňuje zdroje z archivu jakýmkoli způsobem zpřístupnit. Východiskem z této složité situace by měla 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. Tato směrnice v jednom ze svých článků doporučuje vládám členských států, aby umožnily zpřístupňování autorských děl (včetně jejich online podoby), 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)) [24].

Tento článek se zástupcům NK ČR podařilo prosadit do novely autorského zákona, která je v poslední fázi své přípravy před prvním čtením v Parlamentu České republiky.

3.2.3 Alternativní řešení

Jelikož je situace získávání, archivace a zpřístupnění online zdrojů takto komplikovaná nevyhovujícími zákony, byli řešitelé projektu nuceni přistoupit na alternativní řešení. Jediným možným východiskem zůstalo oslovování jednotlivých vydavatelů a uzavírání smluv o poskytování elektronických online zdrojů. Smlouva byla připravena ve spolupráci s právníky na základě doporučení CENL/FEP (*viz 2.1.1.1*).

Jednotliví vydavatelé jsou oslovováni prostřednictvím elektronické pošty a pokud na nabídku WebArchivu zareagují, je jim zaslána smlouva v tištěné podobě. V současné době má tedy Národní knihovna ČR v rámci projektu WebArchiv uzavřeno 40 Smluv o poskytování elektronických online zdrojů. Seznam spolupracujících vydavatelů a jejich zdrojů lze najít na webových stránkách projektu, konkrétně na: http://www.webarchiv.cz/partneri.html.

Již delší dobu se však tento zdlouhavý proces oslovování vydavatelů jeví jako nevyhovující. Snahou je vyloučit z procesu uzavírání smluv tištěnou podobu smlouvy a nahradit ji elektronickým souhlasem. Tím není míněn elektronický podpis (z důvodu stále ještě malého rozšíření), nýbrž potvrzení vydavatele, že souhlasí s archivací a zpřístupněním svého zdroje z archivu prostřednictvím elektronické pošty. Souhlas by měl vydavatel vyjádřit pouze vyplněním údajů o své osobě, o svém zdroji a e-mailem by tuto informaci měl odeslat do Národní knihovny. Nabízené řešení je využíváno při oslovování vydavatelů například v rámci projektu PANDORA. Bohužel pracovníci právního oddělení NK ČR tento způsob „uzavírání smlouvy“ zatím odmítají.

3.2.4 Kritéria výběru zdrojů

V první fázi se řešitelé WebArchivu z celého spektra zdrojů, které prostor českého webu nabízel, zaměřili pouze na malou skupinu elektronických časopisů, existující pouze v online formě. Výběr časopisů byl proveden na základě spolupráce s vysokoškolskými a dalšími odbornými knihovnami. Základním kritériem výběru byla neomezená dostupnost na webu, vydávání minimálně po dobu jednoho roku a nereklamní charakter časopisu.

Postupem času se množina výběrových kritérií rozšířila, stanovení kritérií bylo pojato poněkud v širším kontextu. Zdroje byly primárně vybírány:

- Podle domény – zejména zdroje v doméně .cz
- Podle obsahu zdroje – důraz je kladen na informační hodnotu zdroje
- Podle typu zdroje – především seriály, konferenční materiály, výzkumné zprávy apod.
- Podle formy – zdroje existující pouze v online podobě
- Podle přístupu – pouze volně přístupné zdroje
- Podle formátu – preferovány jsou všeobecně podporované formáty

Kritéria výběru zdrojů je nutné upravit a rozpracovat do větších podrobností.

Potřeba spolupráce

V rámci procesu oslovování potencionálních vydavatelů online zdrojů začala Národní knihovna v roce 2003 spolupracovat 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. Součástí tohoto formuláře je i otázka, zda vydavatel souhlasí se zařazením svého zdroje do WebArchivu. Rovněž jsou zde základní informace o projektu a jeho cílech.

Tímto krokem došlo k usnadnění práce při vyhledávání významných a zajímavých online seriálů, které odpovídají zvoleným kritériím. Bohužel zdaleka ne všechny seriály, kterým je přiděleno ISSN, tato kritéria splňují. Zejména po obsahové stránce je mnoho zdrojů nevhodných (patří mezi ně servery se společenskou tematikou, stránky věnované hrám, mobilním telefonům apod.). Dalším cílem této spolupráce bylo dosažení větší propagace a rozšíření povědomí o projektu mezi vydavateli online zdrojů.

Pouhá spolupráce se střediskem ISSN však rozhodně nestačí. Při takovém množství různých typů zdrojů je nutné rozdělit odpovědnost za uchování kulturního dědictví v elektronické podobě mezi více subjektů. Národní knihovna bez legislativního zajištění není schopná registrovat veškerou publikační produkci hodnou archivace. Je proto třeba rozdělit kompetence a definovat oblasti zájmu, kterým se další organizace budou věnovat.

V letošním roce se podařilo navázat kontakt s pracovníky Národního archivu ČR. Jednání se týkalo možnosti spolupráce v oblasti archivace online dokumentů veřejné správy. Bylo přislíbeno, že Národní archiv bude sledovat významné webové stránky z této oblasti a následně zašle do Národní knihovny seznam relevantních URL. Tato množina dokumentů bude následně sklizena a uložena v rámci projektu WebArchiv.

Je třeba, aby se počet spolupracujících organizací rozrostl o další. Je plánováno oslovení akademických organizací (pravděpodobně jejich knihoven), dalších speciálních archivů (např. Národní filmový archiv), odborných knihoven (Národní lékařská knihovna, Parlamentní knihovna, Knihovna Akademie věd, aj.) a krajských knihoven, aby se zaměřily na online zdroje národního významu tematicky či regionálně spadající do oblasti jejich zájmu.

3.3 Vývoj finančního a personálního zabezpečení projektu WebArchiv

Jak již bylo řečeno v předchozích kapitolách, výběr strategie archivace závisí na stavu legislativy (povinný výtisk a autorské právo), možnostech spolupráce s dalšími organizacemi zabývajícími se archivací webu jak na národní tak na mezinárodní úrovni, ale významnou měrou také na finančních možnostech knihovny a jejím personálním zabezpečení, tj. počtu pracovníků a jejich profesním zaměření.

Dosavadní podmínky umožňují provádět činnosti související s archivací českého webu na úrovni experimentální. Po celou dobu trvání (od roku 2000) jsou procesy řešené v rámci projektu financovány téměř výhradně z jednotlivých grantových prostředků, tedy mimo rozpočet NK ČR. K tomu, aby dosavadní úspěšné řešení archivace českého webu mohlo postupně přejít z fáze testování do praktického provozu, nemůže nikdy dojít bez mimořádných finančních prostředků jak na velkokapacitní paměťová média a další potřebnou techniku, tak na významné personální posílení.

Tab.2: Vývoj finančního a personálního zabezpečení WebArchivu

	2000	2001	2002	2003	2004	2005-plán
Název grantu	Výzkum a vývoj MK ČR[1]	Výzkum a vývoj MK ČR	VISK[2] 3	VISK 8/B, v rámci grantu JIB[3]	VISK 8/B, v rámci grantu JIB	VISK 8/B, v rámci grantu JIB
					Výzkumný záměr MK ČR, v rámci grantu JIB	Výzkumný záměr MK ČR, v rámci grantu JIB
Investice (Kč)	330.000	0	650.000	108.000	120.000 + 130.000	80.000 + 35.000
Neinvestice (Kč)	1.057.000	956.000	350.000	440.000	400.000	400.000
Granty celkem (Kč)	1.387.000	956.000	1.000.000	548.000	520.000	480.000
Příspěvek	440.000	290.000	551.000	120.000	120.000	120.000

NK[4] (Kč)						
CELKEM	1.827.000	1.246.000	1.551.000	668.000	770.000	635.000
Počet zaměstnanců	1	1	2	2	2	2 1/4

Vysvětlivky: [1] – Ministerstvo kultury České republiky; [2] – Veřejné informační služby knihoven; [3] – Jednotná informační brána; [4] – Národní knihovna České republiky

3.4 Návrh optimálního řešení pro WebArchiv

Jelikož se jako nejvhodnější metoda, na kterou ukazuje i zahraniční praxe, jeví kombinovaný přístup k archivaci webu na národní úrovni, bude tato využívána i v rámci českého projektu WebArchiv. K tomu je třeba pokračovat ve vývoji a konfiguraci Heritrixu pro české prostředí, zajistit dostatečnou kapacitu pro uložení stažených dokumentů, připravit kvalitní a jasnou politiku výběru zdrojů do digitálního archivu a začít intenzivně spolupracovat s dalšími institucemi.

3.4.1 Řešení v krátkodobém horizontu

I nadále se bude pokračovat v automatickém sběru digitálních dokumentů. K dispozici je již výkonný harvester Heritrix, který by z testovací fáze měl postupně přejít do plného provozu.

Po stanovení politiky výběru zdrojů do digitálního archivu bude pozornost věnována i výběrové metodě archivace. Kritéria, kterým je věnována kapitola 6, vymezí typy dokumentů, které budou pomocí Heritrixu taktéž sbírány. U těchto zdrojů je třeba určit optimální frekvenci jejich sběru, aby mohly být archivovány prodělané změny v dokumentu v co největším rozsahu.

Během roku budou sledována i zajímavá témata s cílem doplnění sbírky archivovaných dokumentů o tematicky zaměřenou kolekci. Kandidátem na takové téma budou jistě např. parlamentní volby 2006.

Problémem však zůstává finanční zabezpečení projektu. Výběrová metoda archivace je nákladná, a to zejména z důvodu vysoké náročnosti na vklad lidské práce. V Austrálii, rámci projektu PANDORA, byla provedena analýza průměrné ceny jednotky archivované pomocí výběrové metody. Výsledkem bylo zjištění, že lidská práce tvoří náklady na tuto jednotku z 94% [70].

I toto zjištění dokládá potřebu dosáhnout navýšení v oblasti personálního zabezpečení projektu WebArchiv. Realizaci australského projektu zajišťuje celkem 8 pracovníků na plný úvazek + 4 pracovníci věnující se na ¼ úvazku technickým záležitostem projetu – zejména úpravám systému a jeho údržbě. Minimálním požadavkem řešitelů projektu WebArchiv je 4 - 5 pracovníků zaměstnaných na plný úvazek, z toho jeden věnující se technickým otázkám.

Finanční prostředky je třeba zajistit z grantů a částečně i z rozpočtu Národní knihovny ČR. Další možností získání nadstandardních financí je z evropského grantu, který WebArchiv získal v rámci programu Culture 2000 (*viz 6.2.2*).

Propagace projektu a uzavírání smluv s vydavateli

Dalším důležitým úkolem je změna dosavadní praxe uzavírání smluv s vydavateli. Je třeba zjednodušit celou agendu tak, aby probíhala výlučně prostřednictvím elektronické pošty, a to včetně udělení souhlasu vydavatele s archivací a zpřístupněním zdrojů. Tím by mohlo být dosaženo většího nárůstu počtu titulů v archivu. Pracovníci WebArchivu se budou snažit i o lepší komunikaci s vydavateli. Pokusí se oslovit Unii vydavatelů[21] s cílem uspořádání semináře na téma WebArchiv. Setkání by mělo řešit následující body:

- Seznámit je s existenci WebArchivu, s jeho cíli, posláním a procesy
- Identifikovat výhody pro obě strany
- Zdůraznit výhody dlouhodobé archivace
- Hledat pro obě strany výhodné řešení

[21] http://www.uvdt.cz

- Poučit je, resp. vysvětlit jim výhody využívání vhodných standardů, formátů, metadat apod.
- Vyjednat podmínky stahování a autorských práv
- Navrhnout ideální znění smlouvy

Pokud by vydavatelé sami chtěli přihlásit svůj zdroj do archivu, pak bude pro tento účel vytvořen webový formulář umístěný na stránkách projektu, jehož prostřednictvím může vydavatel svůj zdroj zaregistrovat.

V rámci lepší propagace projektu jsou plánovány další přednášky na odborných akcích, publikování článků v odborných elektronických časopisech, vystoupení v médiích. Větší povědomí o projektu by mělo být zajištěno i díky rozvoji spolupráce s dalšími institucemi.

Optimální řešení

Pokud se podaří navýšit počet zaměstnanců a stanovit správnou politiku výběru zdrojů do archivu, pak by optimálním řešením mohlo být:

- Automatický harvesting všech dokumentů v rámci národní domény .cz několikrát ročně (přibližný odhad 4x ročně)
- Pravidelný sběr souboru vybraných dokumentů – dle jejich periodicity
- Sběr vytipovaných relevantních zdrojů mimo doménu .cz
- Sběr vybraných dokumentů z oblasti Deep webu
- Tematicky zaměřený sběr na důležité téma

3.4.2 Řešení v dlouhodobém horizontu

V dlouhodobém horizontu je třeba snažit se zejména o změnu zákonů o povinném výtisku. Studie požadavků a návrhů řešitelů WebArchivu na změny by měla být vypracována do poloviny příštího roku. Poté bude předána kompetentním zástupcům Národní knihovny a předložena Ministerstvu kultury.

Pokud se podaří vyjednat spolupráci s ostatními institucemi zabývajícími se ochranou kulturního dědictví, bude třeba vypracovat společný postup prací a definovat přesné vymezení kompetencí.

Důležitým cílem bude získání členství v konsorciu IIPC, snaha podílet se na vývoji moderních softwarových nástrojů a ty následně využít pro potřeby archivu.

V dlouhodobém horizontu se počítá i s rozvojem stávajících technologií a usnadněním celého procesu archivace.

4 KRITÉRIA VÝBĚRU ELEKTRONICKÝCH ONLINE ZDROJŮ

Jak z našich tak i ze zahraničních zkušeností vyplývá, že nejoptimálnější metodou pro tvorbu digitálního archivu je kombinace plošné a výběrové metody archivace doplněná o tematickou kolekci. Jak probíhá harvesting, bylo již popsáno, nyní je třeba zaměřit se na definování kritérií, dle kterých budou zdroje vybírány. Víme, že harvesting má své limity, nedokáže např. archivovat zdroje z oblasti Deep webu.

Země, které praktikují výběrovou metodu archivace, se snažily definovat kritéria výběru. Soubor kritérií však žádná z nich nemá zatím propracován dokonale, jedná se spíše o pracovní verze, které se budou s postupem času dále vyvíjet.

Proces výběru je intelektuální práce – je třeba rozhodnout, co uchovat a co vyloučit. Kritéria výběru jsou nejčastěji zaměřena na zhodnocení obsahu dokumentu – jeho předmět či disciplínu ve vztahu ke sbírkovým cílům organizace, jeho kvalitu a unikátnost, jeho přístupnost ve smyslu dostupného hardwaru a softwaru. Musí prezentovat určitou, zejména budoucí hodnotu.

To, do jaké hloubky mají jednotlivé organizace kritéria rozpracována, závisí na vytvořených předpokladech (hmotných, legislativních, na počtu spolupracujících institucí). Při malých finančních možnostech, absenci legislativy a malém počtu zaměstnanců jsou kritéria definována spíše na obecné úrovni, země s lepšími možnostmi mívají kritéria propracována do větší hloubky.

Významnou roli v procesu evaluace zdrojů hrají zaměstnanci organizace. Ti, kteří zdroje vybírají, by měli dobře rozumět internetovým technologiím, aby správně rozpoznali, zda je zdroj možné dlouhodobě uchovat či nikoli.

Často se setkáváme se situací, kdy se archivací online zdrojů na národní úrovni nezabývá pouze jediná instituce, ale existuje zde určitá forma spolupráce s dalšími. Partnerské organizace si mezi sebou vymezí kompetence a oblast zájmu, na kterou se zaměří. Oblasti zájmu korespondují s cíli a zaměřením fondu organizace. Spolupráce národních knihoven probíhá většinou s národními archivy, akademickými organizacemi, odbornými a regionálními knihovnami či instituty. Každá

z partnerských organizací si mnohdy základní definovaná kritéria přizpůsobí vlastním požadavkům a stanoví si vlastní politiku výběru.

4.1 Obecná kritéria

Množství online zdrojů, kterými bychom se měli zabývat, je daleko větší, než je možné zvládnout. Ale nejen kvantita, ale i kvalita je důvodem pro stanovení kritérií.

4.1.1 Vymezení publikace

Hranice publikace jsou často velmi neostré, je velice obtížné je vymezit. Často se stává, že publikace obsahuje množství interních i externích hypertextových odkazů. Všechny odkazy na stránce jsou prozkoumány a je zjišťováno, jaké části tvoří stránku jako takovou pro účely archivace a katalogizace. Většinou jsou archivovány pouze interní linky. V některých případech se preferuje rozložit objemné stránky na menší entity a vybrat pouze takové, které odpovídají požadavkům. Někdy se však stane, že jednotlivé části rozsáhlé publikace nemají samostatně význam a pouze dohromady tvoří hodnotný zdroj informací. V takovém případě, pokud to odpovídá kritériím, je publikace archivována jako celek.

4.1.2 Online x tištěná verze

Pokud má zdroj obě verze, pak je preferována tištěná před online verzí. Online verze jsou vybrány v případě, že obsahují další významné informace či jinou přidanou hodnotu.

Pokud je dostupná verze online a zároveň na CD-ROM nebo disketě, pak bude vybrána online verze, jelikož její dlouhodobá archivace je snazší než u optických či magnetických médií.

4.1.3 Časové verze

Není možné archivovat všechny verze/edice/vydání online zdroje, jelikož se v mnoha případech mění velice často. Musí být rozhodnuto, s jakou frekvencí bude

zdroj stahován a archivován. Design publikace, významnost obsažených informací a stabilita stránek – to jsou faktory, které ovlivňují toto rozhodnutí. Některé významné zdroje jsou archivovány tak často, jak to je možné, některé mohou mít jen jednu archivovanou verzi.

4.1.4 Volně přístupné zdroje x komerční zdroje

Přednostně jsou vybírány k archivaci volně přístupné zdroje, jelikož u nich je větší pravděpodobnost, že vydavatel bude souhlasit s archivací. U komerčních publikací je však často zaručena vysoká kvalita obsahu, proto je snaha zařadit co nejvíce těchto publikací do archivu. Je zde ale vyšší riziko neúspěchu – vydavatel nemusí souhlasit s archivací a zpřístupněním svého zdroje, může získat pocit, že by mohlo dojít k ohrožení jeho komerčních zájmů.

4.1.5 Formáty

Formáty online publikací by neměly být překážkou pro výběr. V současnosti ale některé zdroje nemohou být archivovány z technických důvodů. Například zatím nelze archivovat dokumenty, které jsou strukturovány jako databáze, problematické jsou rovněž zdroje obsahující JavaScript či Flash. Knihovny se v reakci na tato omezení snaží nalézat nová technická řešení.

4.1.6 Obsahová kritéria

Základní obsahová kritéria jsou většinou vesměs shodná s kritérii, dle kterých jsou vybírány tradiční dokumenty pro národní bibliografii.

Jsou archivovány zdroje národního významu, jejichž obsah:

- Se týká dané země
- Zahrnuje významné informace z oblasti sociálního, politického, kulturního, náboženského, vědeckého nebo ekonomického života dané země nebo je tento obsah vytvořen národním autorem
- Je vytvořen národním autorem a toto dílo dosahuje mezinárodního významu

Zdroj může být uložen na domácím či zahraničním serveru. Pouhé národní autorství nebo editorská práce nejsou postačujícími faktory pro národní archivaci. Pro online publikace je nejdůležitější obsah [65].

4.2 Kvalitativní kritéria

Kvalita zdrojů zveřejněných na Internetu se různí. Je to dáno zejména svobodou Internetu, snadnou možností publikování čehokoli, prakticky bez jakéhokoli filtračního mechanismu. Proto je nutné zaměřit se na několik parametrů, které by měl zdroj pro výběr do archivu splňovat.

4.2.1 Validita zdroje

Důležitou charakteristikou zdroje je obsah a jeho platnost. Je třeba posoudit, jestli skutečný obsah zdroje odpovídá tomu předpokládanému, zda obsažené informace jsou výsledkem výzkumu či bádání a zdroj obsahuje odkazy, bibliografii či seznam použité literatury. Je také třeba, aby bylo možné provést jeho verifikaci, tj. informace musí být možné zkontrolovat. Validní zdroj by měl obsahovat základní popisné informace, zejména údaje o autorech, adresu, telefon apod.

4.2.2 Důvěryhodnost zdroje

Vysoká priorita je přisuzována věrohodným publikacím. Posuzován je zejména autor, vydavatel či poskytovatel zdroje, jeho renomé. Musí být uvedeny kontaktní údaje, zkoumají se informace o autorovi/autorech, jejich citovanost. Je třeba si povšimnout, zda je zdroj podporován či sponzorován nějakou institucí a o jakou instituci se jedná. Důraz je kladen také na to, zda zdroj prošel redakční či editorskou kontrolou.

4.2.3 Věcnost zdroje

Hodnotí se především množství informací, které příslušný zdroj obsahuje. Kriterium věcnosti porovnává, jak jsou obsažené informace v kontextu ostatních

významné, podstatné. Pokud jde například pouze o prosté oznámení, reklamu, odkaz či kontakt, nelze daný zdroj obvykle považovat za podstatný. Přednost se dává plnotextovým zdrojům před bibliografickými.

4.2.4 Přesnost zdroje

Vysoká priorita je dávána přesnosti poskytovaných informací a je důležité, zda-li je umožněna její kontrola, například podle citovaných zdrojů, odkazů, bibliografie (jsou-li obsaženy). Přesnost zdroje, resp. jím poskytovaných informací se hodnotí i po stránce formální – gramatická správnost, hláskování, výskyt typografických chyb apod.

4.2.5 Úplnost zdroje

Význam zdroje udává hloubka zpracování daného tématu. Sleduje se detail poskytovaných informací, zda-li jsou povrchní nebo vyčerpávající. Důležitým aspektem je úvodní část obsahující základní údaje o zdroji a tematice. Následně se hodnotí, zda jsou pokryty všechny aspekty tématu, zda lze najít vše, co je od zdroje možné očekávat.

4.2.6 Jedinečnost zdroje

Jedinečnost udává množství primárních informací, které zdroj obsahuje. Pokud online publikace obsahuje pouze povrchní informace, které jsou jednoduše dostupné v jiných publikacích (ať už v tištěných či elektronických) nebo jejich obsah je z velké části kompilací, nejsou tyto publikace archivovány.

4.2.7 Skladba a organizace zdroje

Pozornost je věnována organizaci informací v rámci zdroje, zda jsou uspořádány do logických celků, zda jsou informace konzistentní. Při skladbě zdroje by měly být brány v úvahu potřeby uživatelů. Důležité hledisko je také dodržování základních pravidel gramatiky, hláskování případně používání žargonu [74].

4.3 Kritéria dle typu dokumentu

Knihovna je schopná archivovat pouze omezený objem zdrojů. Při stanovování kritérií pro jejich výběr hrají významnou roli kategorie dokumentů. Následující výčet ukazuje, jaké typy dokumentů jsou všeobecně preferovány, které jsou archivovány pouze zřídka a také ty, jenž jsou prozatím z archivace vyřazeny.

4.3.1 Archivace je zaměřena na následující kategorie:

- Dokumenty veřejné správy a samosprávy
- Zdroje akademických institucí
- Materiály z konferencí
- Elektronické časopisy
- Tematicky zaměřené zdroje
- Zdroje unikátní z pohledu designu, použitých technologií

4.3.2 Kategorie dokumentů, které jsou zřídka archivovány:

- Databáze a zdroje z oblasti Deep webu
- Elektronické knihy, literární díla, hudebniny
- Zpravodajské servery
- Mapy
- Osobní stránky a texty

4.3.3 Kategorie dokumentů, které většinou nejsou archivovány (výjimky možné):

- Blogy
- Web-kamery
- Datové soubory
- Diskusní skupiny, chaty

- Koncepty a připravovaná díla (i když odpovídají výběrovým kritériím)
- Dokumenty dostupné na intranetu
- Hry
- Portály a jiné stránky, které slouží pouze pro setřídění informací na internetu
- Stránky reklamního charakteru a propagační stránky
- Stránky, které jsou pouhou kompilací informací z jiných zdrojů a nemají vlastní originální obsah

4.3.4 Kritéria výběru prioritních kategorií dokumentů

Vládní publikace

Jelikož se mnoho dokumentů veřejné správy, kromě jiného z důvodu naplňování principu e-Governmentu, přesunulo na web, je třeba je archivovat. Tyto publikace jsou považovány za velmi významné, často obsahují unikátní informace. V mnoha zemích dokonce funguje povinnost opatřovat tyto dokumenty metadaty, tudíž je také jednodušší je stahovat a archivovat. V některých zemích se začalo uvažovat o automatickém sběru, jednak z důvodu obsažených metadat ve zdroji a také díky existenci speciálně vyčleněných národních domén druhého řádu jako například .gov.au či .gov.uk.

Nespornou výhodou dokumentů veřejné správy je i fakt, že se na ně nevztahuje autorskoprávní ochrana, proto je lze archivovat bez předchozího vyjednávání s jejich vydavateli.

Publikace akademických organizací

Akademické organizace již dnes často budují archivy vlastních publikací, např. mnoho vysokých škol dnes zaměřuje na archivaci kvalifikačních prací. Tyto typy institucí jsou producenty velmi významných zdrojů zejména z oblasti výzkumu a vývoje. Do doby než se všechny instituce zapojí do procesu archivace, zůstává zodpovědnost na knihovnách.

Materiály z konferencí

Na webu je každým rokem publikováno velké množství literatury pocházející z konferencí. Také v rámci těchto materiálů je třeba udělat selekci.

Archivují se většinou pouze stránky, které obsahují podstatného množství plných textů příspěvků přednesených na konferenci. Pokud jsou k dispozici pouze powerpointové prezentace, nejsou stránky archivovány.

Priorita je přisuzována konferencím, které pořádají profesionální asociace, vládní či akademické organizace. Jsou preferovány velké konference před malými semináři [65].

Elektronické časopisy

Elektronické časopisy jsou považovány za jednu z nejvýznamnějších kategorií. Archivace ze zaměřuje na:

- Časopisy s určitou tradicí
- Časopisy jejichž vydavatelé patří mezi renomované
- Časopisy věnované odborným tématům
- Časopisy s dobrou editorskou kontrolu
- Komerční časopisy [65]

Tematicky zaměřené zdroje

Měla by být definována klíčová témata, která budou archivována. Většinou jsou vybírány zdroje dokumentující významné sociální, politické, kulturní, náboženské, vědecké nebo ekonomické události.

Zdroje unikátní z pohledu designu či použitých technologií

Tato kategorie dokumentů se jako jediná primárně netýká obsahu. Vychází se z předpokladu, že v budoucím horizontu by mohlo být pro některé odborníky zajímavé, jakým vývojem prošly webové aplikace. Archivace je zaměřena na stránky

se zajímavým designem, používající výjimečné technologie či stránky s nevšedním způsobem uspořádání.

Databáze a zdroje z oblasti Deep webu

V oblasti Deep webu je uloženo mnoho významných informačních zdrojů, které jsou vhodné pro archivaci. Z nich je třeba vybrat ty nejzajímavější (často strukturované jako databáze) a pokusit se je sklidit a dlouhodobě archivovat. První pokusy právě probíhají ve Francii.

Elektronické knihy, literární díla, hudebniny

Z literárních děl jsou vybírána pouze díla zásadního charakteru – monografie či sbírky básní. Jednotlivé či nahodilé literární pokusy archivovány nejsou.

V oblasti hudebnin se musí jednat o originální dílo skladatele dané země publikované na webu známým vydavatelstvím nebo dílo již uznávaného skladatele či dílo vztahující se svým obsahem k dané zemi [65].

Zpravodajské servery, noviny

Jsou vybírány pouze takové servery, které neduplikují informace dostupné v tištěné verzi. Existují případy, kdy elektronické verze a tištěné verze stejného deníku tvoří jiní novináři, pak se noviny samozřejmě liší i v obsahu a stojí zato je archivovat.

U těchto zdrojů je kladen důraz na věrohodnost, kvalitu a originalitu obsahu. Jelikož se zejména zpravodajské servery aktualizují mnohokrát denně, je třeba najít vhodnou frekvenci jejich sběru.

Mapy

Online mapy jsou archivovány pouze v případě, že informace reprezentovaná na mapě není dostupná v jiné podobě nebo online verze poskytuje další významné

informace. Data musí pocházet od věrohodného zdroje, zveřejněná data musí být ve standardní kvalitě a obsahovat dobrý kartografický popis [65].

Stránky organizací, osobní stránky

Tento typ dokumentů je pro archivaci vybírán spíše výjimečně. V případě stránek organizací jsou vybírány takové, které poskytují zásadní informace o projektech, výzkumu apod. Osobní stránky jsou vybírány pouze v případě, že se jedná o stránky mimořádné kvality, výjimečné odborné hodnoty, stránky poskytující informace, které nejsou nikde jinde publikovány.

4.4 Shrnutí

Aplikace tohoto přístupu negativně ovlivní rozmanitost archivovaných zdrojů. Na druhou stranu však umožní určitou oblast archivovat do hloubky a zaznamená její historický vývoj.

Prioritní kategorie dokumentů by časem měla být rozšířena o další typy zatím téměř nearchivované. Například blogy se stávají čím dál tím víc oblíbeným nástrojem pro komunikaci dokonce i odborných informací. Některé kategorie dokumentů jsou dnes vyřazeny z archivace z důvodu technických překážek. V budoucnu se jistě najdou řešení jak bezpečně archivovat i tyto zdroje.

Žádná z kategorií nemůže být vyčerpávajícím způsobem archivována a každá z nich by vyžadovala vlastní selekční kritéria, aby bylo zcela jasné, co se má sbírat.

Aby všichni zúčastnění na projektu pochopili definovaná kritéria ve stejném smyslu, bude pro ně nutné připravit školení. Ta budou zaměřena zejména na oblast informačních technologií, jelikož pro identifikaci a výběr zdrojů do archivu je třeba mít v této oblasti dobrý přehled.

Kritéria výběru umožní knihovně jednodušší a lepší propagaci projektu, jelikož bude moci deklarovat, na jakou oblast se zaměřuje. Tím by mohlo dojít i k jasnější komunikaci s vydavateli, vědci a dalšími zainteresovanými stranami.

V budoucnu se navíc počítá s účastí dalších organizací, které budou zodpovědné za archivaci zdrojů ve svém sektoru.

5 KRITÉRIA VÝBĚRU ELEKTRONICKÝCH ONLINE ZDROJŮ PRO WEBARCHIV A NÁVRH MOŽNÝCH ŘEŠENÍ

Při tvorbě jakékoli sbírky si musí instituce položit zejména otázku: Pro KOHO tuto sbírku vytváří? Odpověď na tento dotaz je však základním problémem při vytváření archivu digitálních zdrojů, jelikož ji neznáme. Digitální archiv má perspektivní využití v horizontu dvaceti a více let a dnes ještě nikdo není schopen odhadnout, co bude odborníky v budoucnu zajímat ani to, kdo bude archiv využívat.

I přes tento handicap je nutné definovat politiku výběru elektronických online zdrojů do digitálního archivu a opírat se o současné zkušenosti.

Z pětiletého období fungování projektu WebArchiv vyplynulo, že není možné, aby Národní knihovna archivovala veškerý obsah českého webu. Je nutné přistoupit na spolupráci s dalšími subjekty, kteří by byly silnými partnery. Zatím jediným partnerem je české středisko ISSN, slibně se vyvíjí spolupráce v oblasti archivace zdrojů veřejné správy s Národním archivem. Vytipovány máme i další instituce - odborné a krajské knihovny, akademické organizace, specializované archivy. Stanovení odpovědnosti je zásadní otázkou, která může velice pomoci při řešení této problematiky. Záleží ovšem na finančních a personálních možnostech a také na vůli a nadšení dané instituce pro takovou spolupráci.

5.1 Kritéria výběru zdrojů v rámci WebArchivu

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

Tato kritéria však bylo třeba rozpracovat detailněji, bylo definováno 7 základních kategorií, které by vybrané zdroje měly splňovat. Tyto kategorie lze najít i na webových stránkách WebArchivu[22] a řešitelé projektu se jimi dosud při výběru

[22] http://www.webarchiv.cz/kriteria.html

zdrojů řídili. Při srovnání s kritérii zmíněnými v předchozí kapitole zde najdeme mnoho společných znaků, několik nedostatků a potřebu detailněji specifikovat některé příliš široké kategorie.

5.1.1 Protokol

Dosavadní vymezení:

Doporučeny jsou zdroje dostupné prostřednictvím protokolu http, ftp a news. Ve druhém a třetím případě je třeba ověřit, zda vybrané zdroje nejsou publikovány též v protokolu http.

Doporučení:

Zaměřit se pouze na protokol http. Ze zkušeností víme, že vybrané zdroje byly vesměs publikované v rámci protokolu http. Prozatím jsme se nesetkali s dokumentem, který by splňoval daná kritéria, byl zveřejněn přes ftp a neměl shodnou verzi zpřístupněnou přes http. Archivování diskusních skupin publikovaných v rámci protokolu news by mohlo být zajímavé zejména z obsahového, ale i jazykového hlediska, avšak s tímto typem zdroje zatím nastávají problémy v oblasti dlouhodobé archivace.

5.1.2 Uložení

Dosavadní vymezení:

Webové zdroje zpřístupněné na serverech domény prvního řádu .cz a servery dalších domén (např. .org nebo .com), pokud jsou registrovány fyzickou nebo právnickou osobou se sídlem v ČR. Pokud jsou známy, lze archivovat i webové zdroje českých autorů či webové zdroje v češtině, které jsou zpřístupněny na zahraničních serverech.

Doporučení:

Stále platí zaměření zejména na doménu .cz. Vymezení oblasti českého webu je velice obtížné, jeho definování bylo naznačeno v kapitole 4.1.1. Rozšíření publikační produkce v rámci nadnárodních domén je však již znatelné, proto aplikace nástroje pro automatické rozpoznání českého jazyka bude zcela nezbytná.

5.1.3 Původ

Dosavadní vymezení:

Vybírat byly pouze zdroje originálně zpřístupněné na webu.

Doporučení:

Stále se zaměřujeme na zdroje, které jsou dostupné pouze v online podobě. Pokud má zdroj tištěnou verzi, je dostupný na CD či jiných nosičích, dáváme přednost těmto médiím. A to zejména z důvodu, že jejich odevzdávání funguje na principu povinného výtisku. Online verze dokumentů dostupných v několika formách jsou vybrány pouze v případě, že obsahují významné další informace či přidanou hodnotu.

Do budoucna zůstává otevřená otázka dlouhodobé archivace různých typů zdrojů. Pokud by se podařilo prosadit zákon o povinném výtisku zahrnující odevzdávání všech dostupných forem dokumentů včetně online zdrojů a ukázalo by se, že archivace digitálních zdrojů je nejjednodušší a nejefektivnější způsob archivace, pak by stálo za to uvážit, zda nepreferovat online formu.

5.1.4 Přístup

Dosavadní vymezení:

Vybírat pouze volně přístupné zdroje, které lze považovat za samostatné publikační jednotky.

Doporučení:

Zatím byly vybírány pouze volně dostupné zdroje. Není však třeba bránit se i komerčním publikacím, jelikož často obsahují hodnotné informace. A bude-li to možné, pak vyjednat s jejich vydavateli alespoň omezený přístup k těmto dokumentům.

Byl již učiněn i první pokus v této oblasti. Byl zaslán dopis vydavatelství Dashöfer Verlag s žádostí o schůzku, kde by byly vyjasněny všechny záměry WebArchivu. Toto vydavatelství publikuje mnoho volně dostupných i komerčních online zdrojů v doméně .cz. Tento pokus byl ale bohužel zatím neúspěšný.

Jak bylo naznačeno v kapitole 4.4, je potřeba stanovit si strategii vyjednávání s vydavateli, a to včetně komerčních, a uspořádat školení, workshop či alespoň informativní schůzku, kde by byly objasněny záměry Webarchivu a zdůrazněn význam dlouhodobé archivace.

5.1.5 Formát

Dosavadní vymezení:

Požadovány jsou formáty, které jsou interpretovány běžnými webovými prohlížeči bez nutnosti instalace plug-inu. Ve výjimečných případech lze archivovat webové zdroje v proprietárních formátech, např. PDF.

Doporučení:

Je třeba se držet všeobecně podporovaných formátů a vyvarovat se zejména formátů proprietárních, k jejichž interpretaci potřebujeme konkrétní software. Mohlo by totiž dojít k narušení integrity daného digitálního dokumentu při aplikaci metod dlouhodobé archivace. Nástroj pro automatizovaný sběr dat - Heritrix si již s mnoho formáty dokáže bez problémů poradit. Podrobnější poznatky v oblasti harvestingu a formátů přinesl výzkum pracovníků IIPC (*viz příloha č. 2*).

5.1.6 Obsah

Dosavadní vymezení:

Webové zdroje odborného, uměleckého a zpravodajsko-publicistického zaměření a výjimečně webové zdroje administrativního zaměření. Vynechány jsou webové zdroje, které slouží pouze k prezentaci soukromých osob nebo institucí.

Doporučení:

Obsah zdroje úzce souvisí s body 6.2 a 6.7. Je třeba se zaměřit na zdroje obsahující významné informace z oblasti sociálního, politického, kulturního, náboženského, vědeckého nebo ekonomického života. Dokument by měl zároveň splňovat následující podmínky: být v češtině, mít českého autora nebo se obsahově věnovat České republice. Výběr osobních stránek a stránek soukromých institucí bude probíhat dle kritérií vymezených v rámci bodu 5.3.4.

5.1.7 Typ zdroje

Dosavadní vymezení:

Doporučeno je zaměřit se na seriály, monografie, konferenční příspěvky, výzkumné a jiné zprávy, akademické práce aj. O tom, zda bude daný typ webového zdroje či konkrétní webový zdroj zařazen do archivu, bylo rozhodováno individuálně podle toho, zda je významný z obsahového hlediska.

Doporučení:

Pokud se podaří rozšířit personální zabezpečení projektu, mohli by se pracovníci WebArchivu pokusit o výběr následujících typů dokumentů:

Dokumenty veřejné správy

V České republice se mnoho dokumentů veřejné správy přesouvá z papírové podoby na web. Je to jeden z cílů Ministerstva informatiky, které je zodpovědné za naplnění záměru e-Governmentu. Pro dokumenty veřejné správy již byla vyčleněna i doména .gov.cz, což by v praxi znamenalo velké usnadnění práce a možnost automatického stahování těchto dokumentů. Metodické doporučení pro tvorbu doménových jmen však bylo s účinností od 1.1.2005 zrušeno.

Pracovníci WebArchivu se podíleli na přípravě dalšího, neméně závažného metodického doporučení Ministerstva informatiky pro popis elektronických informačních zdrojů. Metodika vychází ze základních 15 prvků Dublin Core, dnes nejrozšířenějšího metadatového formátu pro popis elektronických zdrojů. Doporučení by mělo být aplikováno v blízké budoucnosti a pro WebArchiv by to opět znamenalo

zjednodušení registrace, výběru a harvestingu. Sklizená metadata by mohla posloužit také jako základ bibliografického záznamu zdroje.

Důležitou roli při archivaci online zdrojů hrají formáty. Při schůzce se zástupci Národního archivu se hovořilo nejen o možnosti spolupráce při registraci a stahování dokumentů veřejné správy, ale jelikož mají pracovníci archivu blízko ke správcům stránek jednotlivých ministerstev a dalších orgánů státní správy a samosprávy, přislíbili, že se pokusí ovlivnit výběr používaných formátů na ty s možností dlouhodobé archivace.

Zdroje akademických institucí

V rámci WebArchivu jsou zatím archivovány pouze webové stránky univerzit či projektů, na kterých se akademické instituce podílejí. Dále jsou to materiály z konferencí, články v časopisech či výzkumné zprávy. Nejsou archivovány tak významné informační zdroje, jakými jsou kvalifikační práce nebo e-printy. Je třeba se v oblasti archivace akademických publikací pokusit přesvědčit jednotlivé instituce, aby začaly s archivací vlastních zdrojů. Národní knihovna může plnit zejména roli poradce nebo se aktivně podílet na vytváření takového archivu.

Materiály z konferencí

Konferenční materiály jsou hodnotným informačním zdrojem. Archivace bude probíhat dle navrženého postupu v kapitole 5.3.4. Jsou preferovány velké konference před malými, kdy pořadateli akce jsou profesionální asociace, vládní či akademické organizace. Archivovány budou stránky konferencí obsahující plné texty příspěvků přednesených na konferenci. Za informačně chudé jsou považovány stránky pouze s prezentacemi v Powerpointu. Již existují i softwarové nástroje pro automatické rozpoznání formátů v rámci dokumentu. Bude-li dokument obsahovat převážně soubory s příponou .ppt (Powerpoint), nebude vybrán pro archivaci.

Elektronické časopisy

Elektronickým časopisům je zatím věnována největší pozornost. Je to dáno především spoluprací se střediskem ISSN, a tím pádem jednoduchým zjišťováním představitelů této kategorie. Tyto zdroje jsou také považovány za informačně hodnotné. Je třeba zaměřit se i na komerční vydavatele. Preference byly popsány v kapitole 5.3.4. Pozornost je tedy věnována zejména časopisům s určitou tradicí, časopisům publikovaným renomovanými vydavateli a také těm s dobrou editorskou kontrolou, časopisům věnovaným odborným tématům a časopisům komerčním.

Tematicky zaměřené zdroje

Vytvoření sbírky různých typů dokumentů na určité aktuální téma národního významu je velice atraktivní a zajímavé. V minulosti byly archivovány například stránky s tematikou povodní 2002. V letošním roce se řešitelé projektu rozhodli pro témata: Dalimilova kronika a vládní krize. Za rok to jistě bude téma parlamentních voleb.

Zdroje unikátní z pohledu designu, použitých technologií

Jelikož se domnívám, že budoucí uživatele digitálního archivu by mohly zdroje zajímat nejen z obsahového hlediska, ale také z hlediska použitých technologií či zajímavě designersky vytvořených stránek, bylo by dobré část pozornosti věnovat i této kategorii zdrojů. Východiskem pro výběr by mohly být žebříčky vyhodnocující stránky dle nejlepšího designu.

Samostatnou kapitolou v této oblasti je u nás zatím pomalu se rozvíjející umělecký žánr zvaný Net.Art (Internetové umění). Již delší dobu jsou sledovány a ukládány stránky významné české multimediální umělkyně - Markéty Baňkové[23].

Pozornost bude muset být věnována použitým formátům, jelikož je velká pravděpodobnost, že zajímavé stránky budou obsahovat formáty, se kterými si dosud nedokážeme poradit.

[23] http://www.bankova.cz

5.2 Další možná řešení problematiky výběru zdrojů

5.2.1 Tematické dělení dle oborových bran resp. Konspektu

V rámci Jednotné informační brány (JIB) vyvíjené Národní knihovnou a Ústavem výpočetní techniky Univerzity Karlovy začaly od roku 2004 postupně vznikat jednotlivé oborové brány. Oborová informační brána je definována jako nástroj zajišťující přístup ke sbírce kvalitních, odborně vybraných a zhodnocených informačních zdrojů pro podporu výzkumu v určitém oboru. Umožňuje systematickou práci s informačními zdroji na základě výběru kvalitních zdrojů, vysoké úrovně jejich popisu podle mezinárodních standardů a prověřování jejich dostupnosti [76]. Je plánováno vytvoření 24 oborových bran. Tematické rozdělení vzniklo na základě předmětové kategorizace informačních zdrojů dle metody Konspektu.

Obr. 6: Tematické rozdělení oborových bran dle metody Konspektu

Oborové brány mají v plánu zahrnout v relativní úplnosti domácí české aktuální zdroje informací oboru zveřejněného na Internetu a zároveň výběrově zahrnout i nejvýznamnější zahraniční informační zdroje. Každou oborovou bránu

bude spravovat garant – organizace, která se danému tématu věnuje do hloubky a která je schopna zajistit konzistenci a koherenci své sbírky.

Zatím se připravují na provoz 2 z uvedených 24 oborových bran. Je to oborová brána pro oblast Hudby, jejímž garantem je Hudební oddělení NK ČR, a Knihovnictví a informatika, o jejíž tvorbu a správu se podělí Knihovnický institut NK ČR a Ústav informačních studií a knihovnictví FF UK.

Jelikož jde o výběr zejména českých online zdrojů z určitého oboru, kdy zdroje již prošly určitou kontrolou je pravděpodobné, že budou mít vysokou informační hodnotu. Pokud budou splňovat i další kritéria WebArchivu, pak mohou být zařazeny do sbírky. Předpokládá se, že počet oborových bran bude narůstat. Tematické dělení Konspektu pokrývá celé universum, podle této metody by bylo možné zdroje dělit do uvedených kategorií a webové rozhraní pro zpřístupnění zdrojů by pak mohlo mít podobu australského projektu PANDORA.

5.2.2 Projekt EU - Culture 2000

V letošním roce se podařilo řešitelům projektu WebArchiv získat grant z evropského programu Culture 2000. Jedním ze tří hlavních témat v roce 2005 byla i záchrana kulturního dědictví v evropském rozsahu. Podmínkou získání grantu byla účast nejméně tří zemí a finanční spoluúčast. Rozpočet projektu je 100.000 euro. Projekt Národní knihovny jakožto hlavního organizátora a Slovinska, Estonska a Slovenska jako spoluorganizátorů nese název Web Cultural Heritage a bude probíhat od října letošního roku do září roku 2006.

Hlavními aktivitami projektu jsou:

- Analýza stávajících kritérií výběru elektronických online zdrojů v severozápadní Evropě, Severní Americe, Austrálii, na Novém Zélandu a Asii (zejména v Japonsku a Číně)
- Zmapování a analýza dostupných softwarových nástrojů a postupů při harvestingu a indexaci internetových zdrojů
- Návrh nových kritérií v návaznosti na provedené analýzy
- Adaptace a konfigurace vybraných softwarových nástrojů

- Testování navržených selekčních kritérií prostřednictvím vybraných softwarových nástrojů
- Workshopy a setkání všech spoluorganizátorů projektu
- Přednesení výsledků projektu na konferenci European Conference on Research and Advanced Technology for Digital Libraries (ECDL) 2006

Hlavním cílem projektu je pak stanovení politiky výběru zdrojů do digitálního archivu na národní úrovni.

6 ZÁVĚR

Projekty archivace webu se snaží nabídnout budoucím generacím reprezentativní vzorek kulturní produkce určitého období v historii Internetu. Tedy nejen nejvýznamnější práce jako vědecké články či studie, ale také celou řadu dalších materiálů z oblasti kultury, sociologie, politiky, techniky a dalších.

Posláním národních knihoven je archivovat širokou škálu materiálů zejména z důvodu, že dnes nelze odhadnout, co bude zajímat budoucí vědce či jiné zájemce o jejich studium. To samé platí pro web. Avšak u webových publikací je úplnost, která je vyžadována u tradičních dokumentů, nedosažitelná. U tištěných publikací se vydavatelé snaží obsah filtrovat a takto zúžený rozsah publikací jsou knihovny schopny shromažďovat. Oproti tomu na webu je publikování možné téměř zdarma a mnoho osob a subjektů této příležitosti beze zbytku využívá. Denní nárůst nových dokumentů na webu je obrovský a kontrola nad jejich produkcí téměř neexistuje. I zde se tak otázka výběru dokumentů stává velice aktuální.

Proces archivace webu je poměrně složitá problematika. V rámci jejich dílčích procesů (registrace, sběr, dlouhodobé uchování a zpřístupnění online dokumentů) je třeba intenzivně řešit problémy nejen z oblasti informačních technologií, ale i otázky knihovnické či legislativní. Instituce zabývající se archivací webu na národní úrovni přistoupily k výběru a sběru dokumentů různými způsoby.

Všechny popsané metody archivace webu mají své výhody a nevýhody. Je zajímavé, že mnoho národních knihoven po několika letech archivace a užívání jimi zvolené metody nyní hodnotí dosažené výsledky a často vidí její nedostatky. Žádná ze současných metod archivace není ideální nebo alespoň taková, abychom ji mohli prohlásit za jednoznačně lepší než ostatní.

V rámci plošné metody archivace je pozornost nyní soustředěna na další vývoj softwarového nástroje pro sklízení dokumentů – Heritrixu, vyvíjeného konsorciem IIPC a rovněž na další aktivity této významné mezinárodní organizace. Cílem je, aby se harvesting, stejně jako dlouhodobá archivace online dokumentů, stal rutinní agendou v rámci provozu národních knihoven.

Při použití výběrové metody archivace jsou nejdůležitějším aspektem správně definovaná kritéria výběru online zdrojů do digitálního archivu. Zemím, které se vydaly touto cestou, mají dobře propracovaná a detailně definovaná kritéria výběru, se daří budovat konzistentní a atraktivní sbírku vybraných online zdrojů národního významu.

Pak je tu i kompromisní řešení, které kombinuje obě zmíněné základní metody. Díky automatizovanému sběru vytváří co nejúplnější časové snímky celé oblasti národního webu a zároveň pravidelně doplňuje archiv prezentující vybranou skupinu významných zdrojů.

Nová technická řešení se daří nalézat poměrně rychle, horší je situace v oblasti legislativy. Mnoho z nadějně se rozvíjejících projektů zatím nemá oprávnění jakýmkoli způsobem vytvořený archiv zpřístupnit, jelikož jim v tom brání autorský zákon.

Web je velmi proměnlivé médium a jeho vývoj v budoucnosti lze jen těžko odhadnout. Možná, že se již brzy vyplní vize sémantického webu, kdy budou informace prezentovány tak, že jim stroje budou rozumět. Idea sémantického webu vychází z potřeby dát obsahu webu jasný smysl a učinit zde dostupná data srozumitelná strojům. Konsorcium W3C (World Wide Web Consortium) stojí v čele snahy vyvinout pro sémantický web potřebné standardy. K těm patří zejména RDF (Resource Description Framework), standard používaný pro vyjádření sémantiky nebo XML, který umožňuje vytvářet strukturu. Velmi obtížným aspektem budování sémantického webu je ovšem vytvoření vhodných ontologií, zjednodušeně řečeno obecných slovníků, které budou moci systémy využít k rozpoznání obsahu webového dokumentu [77]. Zejména z těchto důvodů zůstává sémantický web prozatím nenaplněným ideálem a hudbou budoucnosti.

6.1 Závěrečné doporučení pro WebArchiv

Pro úspěšné fungování projektu WebArchiv je v první řadě nutné zajistit takové personální a finanční zabezpečení projektu, aby bylo možné přejít postupně z fáze testování do praktického a rutinního provozu.

Řešitelům WebArchivu se již podařilo překonat mnoho těžkých překážek a zbývá ta nejtěžší a nejdůležitější – zpřístupnit archivovaná data veřejnosti. Tím dojde k naplnění smyslu všech dosud realizovaných činností a projekt snad dosáhne prvních uznání.

O významu projektu není pochyb. Žijeme v generaci některými označované za „press delete“ a byla by škoda přijít o tak významnou část národního kulturního dědictví.

SEZNAM POUŽITÉ LITERATURY

1. ABITEBOUL, S., COBÈNA, G., MASANÈS, J. A first experience in archiving the French Web. In *Research and advanced technology for digital libraries: 6th European conference, ECDL 2002, Rome, Italy, September 16–18, 2002* [online]. Berlin: Springer, 2002 [cit. 2005-05-15]. 15 s. Dostupný na WWW: <ftp://ftp.inria.fr/INRIA/Projects/verso/gemo/GemoReport-229.pdf>.

2. *AOLA: Austrian On-Line Archive* [online]. Vienna: Technical University, 2001 [cit. 2005-04-25]. Dostupné na WWW: <http://www.ifs.tuwien.ac.at/~aola/>.

3. *Archivserver DEPOSIT.DDB.DE* [online]. Die Deutsche Bibliothek, last updated 4.11 2004 [cit. 2005-04-25]. Dostupné na WWW: <http://deposit.ddb.de/>.

4. ARMS, W. *Web Preservation Project: final report* [online]. Library of Congress, September 2001 [cit. 2005-05-19]. Dostupné naWWW: <http://www.loc.gov/minerva/webpresf.pdf>.

5. ARVIDSON, A. The collection of Swedish Web pages at the Royal Library: the Web heritage of Sweden. In *68th IFLA Council and General Conference, Glasgow, UK, 18-24 August 2002* [online]. [cit. 2005-02-25]. 3 s. Dostupný na WW: <http://www.ifla.org/IV/ifla68/papers/111-163e.pdf>.

6. ARVIDSON, A., PERSSON, K. Harvesting the Swedish Web space. In *What´s next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001* [online]. [cit. 2005-02-25]. Dostupný na WW: <http://bibnum.bnf.fr/ecdl/2001/sweden/sld001.htm>.

7. ARVIDSON, A., PERSSON, K., MANNERHEIM, J. The Kulturarw3 project - the Royal Swedish Web Archiw3e: an Example of „Complete“ Collection of Web Pages. In *66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 August 2000* [online]. [cit. 2005-02-26]. Dostupný na WWW: <http://www.ifla.org/IV/ifla66/papers/154-157e.htm>.

8. *Austrian On-Line Archive. Web Archiving Bibliography* [online]. Austrian National Library, last updated April 2004 [cit. 2005-04-10]. Dostupné na WWW: <http://www.ifs.tuwien.ac.at/~aola/links/WebArchiving.html>.

9. BARTOŠEK, M. Aktuální oblasti výzkumu digitálních knihoven. In *INFOS 2003: medzinárodné informatické sympózium, Stará Lesná, 7. – 10. apríla 2003* [online]. [cit. 2005-04-05]. Dostupný na WWW: < http://www.aib.sk/infos/infos2003/18.htm>.

10. BARTOŠEK, M. *Výzkumný záměr: digitální knihovny* [online]. Brno: Masarykova univerzita, last updated 5.9.2005 [cit. 2005-09-05]. Dostupné na WWW: <http://wwwdata.muni.cz/to.cs/research/cez_item.asp?ID=23>.

11. BOHÁČEK, Martin. Autorské právo a elektronický obchod po vstupu ČR do ES z hlediska knihoven. In *Inforum 2003: 9. ročník konference o profesionálních informačních zdrojích* [online]. Praha : Albertina icome, 2003 [cit. 2005-03-10]. Dostupné na WWW: <http://www.inforum.cz/inforum2003/prispevek.asp?CisloSekce=20&Kod=123>.

12. BRATKOVÁ, E. K otázkám pojmu, třídění a typologie internetových a webovských informačních zdrojů. *Národní knihovna: knihovnická revue*. 1998, roč. 9, č. 5, s. 238-262.

13. BRATKOVÁ, E. *Vyhledávání informací z digitálních virtuálních knihoven: studijní materiál*. Praha: Ústav informačních studií a knihovnictví FF UK, leden 2002. 39 s.

14. BRYGFJELD, S. A. Access to Web Archives: the Nordic Web Archive Access Project. In *68th IFLA Council and General Conference, Glasgow, UK, 18-24 August 2002* [online]. [cit. 2005-02-25]. 4 s. Dostupný na WW: <http://www.ifla.org/IV/ifla68/papers/090-163e.pdf>.

15. CATHRO, W., WEBB, C., WHITING, J. Archiving the Web: the PANDORA Archive at the National Library of Australia. In *Preserving the Present for the future Web Archiving Conference, Copenhagen, 18-19 June 2001* [online]. [cit. 2005-03-04]. Dostupný na WWW: <http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html>.

16. CELBOVÁ, L. Bibliografické standardy pro popis elektronických zdrojů. In *Moderní informační a komunikační technologie v knihovnictví, STK, 13.11.2001* [online]. [cit. 2005-01-09]. Dostupný na WWW: <http://webarchiv.nkp.cz/stk2001_lc.zip>.

17. CELBOVÁ, L. Mezinárodní konference se zabývala otázkou, zda archivovat web komplexně, nebo výběrově. *Ikaros* [online]. 2005, roč. 9, č. 1 [cit. 2005-01-09]. Dostupný na WWW: <http://www.ikaros.cz/Clanek.asp?ID=200501007>.

18. CELBOVÁ, L. Politika výběru elektronických dokumentů publikovaných v prostředí Internetu pro účely ČNB. *Ikaros* [online]. 2001, roč. 5, č. 8 [cit. 2005-01-16]. Dostupný na WWW: <http://www.ikaros.cz/Clanek.asp?ID=200208342>.

19. CELBOVÁ, L. *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet : souhrnná zpráva za rok 2000* [online]. Praha : Národní knihovna ČR, 2000 [cit. 2005-07-06]. Dostupné na WWW: <http://webarchiv.nkp.cz/zprava2000.pdf>.

20. CELBOVÁ, L. *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet : závěrečná zpráva za léta 2000-2001* [online]. Praha : Národní knihovna ČR, leden 2002 [cit. 2005-07-06]. Dostupné na WWW: <http://webarchiv.nkp.cz/zprava2001/zprava2001.pdf>.

21. CELBOVÁ, L. Stanou se online dostupné elektronické zdroje integrovanou součástí digitálních knihoven? *Národní knihovna: knihovnická revue*. 2001, roč. 12, č. 2, s. 91-98. Dostupný též na WWW: <http://webarchiv.nkp.cz/nk2_2001.pdf>.

22. CELBOVÁ, L. *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. 2005-07-06]. Dostupné na WWW: <http://webarchiv.nkp.cz/zprava2002/zprava2002.pdf>.

23. 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 WWW: <http://knihovny.cvut.cz/akp2003/index.htm>.

24. CELBOVÁ, L., SIMONOVÁ, M., ŽABIČKA, P. WebArchiv – od výzkumu k tvrdé realitě. In *Knihovny současnosti 2003* [online]. Brno : Sdružení knihoven ČR, 2003 [cit. 2005-07-06]. Dostupné na WWW: <http://www.webarchiv.cz/sec2003.pdf>.

25. CELBOVÁ, L., ŽABIČKA, P. 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>.

26. CELBOVÁ, L., ŽABIČKA, P. 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>.

27. CLAUSEN, L. R. *Handling file formats*. [online]. Copenhagen (Dánsko): The Royal Library, May 2004 [cit. 2005-04-19]. Dostupné na WWW: <http://netarchive.dk/website/publications/FileFormats-2004.pdf >.

28. COCH, J., MASANÈS, J. Language engineering techniques for web archiving. In *4th International Web Archiving Workshop (IWAW'04*) *Bath, UK, 16 September 2004* [online]. [cit. 2005-04-19]. Dostupné na WWW: <http://www.iwaw.net/04/index.html >.

29. Conference of European National Librarians, Federation of European Publishers (CENL/FEP). Mezinárodní deklarace k odevzdávání elektronických dokumentů do konzervačního fondu. *Národní knihovna: knihovnická revue* [online]. 2001, roč. 12, č. 4, s. 289 – 293. [cit. 2005-01-25] Dostupný na WWW: <http://full.nkp.cz/nkkr/NKKR0104/0104289.html>.

30. *DACHS* [online]. Heidelberg (Německo): University of Heidelberg, last updated 1 September 2005 [cit. 2005-09-01]. Dostupné na WWW: <http://www.sino.uni-heidelberg.de/dachs/>.

31. DAY, M. *Collecting and preserving the World Wide Web: a feasibility study undertaken for the JISC and Wellcome Trust* [online]. 25 February 2003 [cit. 2005-03-19]. 91 s. Dostupné na WWW: <http://library.wellcome.ac.uk/projects/archiving_feasibility.pdf>.

32. *Digital Preservation* [online]. Washington: Library of Congress, 2003 [cit. 2005-03-17]. Dostupné na WWW: <http://www.digitalpreservation.gov>.

33. *E-Depot* [online]. Amsterdam (Holandsko) : National Library of the Netherlands [cit. 2005-03-17]. Dostupné na WWW: <http://www.kb.nl/dnp/e-depot/dm/dm-en.html>.

34. *Electronic Collection* [online]. Ottawa (Kanada): Library and Archive Canada, last updated 2005-07-24 [cit. 2005-08-05]. Dostupné na World Wide Web: <http://www.collectionscanada.ca/electroniccollection/003008-200-e.html>.

35. GATENBY, P. Archiving the web: The national collection of Australan online publications. In *International Symposium of web archiving, National Diet Library, Tokyo, Japan, 30 January 2002* [online]. Canberra (Austrálie): National Library of Australia, 2002 [cit. 2005-06-19]. Dostupné na WWW: <http://www.nla.gov.au/nla/staffpaper/2002/phillips1.html>.

36. GATENBY, P. Collecting and Managing Web Resources for Long-Term Access : Web Harvesting and Guidelines to Support Preservation. In *70th IFLA General Conference and Council, 22 - 27 August 2004, Buenos Aires, Argentina* [online]. [cit. 2005-02-25]. Dostupné na WWW: <http://www.ifla.org/IV/ifla70/papers/026e-Gatenby.pdf>.

37. *Guidelines for Selecting, Archiving and Preserving Websites pertinent to Tasmanian Government Information and Cultural Heritage* [online]. State Library of Tasmania, last modified 1 September 2005 [cit. 2005-09-01]. Dostupné na WWW: <http://odi.statelibrary.tas.gov.au/About/selpolicy.asp>.

38. *Guidelines for the Preservation of Digital Heritage* [online]. UNESCO, March 2003 [cit. 2005-09-01]. Dostupné na WWW: <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>.

39. HAKALA, J. Archiving the web: european experiences. *Tietolinja* [online]. 2003, č. 2 [cit. 2005-01-09]. Dostupný na WWW: <http://www.lib.helsinki.fi/tietolinja/0203/Webarchive.html>.

40. HAKALA, J. Collecting and Preserving the Web: developing and testing the NEDLIB Harvester. *RLG DigiNews* [online]. 2001, vol.5, no. 2 [cit. 2005-04-19]. Dostupné na WWW: <http://www.rlg.org/preserv/diginews/diginews5-2html#feature2>.

41. HAKALA, J. Harvesting the Finnish Web space - practical experiences. In *What´s next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001* [online]. [cit. 2005-05-19]. Dostupné na WWW: <http://www.bnf.fr/pages/infopro/ecdl/finland/sld001.htm>.

42. HARVEY, R. Now you see it, now you don´t: maintaining digital learning objects for the future. *E-JIST* [online]. 2000, vol. 5, no. 2 [cit. 2005-01-16]. Dostupný na WWW: <http://www.usq.edu.au/electpub/e-jist/docs/Vol5_No2/harvey.html>.

43. HENRIKSEN, B. N. The Danish project netarchive.dk [online]. In *2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002* [online]. 2002 [cit. 2005-05-19]. Dostupné na WWW: <http://bibnum.bnf.fr/ecdl/2002/>.

44. CHARLESWORTH, A. *Legal issues relatin to the archiving of Internet resource of the UK, EU, USA and Australia: a study undertaken for the JISC and wellcome trust* [online]. 25 Feruary 2003 [cit. 2005-04-19]. 79 s. Dostupné na WWW: <http://library.wellcome.ac.uk/projects/archiving_legal.pdf>.

45. CHRISTENSEN-DALSGAARD, B. WebArchive Activities in Denmark. *RLG DigiNews* [online]. 2004, vol.8, no. 3 [cit. 2005-04-19]. Dostupné na WWW: <http://www.rlg.org/en/page.php?Page_ID=17661>.

46. *Internet Archive* [online]. Alexa Internet, last updated 10 March 2001 [cit. 2005-03-17]. Dostupné na WWW: <http://www.archive.org/>.

47. JODELIS, R. Harvesting and archiving Electronic Resources in Lithuania: towards Virtual Library. In *Inforum 2003* [online] Praha: Albertina income Praha, 2003 [cit. 2005-04-19]. Dostupné na WWW: <http://www.inforum.cz/inforum2003/prospevky/Jodelis_Remigijus.pdf>.

48. KAVČIČ-ČOLIC, A. Archiving the Web – some legal aspects. In *68 th IFLA Council and General Conference, August 18-24, 2002* [online]. Glasgow (Scotland), 2002 [cit. 2005-04-19]. 7 s. Dostupné na WWW: <http://www.ifla.org/IV/ifla68/papers/116-163e.pdf>.

49. KAVČIČ-ČOLIČ, A., GROBELNIK, M. Archiving the Slovenian Web: Recent Experiences. In *4th International Web Archiving Workshop (IWAW'04*) *Bath, UK, 16 September 2004* [online]. [cit. 2005-04-19]. Dostupné na WWW: <http://www.iwaw.net/04/index.html >.

50. KENNEY, A. R. et al. Preservation Risk Management for Web Resources. *D-Lib Magazine* [online]. 2002, vol.8 no. 1 [cit. 2005-04-19]. Dostupné na WWW: <http://www.dlib.org/dlib/january02/kenney/01kenney.html>.

51. KOERBIN, P. The PANDORA Digital Archiving System (PANDAS) and Managing Web Archiving in Australia: a case study. In *4th International Web Archiving Workshop (IWAW'04*) *Bath, UK, 16 September 2004* [online]. [cit. 2005-04-19]. Dostupné na WWW: <http://www.iwaw.net/04/index.html >.

52. *Kulturarw3* [online]. Stockholm (Švédsko): The Royal Library, last updated March 1, 2005 [cit. 2005-04-05]. Dostupné na WWW: <http://www.kb.se/kw3/ENG/>.

53. LAMPOS, C. et al. Archiving the Greek Web. In *4th International Web Archiving Workshop (IWAW'04*) *Bath, UK, 16 September 2004* [online]. [cit. 2005-04-19]. Dostupné na WWW: <http://www.iwaw.net/04/index.html >.

54. MACH, P. et al. *Dlouhodobé uchovávání a zpřístupňování dokumentů v digitální podobě s celostátní působností.* Praha: České vysoké učení technické, 2001, 53 s.

55. MACH, P. et al. *Dlouhodobé uchovávání a zpřístupňování dokumentů v digitální podobě s celostátní působností: studie.* Praha: České vysoké učení technické, 2002, 91 s.

56. MANNERHEIM, J. The WWW and our digital heritage – the new preservation tasks of the library comunity. In *66th IFLA Council and General Conference, Jerusalem, Israel, 13-18 August 2000* [online]. [cit. 2005-02-25]. Dostupný na WW: < http://ifla.org/IV/ifla66/papers/158-157e.htm>.

57. MARILL, J. et al. *Web Harvesting Survey* [online]. International Internet Preservation Consortium, July 2004 [cit. 2005-04-19]. Dostupné na WWW: <http://www.netpreserve.org/publications/iipc-r-001.pdf>.

58. MASANÈS, J. Archiving the Deep Web. In *2nd ECDL Workshop on Web Archiving, Rome, Italy, 19 September 2002* [online]. [cit. 2005-04-19]. Dostupné na WWW: <http://bibnum.bnf.fr/ecdl/2002/>.

59. MASANÈS, J. The BnF´s project for Web archiving. In *What´s next for digital deposit libraries? ECDL Workshop, Darmstadt, Germany, 8 September 2001*, [online]. [cit. 2005-04-19]. Dostupné na WWW: <http://www.bnf.fr/pages/infopro/ecdl/france/slg001.htm>.

60. *Minerva: Mapping the Internet Electronic Resource Virtual Archive* [online]. Washington : The Library of Congress [cit. 2005-04-17]. Dostupné na WWW: <http://www.loc.gov/minerva/>.

61. MOHR, G., KIMPTON, M., STACK. Introduction to Heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW'04*) *Bath, UK, 16 September 2004* [online]. [cit. 2005-04-19]. Dostupné na WWW: <http://www.iwaw.net/04/index.html >.

62. MURRAY, B. H., MOORE, A. *Sizing the Internet* [online]. A white paper, Cyveillance, July 2000. [cit. 2004-11-30]. Dostupný na WWW: <http://www.cyveillance.com/web/downloads/Sizing_the_Internet.pdf>.

63. *New Decree for Kulturarw3* [online]. Stockholm (Švédsko): Royal Library, updated June 10, 2002 [cit. 2005-04-19]. Dostupné na WWW: <http://www.kb.se/Info/Pressmed/Arkiv/2002/020605_eng.htm>.

64. *NWA* [online]. Oslo (Norsko): National Library of Norway, updated June 10, 2005 [cit. 2005-07-19]. Dostupné na WWW: <http://nwa.nb.no/>.

65. *Online Australan Publications: selection guidelines for archiving and preservation by the National Library of Australia* [online]. Canberra (Austrálie): National Library of Australia, last updated 17 July 2005 [cit. 2003-08-19]. Dostupné na WWW: <http://pandora.nla.gov.au/selectionguidelines.html>.

66. *Our digital Island* [online]. State Library of Tasmania, last updated 21 June 2004 [cit. 2005-04-05]. Dostupné na WWW: <http://odi.statelibrary.tas.gov.au/>.

67. PADI: *Preserving Access to Digital Information* [online]. Canberra (Austrálie): National Library of Australia, last updated March 2005 [cit. 2005-04-19]. Dostupné na WWW: <http://www.nla.gov.au/padi/topics/65.html>.

68. *PANDAS Manual* [online]. Canberra (Austrálie): National Library of Australia, last updated 15 January 2004 [cit. 2005-04-19]. Dostupné na WWW: <http://pandora.nla.gov.au/manual/pandas/index.html>.

69. *PANDORA Project : Preserving and Accessing Networked Documentary Resources of Australia* [online]. Canberra (Austrálie) : National Library of Australia, last updated 6 March 2005 [cit. 2005-04-05]. Dostupné na WWW: <http://pandora.nla.gov.au/index.html>.

70. PHILLIPS, M. Selective Archiving of Web Resources: a Study of Acquisition Costs at the National Library of Australia. *RLG DigiNews* [online]. 2005, vol. 9. no. 3 [cit. 2005-09-5]. Dostupný na WWW: <http://www.rlg.org/en/page.php?Page_ID=20666#article0>.

71. PHILLIPS, M. *Collecting Australian Online Publications* [online]. Canberra (Austrálie) : National Library of Australia, May 2003 [cit. 2005-04-05]. Dostupné na WWW: <http://pandora.nla.gov.au/bsc49.doc>.

72. *Preservation Metadata and the OAIS Information Model: A Metadata Framework to Support the Preservation of digital Objects* [online]. OCLC/RLG Working Group on Preservation Metadata, 2002 [cit. 2005-04-20]. Dostupné na WWW: <http://www.oclc.org/research/projects/pmwg/pm_framework.pdf >.

73. *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet : souhrnná zpráva za rok 2000* [online]. Praha : Národní knihovna ČR, 2000 [cit. 2003-03-17]. Dostupné na World Wide Web: <http://webarchiv.cz/dokumenty.html>.

74. SIMONOVÁ, M., KAČÍRKOVÁ, P. *Návrh kritérií výběru elektronických zdrojů pro oborové informační brány*. Praha: Národní knihovna ČR, 2003. 10 s.

75. *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. 2005-07-10]. Dostupné na WWW: <http://www.nkp.cz/o_knihovnach/Dir01_29_ECcz.pdf>.

76. STOKLASOVÁ, B. Oborové informační brány. In *Knihovny současnosti 2003* [online]. Brno : Sdružení knihoven ČR, 2003 [cit. 2005-07-06]. Dostupné na WWW: <http://jib-info.cuni.cz/dokumenty/sec2003/sec2003_bst.ppt>.

77. THIBODEAU, P. Sémantický web v kostce. *Science World* [online]. 1. 10. 2003 [cit. 2005-01-09]. Dostupný na WWW: <http://www.scienceworld.cz/sw.nsf/0/D08DFE698E30CA80C1256E970048FCEE?OpenDocument&cast=1>.

78. *UK Web Archiving Consortium* [online]. [cit. 2005-03-17]. Dostupné na WWW: <http://www.webarchive.org.uk/>.

79. VAN NUYS, C. The Paradigma Project. *RLG DigiNews* [online]. 2003, vol. 7, no.2 [cit. 2005-04-19]. Dostupné na WWW: <http://www.rlg.org/preserv/diginews/diginews7-2.html>.

80. VAN NUYS, C. et. al. The Paradigma Project and its Quest for Metadata Solutions and User Services. In *70th IFLA General Conference and Council, 22 - 27 August 2004, Buenos Aires, Argentina* [online]. [cit. 2005-02-25]. Dostupné na WW: <http://www.ifla.org/IV/ifla70/papers/009e-Nuys.pdf>.

81. *Virtual Remote Control* [online]. New York: Cornell University [cit. 2005-04-19]. Dostupné na WWW: <http://irisresearch.library.cornell.edu/VRC/>.

82. VOJTÁŠEK, F. Archivace a zpřístupnění elektronických dokumentů se zaměřením na webové zdroje. In *Knihovny současnosti 2001: sborník z 9. konference, konané ve dnech 11. – 13. září 2001 v Seči u Chrudimi.* Brno : Sdružení knihoven ČR, 2001, s. 136-147. ISBN 80-86249-14-X.

83. VOJTÁŠEK, F. Archivace tradičních a elektronických dokumentů: stejný cíl, různé metody. In *Inforum 2001* [online]. Praha: Albertina icome Praha, 2001 [cit. 2005-04-19]. Dostupné na WWW: <http://www.inforum.cz/informu2001/prispevky/vojtasek.htm>.

84. VOLMUTHOVÁ, Z. *Povinný výtisk online publikací a jeho legislativní zajištění*, Praha, 2003. Diplomová práce. Univerzita Karlova v Praze, Filozofická fakulta, Ústav informačních studií a knihovnictví 2003. Vedoucí práce PhDr. Eva Bratková.

85. *WebArchiv* [online]. Praha : Národní knihovna ČR, posl. aktual. 24. října 2003 [cit. 2005-07-06]. Dostupné na World Wide Web: <http://www.webarchiv.cz>.

86. YAN, H., HUANG, L., CHEN, C. A New Data Storage and Service Model of China Web InfoMall. In *4th International Web Archiving Workshop (IWAW'04*) *Bath, UK, 16 September 2004* [online]. [cit. 2005-04-19]. Dostupné na WWW: <http://www.iwaw.net/04/index.html >.

87. *Zákon č. 121/2000 Sb.* o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon). *Sbírka zákonů 2000, částka 36, str. 1658 (2000).*

88. *Zákon č. 37/1995 Sb.* o neperiodických publikacích. *Sbírka zákonů 1995, částka 8, str. 459 (1995).*

89. *Zákon č. 46/2000 Sb.* o právech a povinnostech při vydávání periodického tisku a o změně některých dalších zákonů (tiskový zákon). *Sbírka zákonů 2000, částka 17, str. 586 (2000).*

90. ŽABIČKA, P. Archiv českého webu v roce 3. *Národní knihovna: knihovnická revue*. 2002, roč. 13, č. 3, s. 168-176. ISSN 1214-0678. Dostupné též na WWW: <http://webarchiv.nkp.cz/nk2002.pdf>.

91. ŽABIČKA, P. Infrastruktura Webarchivu v roce 2002. In *Inforum 2002* [online]. Praha: Albertina icome Praha, 2002 [cit. 2005-07-06]. Dostupné na WWW: <http://www.inforum.cz/inforum2002/prednaska8.htm>.

92. ŽABIČKA, P. Konference ECDL 2002. *Ikaros* [online]. 2002, roč. 6, č. 10 [cit. 2005-07-06]. Dostupné na WWW: <http://www.ikaros.cz/Clanek.asp?ID=200209068>.

93. Žabička, Petr. NEDLIB Harvester - technika "sklizně" informací. *Ikaros* [online]. 2000, roč. 4, č. 10 [cit. 2005-06-15]. Dostupné na WWW: <http://ikaros.ff.cuni.cz/20 00/c10/harvest.htm>.

94. ŽABIČKA, P. Webarchiv – digitální knihovna českého webu. In *RUFIS 2002*. Brno : ApS Brno, 2002, s. 121-129. ISBN 80-86510-40-9. Dostupné též na WWW: <http://webarchiv.nkp.cz/rufis2002_pz.pdf>.

PŘÍLOHY

Příloha č. 1: Přehled nejvýznamnějších projektů současnosti

Příloha č. 2: Zpráva IIPC - Web Harvesting Survey

Příloha č. 3: Ukázka vyhledávače NWA Toolset

Příloha č. 1

6.2 Přehled nejvýznamnějších národních projektů současnosti

Země	Název projektu/Nositel projektu	Metoda[1]	Přístup[2]
Austrálie	**PANDORA / National Library of Australia**	**V**	**A**
Austrálie (Tasmánie)	**Our Digital Island / State Library of Tasmania**	**V**	**A**
Česká republika	**WebArchiv / Národní knihovna ČR + Moravská zemská knihovna**	**V + H**	**N**
Čína	**Web InfoMall / Peking University**	**H**	**A**
Dánsko	**Netarchive.dk / Kongelige Bibliotek, Kodaň + Statsbibliotek, Århus**	**V + H**	**O**
Estonsko	**Erik@ / Eesti Rahvusraamatukogu**	**V**	**N**
Finsko	**EVA / Helsinki University Library**	**H**	**N**
Francie	**Bibliothéque nationale de France**	**V + H**	**N**
Island	**Landsbókasafn Íslands Háskólabókasafn**	**V**	**N**
Japonsko	**WARP / National Diet Library**	**V**	**A**
Kanada	**EPPP / National Library of Canada**	**V**	**A**
Litva	**LIBIS Centre of Lithuanian National Library**	**H**	**O**
Lotyšsko	**National Library of Latvia**	**H**	**N**
Německo	**Archiveserver Deposit.ddb.de/Die Deutsche Bibliothek**	**V**	**O**
Nizozemí	**e-Depot / Koninklijke Bibliotheek (Národní knihovna Nizozemí)**	**V**	**N**
Norsko	**PARADIGMA/Nasjonalbiblioteket (Národní knihovna Norska)**	**H +V**	**N**

Nový Zéland	**National Library of New Zealand**	**H + V**	**N**
Rakousko	**AOLA / Österreichischen Nationalbibliothek + Technische Universität Wien**	**H**	**N**
Řecko	**Athens University of Economics and Business**	**H**	**N**
Slovinsko	**Narodna in univerzitetna knjižnica, Ljubljana**	**H**	**N**
Švédsko	**Kulturarw[3] / KB**	**H**	**O**
Švýcarsko	**e-Helvetica / Schweizerischen Landesbibliothek**	**V**	**N**
USA	**Internet Archive / Alexa Internet**	**H**	**A**
USA	**MINERVA / LoC**	**V**	**A**
Velká Británie	**UKWAC / British Library (organizátor)**	**V**	**A**

Data platná k 20. 9. 2005

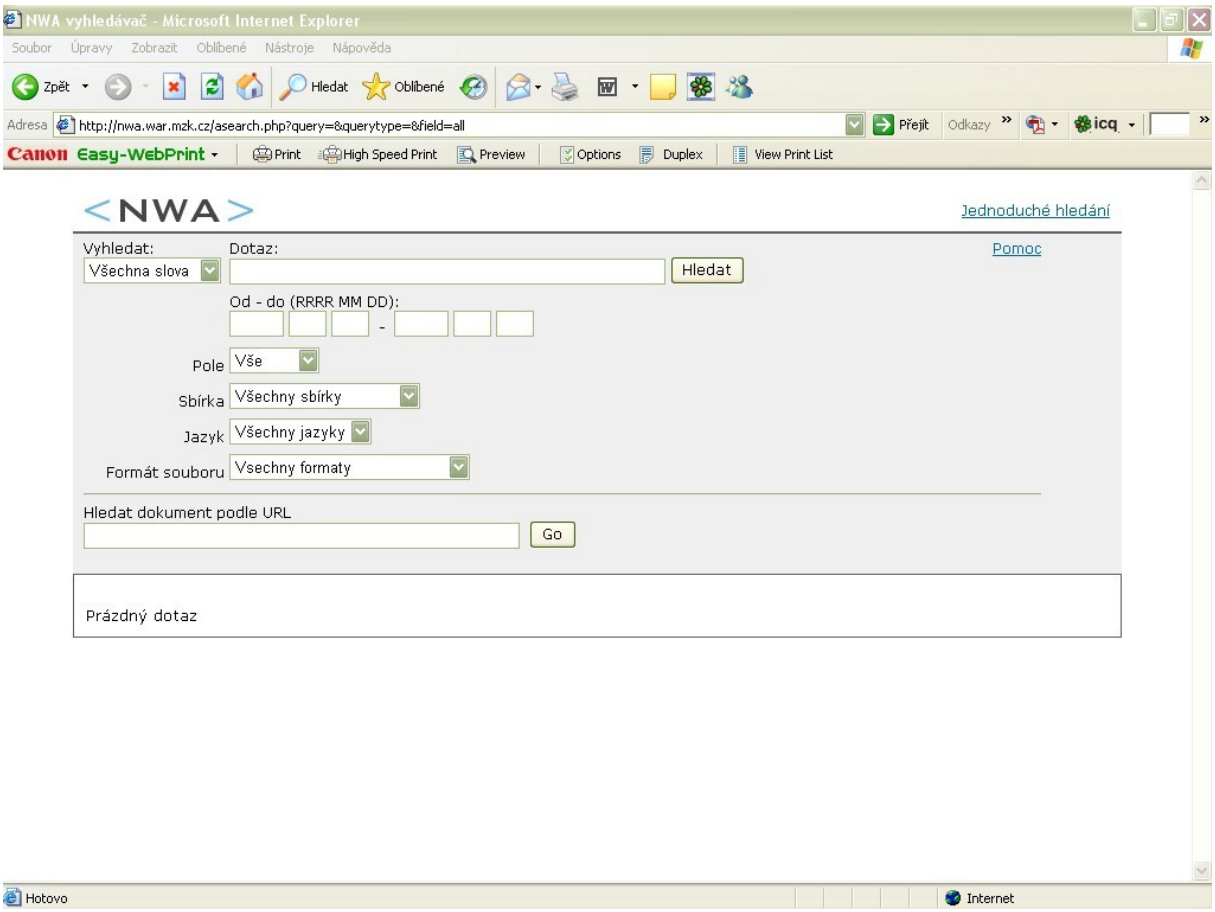
Vysvětlivky:

Použitá metoda při vytváření archivu: **H = Harvesting, V = Výběr**

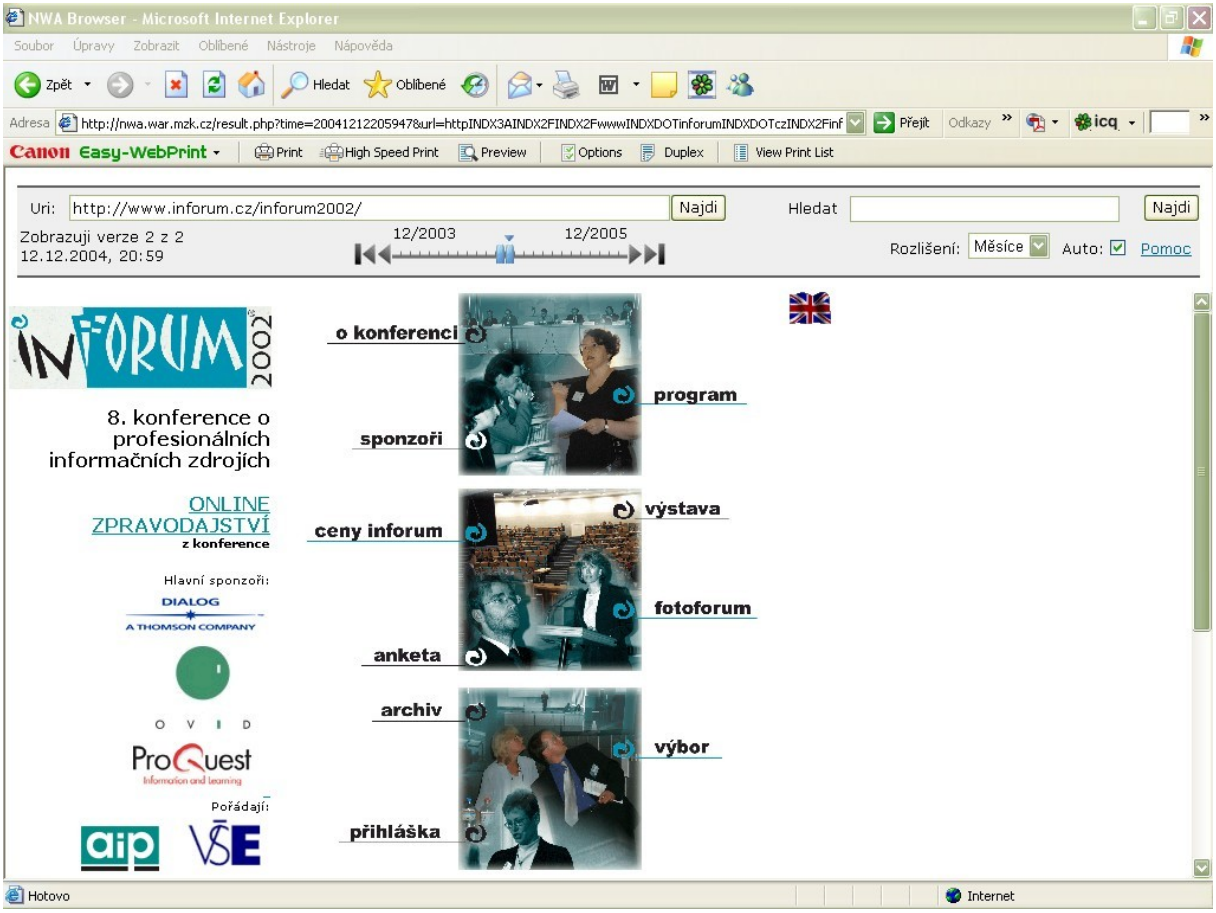
Přístup do archivu: **A = Ano N = Ne O = Omezený**

Příloha č. 3

Ukázka fulltextového vyhledávače NWA v českém prostředí


NWA vyhledávač - Microsoft Internet Explorer
Soubor Úpravy Zobrazit Oblíbené Nástroje Nápověda
Zpět Hledat Oblíbené
Adresa http://nwa.war.mzk.cz/asearch.php?query=&querytype=&field=all
Přejít Odkazy
Canon Easy-WebPrint
Print High Speed Print Preview Options Duplex View Print List
<NWA>
Jednoduché hledání
Vyhledat:
Dotaz:
Pomoc
Všechna slova
Hledat
Od - do (RRRR MM DD):
Pole
Vše
Sbírka
Všechny sbírky
Jazyk
Všechny jazyky
Formát souboru
Vsechny formaty
Hledat dokument podle URL
Go
Prázdný dotaz
Hotovo
Internet

Ukázka archivované verze Infora z roku 2002



EVIDENCE VÝPŮJČEK

Prohlášení:
Dávám svolení k půjčování této diplomové práce. Uživatel potvrzuje svým podpisem, že bude tuto práci řádně citovat v seznamu použité literatury.

V Praze, 19.8. 2005.

Markéta Škodová ( )

Jméno	Katedra / Pracoviště	Datum	Podpis