ARCHIV ČESKÉHO WEBU V ROCE 3

Petr Žabička
Moravská zemská knihovna, Brno

Projekt Webarchiv (*webarchiv.nkp.cz*) se letos dostává do třetího roku své existence. Na konci loňského roku byla odevzdána závěrečná zpráva projektu výzkumu a vývoje „Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet“ [1], která stručně shrnuje výsledky dosažené v prvních dvou letech řešení projektu. Díky úzké spolupráci Národní knihovny ČR (dále též NK) s Ústavem výpočetní techniky Masarykovy univerzity v Brně se však vývoj nezastavil a práce nastoupeným směrem pokračují v rámci řešení výzkumného záměru ÚVT MU „Digitální knihovny“ a nově jsou podpořeny i grantem, který NK získala pro rok 2002 v programu VISK3. V tomto článku se blíže podíváme na problematiku, kterou se projekt zabývá.

Cílem projektu Webarchiv je, jak již jeho název napovídá, zajištění trvalého uchování domácích elektronických, online publikovaných informačních zdrojů jako součásti národního kulturního dědictví. V článku [2] rozebírá L. Celbová jednotlivé okruhy problémů, které je v této souvislosti nutno řešit. Ačkoli se v tomto textu budeme zabývat spíše technickými aspekty celé problematiky, pokusíme se zmíněné rozdělení do jednotlivých okruhů co nejvíce dodržet. Budeme se tedy zabývat následujícími tématy:

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

V jejich rámci pak budou popsány jak zatím dosažené výsledky projektu, tak i nyní řešené problémy a plány dalšího výzkumu a vývoje. Samotná technická infrastruktura byla již popsána v několika článcích, naposledy pak přehledným způsobem v příspěvku „Infrastruktura Webarchivu v roce 2002“ [3] na konferenci Inforum 2002. Tam byli také posluchači seznámeni s prvními informacemi o letos probíhající úplné sklizni domény *.cz*, která běží již od 23. 4. 2002 a která by měla dát řešitelskému týmu odpověď na mnoho otázek souvisejících s dalším vývojem projektu.

Legislativa

V současné době existuje několik kritických míst, která mohou, ať už v pozitivním nebo negativním smyslu, ovlivnit další řešení projektu. V této kapitole si ukážeme, jaký vliv mohou mít různé výklady zákona na další vývoj projektu.

Prvním místem, kde dochází ke střetu obecného zájmu na zachování kulturního dědictví budoucím generacím s u nás platnou legislativou, je problematika povinného výtisku. Zákon u nás totiž nezakotvuje jednoznačně povinnost vydavatele odevzdávat povinný výtisk elektronicky publikovaných dokumentů. Zkušenost ze Švédska ukazuje důležitost takového právního zakotvení: automatická archivace (sklízení) online publikovaných elektronických zdrojů národní knihovnou, která je jedním ze způsobů, jak získat povinný výtisk tohoto typu dokumentů, zde musela být na mnoho měsíců přerušena právě proto, že úřady se nedokázaly shodnout na tom, zda je taková činnost legální, a patovou situaci vyřešilo až přijetí příslušného zákona [4]. Naopak v Dánsku, kde je podobný zákon v platnosti již delší dobu, nemá většina vydavatelů o jeho existenci ani tušení a jediným efektivním řešením problému je tak opět jedině automatická archivace všech publikovaných dokumentů.

Nedotaženost zákona o povinném výtisku u nás v tomto směru otevírá cestu různým výkladům omezení daných zákonem o autorském právu. Automatickou identifikaci a archivaci online publikovaných dokumentů lze srovnávat s běžně používanou technologií indexování webu, jak ji provádějí prohledávače Internetu. Přesto ale není jisté, zda bude bez opory v zákoně možné využívat stávající strategii. Existující infrastruktura je však nastavitelná tak, že bude možné zachovat alespoň omezený rozsah sklízení i v případě, že by bylo nutné podřídit se určitým zákonným omezením. Jediným důsledkem takových omezení by pak bylo velmi výrazné zmenšení rozsahu sbírky, tvořené pak víceméně na základě dobrovolně dodávaných dokumentů. Na druhou stranu by se díky takovému zásahu výrazně zmenšila i finanční náročnost hardwaru pro uložení takového archivu.

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

Je možné prohlásit, že právo občana na informace by mělo být naplněno i existencí digitální knihovny, obsahující elektronicky publikované dokumenty v nezměněné podobě. Zajištění integrity takové knihovny musí být proto jedním z prioritních úkolů jejího provozovatele. V případě Webarchivu je tato kontrola zajištěna použitým systémem jednoznačných identifikátorů dokumentů na bázi kontrolního součtu MD5 [5], což je v současné době pro tento účel nejdostupnější mechanismus, který však musí být

podpořen dalšími opatřeními jak na úrovni technické, tak organizační. V případě podcenění tohoto aspektu se totiž můžeme dostat do téměř orwellovské situace, ve které nebude možné nijak ověřit autenticitu v minulosti publikovaných informací a dokumentů – ať už proto, že z veřejného prostoru zcela zmizí, nebo proto, že budou z různých důvodů pozměněny nebo zcela nahrazeny jiným obsahem.

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

Kritéria výběru zdrojů a strategie jejich archivace

Stanovení podmínek, které musí splňovat elektronické zdroje kandidující na včlenění do budovaného digitálního archivu, je jedním z nejkritičtějších okamžiků každého podobného projektu. Při stanovování těchto podmínek je nutno brát v úvahu jak objem finančních prostředků, které jsou pro tuto činnost k dispozici, tak i aktuální stav rozvoje celé oblasti informačních a komunikačních technologií. V následujících odstavcích si ukážeme, jak tyto okolnosti ovlivnily řešení projektu.

Protokoly, formáty

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

Je zřejmé, že pokus archivovat online elektronické zdroje, dostupné jinak než prostřednictvím Internetu, by byl velmi nákladný a jeho přínos by byl mizivý. Takové jednoznačné tvrzení již však nelze pronést, máme-li na mysli newebové Internetové zdroje. Většinou totiž dopředu nelze určit, která technologie začne mít v budoucnosti význam a která je jen drobnou epizodou v dějinách Internetu (tak skončily v propadlišti dějin technologie typu „push", kterým byla kdysi prorokována velká budoucnost, tak se naopak objevuje přes aktivní odpor zábavního průmyslu stále větší množství různých typů sítí peer-to-peer). Přesto lze zatím stále obhájit názor, že většině populace je reálně přístupná jen ta část zdrojů, ke kterým se dostanou prostřednictvím běžného prohlížeče. Pokud tedy pomineme relativně velkou množinu mailových a newsových diskusních skupin, zůstává před námi dvojice protokolů http a ftp (protokol gopher lze již považovat za mrtvý, protokol https pak díky tomu, že je určen pro šifrovaný přenos dat, za protokol určený k přenosu neveřejných a důvěrných informací, tedy informací, dostupných sice elektronicky, ale ne nutně veřejně).

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

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

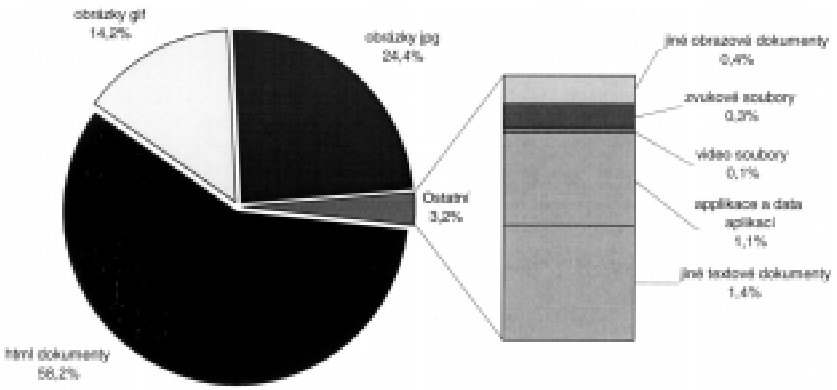
Prostorový a časový rozsah

Dosavadní zkušenosti ukazují, že z hlediska dlouhodobé konzervace nejvýznamnější část dokumentů je dostupná přes protokoly http a ftp a je uložena především v souborech formátů html, jpg a gif. Přes protokol ftp je však mimo jiné zpřístupněno i obrovské množství zrcad-

Tabulka 1: Zastoupení souborů v archivu podle formátů

	počet souborů [tis.]	počet souborů [%]	průměrná velikost souboru [kB]	celková velikost souborů [GB]	celková velikost [%]
html dokumenty	4.092	58,15	16,75	65,39	28,40
obrázky jpg	1.719	24,43	27,09	44,42	19,29
obrázky gif	1.002	14,24	7,60	7,26	3,15
jiné textové dokumenty	96	1,36	318,79	29,08	12,63
aplikace a data aplikací	78	1,11	599,18	44,74	19,43
jiné obrazové formáty	25	0,35	114,46	2,70	1,17
zvukové soubory	21	0,30	1 082,90	21,94	9,53
video soubory	4	0,06	3 953,34	14,70	6,38
celkem	7.037	100,00	34,31	230,23	100,00

Graf 1: Relativní četnost souborů v archivu podle typů



lených zahraničních archivů. Proto je vhodné sklízení v případě protokolu ftp zaměřit jen na relevantní dokumenty, tedy dokumenty přímo odkazované ze stránek přístupných přes protokol http. Dalším kritériem, které může velmi významně ovlivnit objem a kvalitu archivu, je pak stanovení rozsahu archivace.

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

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

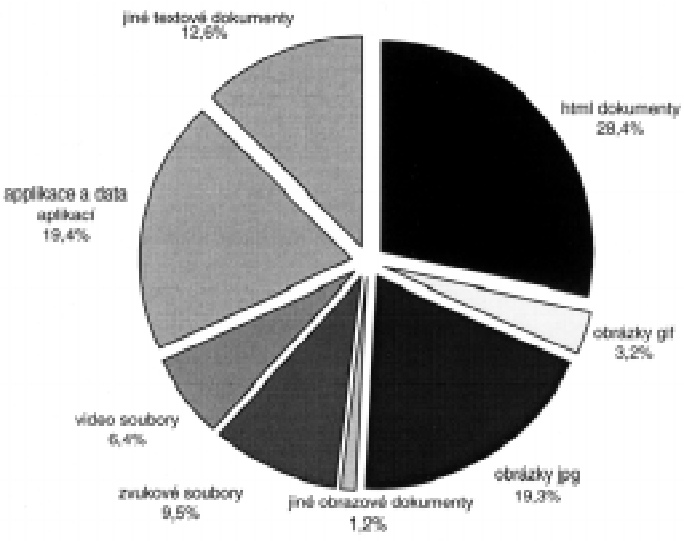
Aby bylo možno tyto postupy realizovat, je nutno nejprve stanovit, jaký je vlastně rozsah českého webu. Ačkoli jej můžeme zjednodušeně definovat jako

1) všechny dokumenty, publikované v doméně *.cz*, je zřejmé, že toto kritérium nemůže pokrýt celou českou online produkci. Proto by bylo vhodné tento rozsah rozšířit o mnoho dalších, vzájemně se prolínajících kategorií:
2) dokumenty v doménách druhé úrovně, registrovaných na subjekt se sídlem v České republice
3) dokumenty publikované na serverech fyzicky umístěných v ČR
4) dokumenty v českém jazyce
5) dokumenty českých autorů
6) dokumenty se vztahem k Česku.

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

Už získání údajů o rozsahu domény *.cz* (případ 1) není triviální. Správce domény nejvyšší úrovně .cz, sdružení CZNIC (www.nic.cz), sice na svých stránkách zveřejňuje přehledové statistiky (z těch plyne, že celkový počet registrovaných domén druhé úrovně se nyní pohybuje okolo 118 000) a zpřístupňuje detaily o jednotlivých doménách, ale kompletní seznam domén druhé úrovně nezveřejňuje. Naštěstí je zatím možné tento seznam standardní cestou získávat z jednoho ze zahraničních sekundárních jmenných serverů pro doménu *.cz* a bude tak možné zkoumat, nakolik účinná je druhá cesta vedoucí k získání těchto údajů, používaná všemi webovými roboty, tedy i Nedlib Harvesterem. Tato metoda je založena na extrakci odkazů vedoucích k dalším dokumentům během postupného procházení všech webových stránek, splňujících zadaná kritéria. Celkový počet tímto způsobem nalezených domén druhé úrovně již překročil hranici 30 000.

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



Graf 2: Zastoupení hlavních typů souborů a archivů podle velikosti

Obdobně ve 3. případě je možné se pokusit zjistit co nejpřesněji rozsahy IP adres používaných našimi primárními poskytovateli připojení (tj. členy sdružení Neutral Internet eXchange, www.nix.cz). O tyto adresy pak bude obohacena databáze povolených adres. Tím zajistíme, že při sklizni nebudou vynechány ty servery, na které není odkazováno jménem, ale jen IP adresou. Teoreticky bychom sice mohli i aktivně skenovat tyto rozsahy IP adres a hledat tak spuštěné www servery, není ale pravděpodobné, že by takto nalezené neregistrované servery obsahovaly hodnotné informace.

Ve 4. případě je situace složitější: procházení celosvětového webu s cílem najít stránky v českém jazyce je sice technicky realizovatelné, zároveň však neefektivní. Je ale možné, že v budoucnu půjde tento problém alespoň částečně vyřešit ve spolupráci s dalšími institucemi zabývajícími se touto problematikou tak, že všechny dokumenty stažené danou národní institucí budou podrobeny automatické analýze pro rozpoznání jazyka a odkazy na nalezené stránky v cizím jazyce by byly předány příslušné národní instituci.

Na rozdíl od výše uvedených bodů by v případech 5 a 6 bylo velmi obtížné, ne-li nemožné, automaticky rozhodnout, zda daný dokument spadá do zájmové oblasti. Zde bude záležet především na knihovnících nebo na vydavatelích samotných, zda takový server nebo dokument zaregistrují. Prostředky pro takovou registraci budou připraveny ještě během letošního léta.

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

Stanovili-li jsme si tedy alespoň přibližně rozsah českého webu, můžeme v jeho rámci začít hledat takovou podmnožinu zdrojů, kterou by bylo vhodné archivovat v co největší úplnosti, a tedy i co nejčastěji. Zde už nevystačíme s čistě technickým řešením, tuto podmnožinu, která bude zároveň kandidátem na zařazení do národní bibliografie, budou muset vytipovávat lidé k tomu určení.

V současné době se nabízí několik způsobů, jak tuto činnost zajišťovat; nejperspektivnějším z nich by mohlo být využití potenciálu projektu Jednotné informační brány CASLIN (*octopus.ruk.cuni.cz*). Jedním z výstupů tohoto projektu totiž je průběžně aktualizovaný předmětově členěný informační portál online elektronických zdrojů. Správa jednotlivých oborů tohoto portálu bude svěřena vždy té knihovně, která má v daném oboru největší zkušenosti. Díky tomu lze očekávat, že každý obor bude v portálu reprezentován nejvýznamnějšími informačními zdroji, které se tak zároveň stanou, pokud splní další kritéria, součástí národní bibliografie.

Je zřejmé, že takto pojatý systém může mnoho serverů z národní bibliografie vyloučit, na druhou stranu je nutno mít na zřeteli to, že každý zdroj zahrnutý do skupiny pro intenzivní sklízení s sebou nese nemalý díl kvalifikované lidské práce spojené s jeho zpracováním a případně analytickým popisem. Finanční náročnost může být v takovém případě samozřejmě snížena, dojde-li k nějaké formě dohody s příslušným vydavatelem.

Vliv technického řešení na rozsah a průběh sklizně

Volbou nejvhodnějšího nástroje pro plošnou archivaci webu se v současné době zabývá několik projektů v různých evropských zemích, za všechny lze zmínit testování v Rakousku nebo v Dánsku (*www.netarkivet.dk*). U nás používaný produkt, NEDLIB Harvester [6], v těchto srovnávacích testech rozhodně nezaostává a potvrzuje se tak, že byl v pilotní fázi projektu zvolen pro archivaci českého webu zcela oprávněně.

Jako každý správný program, i Harvester je samozřejmě do značné míry konfigurovatelný. Vedle seznamu výchozích webových stránek a omezení rozsahu sklizně pomocí URL nebo jejich částí lze nastavit i řadu dalších parametrů. Mezi ně patří především povolení nebo zakázání podpory protokolu ftp, logování zamítnutých URL, akceptování omezení pro roboty na jednotlivých serverech (robots.txt), podpora URL s parametrem, nebo maximální hloubka zanoření v rámci jednoho serveru. Zvláště poslední dva parametry mohou velmi významně ovlivnit rozsah a kvalitu sklizně.

Podpora URL s parametry umožňuje omezit sklízení jen na ta URL, která neobsahují znak '?', jenž v URL uvozuje seznam parametrů. Díky tomu lze sice do značné míry zabránit problémům spojeným s nekonečnými smyčkami při procházení serverů, na druhé straně se tak ale nepříjemně omezuje rozsah sklizně. Jako typický příklad lze uvést server root.cz, jehož jedinou stránkou, na kterou se dá dostat pomocí URL bez parametru, je jeho hlavní stránka. Protože podobně funguje většina elektronických periodik, vyřadili bychom ignorováním URL s parametry právě ty zdroje, které jsou z hlediska našeho kulturního dědictví nejcennější. K zabránění vzniku nekonečných smyček, kdy harvester donekonečna prochází automaticky generované stránky na nějakém serveru jen proto, že naráží na stále další dynamicky generovaná URL ukazující ve skutečnosti na stále stejný cíl, slouží právě parametr maximální hloubka zanoření. Ten je nyní nastaven na 50 kroků a umožňuje tak sklízet bez problémů velkou většinu toho, co český web nabízí.

Je ale pravděpodobné, že mnohé dynamicky generované stránky se v archivu vyskytnou několikrát jen proto, že se navzájem nepatrně liší. Typickým příkladem jsou zde opět webové stránky knihovního systému Aleph, které obsahují ve svém URL i ve všech dalších odkazech dynamicky generovaný identifikátor sezení (session-id), takže URL může vypadat například takto:
http://aleph.mzk.cz/ALEPH/YIG1EJP2FBE7SEA4VNNM1KV97Q5T78FFN22M3ENFSHSUDS66S8-01211/file/start-0.

Pokud se harvester na takovouto stránku vrátí s časovým odstupem delším než několik minut, původní sezení je už na straně Alephu uzavřeno a je vygenerován nový identifikátor ve formě nového URL. To je pak opět navštíveno s delším časovým odstupem a opakovaně archivo-

váno. Tento cyklus se opakuje tak dlouho, dokud není vyčerpán povolený počet zanoření. Až podrobnou analýzou výsledků sklizně ale bude možné rozhodnout, jak časté takové případy jsou a jak se před nimi bránit nejen v tomto, ale i v dalších podobných případech. Zde je však nutno poznamenat, že k podobným problémům dochází pouze v případě, že správce daného serveru ve vlastním zájmu v souboru robots.txt nezakáže všem robotům přístup na inkriminovaná URL.

Web Národní knihovny je jedním z těch, na kterých bude možné po skončení sklizně ověřit kvalitu algoritmů harvesteru. Tento web je totiž dostatečně rozmanitý na to, aby se na něm vyskytovala většina dnes běžně používaných webových technologií, zároveň je ale tak malý, že bude snadné porovnat jeho skutečný rozsah s tím, co sklidil harvester. Již nyní je ale zřejmé, že při sklízení webu Národní knihovny harvester mnoho důležitých informací vůbec nenašel, protože byly skryty za různými druhy prohledávacích rozhraní – jako příklad můžeme jmenovat naskenované lístkové katalogy, obsah databází Alephu a další.

Je samozřejmé, že ať už je pro archivaci zvolen jakýkoli produkt, je jím vytvořený archiv poplatný jeho limitům. Ani NEDLIB Harvester není v tomto směru samozřejmě výjimkou, a tak existuje několik prozatím nepřekročitelných omezení. Tím nejvýraznějším omezením Harvesteru je absence podpory javascriptu. Důsledkem tohoto stavu je to, že v archivu zcela chybějí ty stránky, na něž vedou jen odkazy generované javascriptem až v prohlížeči (typickým příkladem takových odkazů jsou odkazy do archivu Neviditelného psa). Zatím méně bolestivým nedostatkem stejného charakteru je absence podpory prezentací ve formátu flash.

Tabulka 2: Nejrozsáhlejší domény v archivu

název domény 2. úrovně	počet souborů	celková velikost souborů [MB]	průměrná velikost souboru [kB]	pořadí podle počtu souborů	pořadí podle celkové velikosti	zaměření domény
3web	22 454	511	23,3	24	58	Nixnet - webhosting
aktualne	22 499	839	38,2	23	37	Webzdarma - webhosting
atlas	19 751	461	23,9	28	67	Atlas - portál, webhosting
borec	23 156	1 027	45,4	21	31	Webzdarma - webhosting
cas	22 608	1 293	58,6	22	21	akademie věd
compaqplus	2 432	2 738	1 152,8	528	10	firemní stránky
cpress	7 189	920	131,0	69	34	vydavatelství, portál
cuni	61 395	1 548	25,8	8	19	vysoká škola
cvut	62 192	9 611	158,2	7	1	vysová škola
d2	68 819	1 788	26,6	5	16	Nixnet - webhosting
datasys	5 848	1 188	208,0	87	26	firemní stránky
euweb	27 656	1 030	38,1	17	30	Webzdarma - webhosting
fbi	25 641	687	27,4	19	49	S4U - webhosting
freemusic	7 265	1 891	266,5	68	14	hudba
gamesweb	66 704	4 195	64,4	6	3	Zoner - hry
gamez	2 293	1 107	494,2	587	27	NetCentrum – hry. **ISSN**
gamezone	6 053	1 049	177,4	82	29	Quick - hry
hyperlink	107 380	2 990	28,5	3	6	Cpress - webhosting
hyperlinx	49 125	1 868	38,9	11	15	Cpress - webhosting
idnes	18 585	747	41,1	29	41	Dnes - zpravodajství
ihned	21 841	867	40,6	25	36	HN - zpravodajství. **ISSN**
jrc	2 537	2 505	1 011,3	471	11	firemní stránky - hry
kgb	60 875	1 481	24,9	9	20	S4U - webhosting
linux	6 398	1 091	174,7	78	28	linux
misto	26 643	903	34,7	18	35	Reflektor - webhosting
mp3records	12 347	3 084	255,8	41	5	hudba
mp3shop	2 689	1 997	760,3	359	13	hudba

muni	59 510	2 879	49,5	10	9	vysoká škola
mysteria	16 405	714	44,5	35	46	Webzdarma - webhosting
nhlpro	41 608	1 600	39,4	12	18	sport
quick	30 049	954	32,5	16	33	Quick - webhosting
senat	3 305	1 017	315,2	203	32	státní instituce
sumanet	345	1 279	3 796,9	3189	22	firemní stránky
techno	17 511	242	14,2	33	119	hudba. **ISSN**
tiscali	17 291	2 918	172,8	34	8	Tiscali - portál, webhosting
unas	37 784	1 211	32,8	14	24	Webzdarma - webhosting
vfu	2 765	2 502	926,7	276	12	vysoká škola
volny	17 601	497	28,9	32	60	Volny - portál, webhosting
vse	17 978	487	27,7	30	62	vysoká škola
vutbr	25 231	1 263	51,3	20	23	vysoká škola
web3	21 020	468	22,8	26	65	Nixnet - webhosting
webpark	153 167	3 932	26,3	2	4	NetCentrum - webhosting
webz	41 469	1 648	40,7	13	17	Webzdarma - webhosting
webzdarma	74 144	2 925	40,4	4	7	Webzdarma - webhosting
worldonline	9 588	1 193	127,4	54	25	Tiscali - webhosting
wz	153 729	6 379	42,5	1	2	Webzdarma - webhosting
xko	35 017	714	20,9	15	45	diskusní fóra
xpoint	19 871	310	16,0	27	85	Cpress - diskusní fóra
zpravodaj	17 738	564	32,6	31	55	Aliaweb - webhosting
výběr celkem	**1 575 501**	**85 110**	**55,3**			

Další nepříjemné omezení není dáno ani tak vlastnostmi softwaru, jako výkonem nyní používaného hardwaru. Ačkoli je nyní Harvester připojen k Internetu rychlostí 100 Mbit/s a mohl by tedy teoreticky za den stáhnout řádově stovky GB dat, server, na kterém je nyní provozován, dovoluje stahovat jen asi 6 GB dat denně. Tento problém bude možné odstranit až pořízením nového serveru v průběhu druhého pololetí letošního roku. Protože zde přichází v úvahu několik hardwarových platforem (především PC server/Linux a Sun/Solaris), proběhnou během léta zátěžové testy uvažovaných serverů, které by měly během několikadenního sklízení ukázat jejich silné i slabé stránky. Jakmile bude vybraný server napevno nainstalován, bude na něj přenesen provoz harvesteru provádějícího aktuální sklizeň a původní server tak bude uvolněn pro vývoj a testování.

Výsledky dosavadního sklízení

V úvodu zmíněná a v době psaní tohoto článku (červenec 2002) již třetí měsíc běžící sklizeň celé domény *.cz* by měla ukázat mimo jiné i to, jaký je skutečný rozsah českého viditelného webu. Výchozími body pro tuto sklizeň byly především hlavní stránky internetových portálů seznam.cz a quick.cz. Přes výše zmíněné problémy se do konce července 2002 podařilo stáhnout 10 057 247 souborů o celkové velikosti přes 241 GB. Tabulka 2 pak ukazuje 50 domén druhé úrovně, které byly k 10. 6. 2002 na prvních 35 místech buď podle celkového počtu z nich stažených souborů, nebo podle jejich celkové velikosti. Tato necelá 2 promile počtu doposud alespoň jednou navštívených domén 2. úrovně tak nyní představují přibližně čtvrtinu objemu dosavadní sklizně.

Tato tabulka také naznačuje, jaké informační bohatství český web skrývá: mezi těmito padesáti největšími doménami, zastupujícími z velké části webhostingové firmy, najdeme 3 servery, kterým bylo přiděleno ISSN, 6 univerzit, 1 univerzitou provozovaný specializovaný server (linux.cz), Českou akademii věd a několik zpravodajských a vydavatelských serverů. Pro zajímavost: doména nkp.cz je nyní s 5680 soubory a 130 MB na 92. místě podle počtu souborů, resp. na 234. místě podle jejich objemu, a proto je z tohoto hlediska škoda, že se některé aktivity NK prezentují mimo tuto doménu.

Dlouhodobé uchování zdrojů

Velikost Harvesterem tvořeného archivu může snadno dosáhnout obrovských rozměrů: jedno kolo stahování představuje v našich podmínkách stovky GB a může překročit i hranici 1TB. Archiv s tak velkým potenciálem růstu není samozřejmě snadné ani levné provozovat. Ač-

koli v současné době již jsou na trhu za nízkou cenu pevné disky o kapacitách více než 100 GB, 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 1TB 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. V pilotní fázi projektu bylo s výhodou využito toho, že takové zařízení již v NK existuje a je jím páskový robot, který hostí i data z mnoha dalších, především digitalizačních projektů. Výhodou páskového robota v pilotní fázi projektu byla především jeho rozšiřitelnost – dokoupením relativně levných pásek bylo možné rozšířit jeho kapacitu tak, aby robot pojmul všechna data získaná Harvesterem. V letošním roce již nebude další rozšiřování jeho kapacity tak levné, protože již byla vyčerpána licencí pokrytá kapacita robota a mimo další pásky bude nutné zaplatit i rozšíření licence. Zde se však ukazuje výhoda využívání jednoho zařízení více projekty a aktivitami v rámci NK, protože sdružením finančních prostředků se daří dosáhnout celkově výhodnějších cen od dodavatelů.

Další výhodou páskového robota je bezpečnost na něm uložených dat, která je zajištěna vysokou mírou redundance – každý dokument je uložen na třech různých páskách. Relativně rychlou dostupnost jak pro zápis, tak pro čtení pak zajišťuje předřazené diskové pole, které funguje jako cache paměť pro souborový systém robota.

Vzhledem k velkému objemu ukládaných dat nejsou archivované dokumenty ukládány do žádné databáze, ale přímo do souborového systému robota. Dalším důvodem podporujícím toto řešení je i usnadnění budoucí migrace archivu na nové platformy – je nutné si uvědomit, že budovaný archiv by měl být trvale dostupný i ve vzdálené budoucnosti, kdy už současný hardware beznadějně zastará. Protože se žádný souborový systém nedokáže rozumně vypořádat s velkým množstvím malých dokumentů, jsou nově získané dokumenty před uložením do archivu spojovány programem tar do balíků po dvou tisících a poté jsou ještě komprimovány programem gzip. Spolu s každým dokumentem jsou do balíku uložena v samostatném souboru i metadata popisující jeho vlastnosti, okolnosti jeho stažení a v případě html dokumentu i všechna metadata, která v něm byla obsažena. Průměrná velikost jednoho takového balíku dat je 56 MB, díky kompresi se ušetří přibližně 15 % prostoru – relativně nízká úroveň komprese je dána převahou komprimovaných formátů souborů uložených v archivu. Velký počet souborů v balíku sice může působit problémy při zpřístupnění archivu, na druhou stranu se s takto vybudovaným archivem lépe manipuluje.

Lze předpokládat, že po hardwarové stránce nebude dlouhodobé uchování archivu obtížné. Růst kapacity paměťových médií při současném poklesu cen dává naději, že celková cena provozu archivu se nebude zvyšovat. Díky již zmíněnému ukládání dat do souborového systému by neměl být problém ani s migrací dat, která bude prováděna prostým zkopírováním na nové médium. Formáty tar a gzip jsou dostatečně zdokumentované a programy pro práci s nimi dostupné včetně zdrojového kódu pro každý existující operační systém, není tedy pochyb o tom, že archivované dokumenty zůstanou trvale přístupné.

Větším oříškem samozřejmě budou samotné archivované soubory. Je sice pravděpodobné, že nejrozšířenějšími formáty zůstanou formáty dlouhodobě interpretovatelné (html, txt, gif, jpg), lze ale mít oprávněné pochybnosti o všech proprietárních formátech, především těch, které nejsou tak rozšířeny jako například formáty firem Adobe nebo Microsoft. I u formátů Microsoftu je však zárukou jejich interpretovatelnosti spíše dostupnost alternativních programů s otevřeným kódem, které umějí s těmito formáty pracovat (OpenOffice), než podpora ze strany Microsoftu. Otázka, zda v budoucnosti takové formáty konvertovat, nebo zda jít cestou emulace [7], však zatím zůstává otevřená.

Ať už bude v budoucnosti vývoj tohoto archivu jakýkoli, lze říci, že využitím NEDLIB Harvesteru získala Národní knihovna vhodný nástroj pro tvorbu konzervačního archivu českého webu.

Bibliografická správa a zpřístupnění zdrojů

Vytvoření takového archivu je sice důležitým, ale zároveň jen prvním krokem na cestě k naplnění jeho smyslu, tedy ke zpřístupnění jeho obsahu. Prvním krokem při zpřístupňování archivu musí samozřejmě být alespoň rámcové stanovení použitých postupů, s nimi spojených pracovních procesů a jejich rozsahu.

Je zřejmé, že ani společným úsilím všech českých knihoven nebude nikdy možné zkatalogizovat celý archiv českého webu – tento úkol bude nutné přenechat „strojům“. Přes značný pokrok v oblasti počítačového porozumění přirozenému jazyku v posledních letech bude pravděpodobně ještě řadu let trvat, než bude možné začít uvažovat o nasazení plně automatizovaného nástroje pro bibliografický popis archivovaných dokumentů.

Fulltext

Pro zpřístupnění archivu nám tak zůstávají technologie fulltextového indexování a automatizované extrakce autorem vytvořených metadat. Koncem roku 2001 byl na MFF UK vypsán ročníkový týmový vývojový projekt na vytvoření indexační a vyhledávací aplikace pro Webarchiv. Tato aplikace by měla zpřístupnit stažené dokumenty v jejich kontextu, tedy s vloženou grafikou ze stejné doby a s odkazy vedoucími primárně opět do archivu. Vyhledávání v archivu by mělo být umožněno nejen na základě URL nebo kontrolního součtu dokumentu, ale i na základě z dokumentu extrahovaných metadat nebo fulltextového vyhledávání. Tato aplikace by měla být navržena tak, aby bylo možné k ní kdykoli připojit moduly pro indexování jiných, než textových typů souborů. Jakkoli se to může zdát na první pohled nereálné, nástroje tohoto typu již existují a jeden z nich, Convera Retrievalware, je dokonce v NK zkušebně provozován. Jedním z dalších cílů

projektu bude proto pokus o jeho využití pro indexování některých typů souborů obsažených v archivu.

Zda bude některá z těchto technologií nasazena v reálném provozu, bude samozřejmě záviset i na dříve zmíněných legislativních otázkách. Je totiž zřejmé, že stávající hardwarová platforma je pro ostrý provoz takového nástroje nevyhovující. To je dáno jednak nemožností souběhu harvestingu a indexace na jednom serveru, kapacitní problémy se ovšem týkají celé nyní používané hardwarové infrastruktury. Pokud bychom se totiž například rozhodli využít pro fulltextové indexování nástroj Retrievalware, bude nutné pro každých 100 GB textových souborů mít k dispozici 150GB diskového prostoru pro tvorbu indexů, resp. 70 GB pro jejich uložení. Pokud navíc bude potřeba zpřístupnit archiv více než jednomu současnému uživateli, bude pravděpodobně nutné výrazně změnit dosavadní systém uložení dat. Páskový robot by pak bylo možné využívat jen jako zálohovací zařízení, protože by nebyl schopen rychlé odezvy na větší počet paralelně přicházejících požadavků. Dalším problémem by se pak mohlo stát samo uložení souborů do velkých komprimovaných balíků. Pokud se pak takový desítky megabajtů velký balík musí dekomprimovat kvůli získání několikabajtového obrázku, začne být i zde rychlost odezvy na považenou, i když balík je již uložen na disku. Je proto možné, že v budoucnosti bude nutné hledat kompromisní řešení cestou snížení počtu souborů v balíku přinejmenším na desetinu, nebo bude nutné zcela oddělit podobu archivu na páskovém robotu od podoby na rychlejším médiu, ke které by měly přístup programy zpřístupňující archiv uživatelům.

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

Metadata

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

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

Dublin Core Metadata Generator byl původně společně s dalšími nástroji převzat s minimálními úpravami od Helsinské univerzitní knihovny, která jej vyvinula v rámci projektů Nordic Metadata I a II (*http://www.lib.helsinki.fi/meta/*). Na základě výsledků zkušebního provozu byl program postupně upravován až do dnešní podoby. Významným pokrokem zde bylo například zavedení podpory extrakce externě uložených metadat ve formátu RDF/XML. Výstupní formát HTML byl upraven tak, aby vygenerovaná metadata byla kompatibilní s XHTML 1.0, zatímco výstup generovaný ve formátu XML/RDF byl zpřehledněn a byla aktualizována použitá syntaxe.

I samotný formulář pro vkládání metadat doznal určitých změn, z nichž nejvýznamnější je volba kvalifikátorů prvku *Subject* tak, aby odpovídaly u nás používaným systémům věcného třídění, a také doplnění funkce automatického vložení jedinečného čísla národní bibliografie ve formátu URN přímo do pole *Identifier*, pokud bylo toto pole předtím prázdné. To zajišťuje uživateli větší pohodlí a výrazně zmenšuje riziko chyb hrozících jinak při kopírování nebo přepisu identifikátoru. Doufáme, že právě cesta získávání URN autory dokumentů během tvorby metadat popisujících tyto dokumenty v budoucnosti učiní používání samostatného formuláře pro přidělování URN zbytečným.

Zmíněné přidělení jednoznačného identifikátoru je umožněno propojením Dublin Core generátoru s generátorem URN. Ten byl nejprve jen lokalizován, ale právě kvůli propojení s DC generátorem musel být později mírně upraven. Již nyní se ale chystá úprava systému přidělování URN tak, aby program přidělující URN fungoval jako samostatný URN server, přičemž budou zveřejněny funkce pro získání URN v často používaných programovacích jazycích, což umožní snadnou integraci této funkce přímo do publikačních systémů vydavatelů online zdrojů. Díky tomu by se přidělování URN mělo stát zcela automatickým procesem.

Řadu pomůcek dostupných na serveru Webarchivu doplnil i kalkulátor MD5**.** Ten umožňuje spočítat kontrolní součet MD5 zadaného textového řetězce. Pokud je tímto řetězcem platné URL nějakého dokumentu, může kalkulátor tento dokument stáhnout a spočítat jeho kontrolní součet. Protože jsou tyto kontrolní součty používány pro identifikaci dokumentů archivovaných Harvesterem, je jedna z možností využití Kalkulátoru zřejmá: může sloužit jako pomůcka při analýze práce Harvesteru i při zkoumání archivu samotného.

Národní bibliografie

Pokud bychom z archivu vydělili ty dokumenty, ke kterým existuje metadatový popis podle standardu Dublin Core, mohli bychom na jejich základě vybudovat menší bibliografickou databázi obsahující případně i plné texty dokumentů. Ani zde by nebylo nutné provádět vývoj na zelené louce, protože přesně taková databáze je již v NK

v oddělení analytického popisu provozována (*full.nkp.cz*) a aplikace, která za touto databází stojí, by se určitě dala upravit tak, aby byla schopna přijímat dokumenty předávané automaticky z WebArchivu.

Ani takto získané záznamy se však nemohou bez vyhodnocení obsahu primárního dokumentu a podrobnějšího zpracování stát součástí České národní bibliografie. Právě takovou skupinou dokumentů by ale mohly být dokumenty získané cestou intenzivního sklízení. Jak jsme si již řekli, jde o dokumenty ze zdrojů vytipovaných samotnými knihovníky při tvorbě předmětového portálu Jednotné informační brány CASLIN. Tyto zdroje, kterých by měly být řádově desítky, nejvýše pak stovky, by mohly být řádně zkatalogizovány ve formátu MARC, na což by případně mohlo navázat kooperativní analytické zpracování vybraných článků, opět standardním způsobem v tomtéž formátu. Pomocí od vydavatele by zde samozřejmě mohlo opět být vložení metadat přímo do zdrojového textu článku.

Mezinárodní spolupráce

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

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

Dalším souvisejícím krokem na poli mezinárodní spolupráce bylo pak navázání kontaktů s týmem Technické univerzity ve Vídni, řešícím problematiku archivace rakouského webu ve spolupráci s Rakouskou národní knihovnou.

Díky navázání těchto kontaktů se pak NK společně s Masarykovou univerzitou a dalšími dvěma českými firmami mohla stát členem skupiny národních knihoven a dalších organizací z třinácti evropských zemí, které společně podaly Vyjádření zájmu (Expression of Interest) o vypsání projektu s názvem „Archiv evropského webu" v rámci 6. rámcového programu Evropské unie. Cílem tohoto projektového záměru je sjednotit roztříštěné národní iniciativy jednotlivých evropských zemí a podpořit tak vytvoření distribuovaného archivu evropského webu, založeného na síti národních archivů jednotlivých zemí. Záměr projektu je však mnohem ambicióznější než to, co bylo zatím možno dosáhnout v rámci ČR. Jeho cílem je vytvořit společné postupy, doporučení a položit základ jednotné infrastruktury v této oblasti. Bude-li tento záměr akceptován, je pravděpodobné, že již v příštím roce dojde v oblasti dlouhodobého uchování elektronických zdrojů k výraznému posunu jak v praktické, tak i v teoretické rovině.

Závěr

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

Literatura:

[1] *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet : závěrečná zpráva za léta 2000-2001* [online]. Praha : Národní knihovna ČR, leden 2002, [cit. 2002-06-15]. Dostupný na WWW: <http://webarchiv.nkp.cz/zprava2001/zprava2001.pdf>.

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

[3] ŽABIČKA, Petr. Infrastruktura Webarchivu v roce 2002. In *Inforum 2002*. Praha : Albertina icome Praha, s.r.o., 2001. Dostupný na WWW: <http://www.inforum.cz/inforum2002/prednaska8.htm>.

[4] LIDMAN, Thomas. *New Decree for Kulturarw3* [online]. Stockholm : The Royal Library, June 10, 2002, [cit. 2002-06-15]. Dostupný na WWW: <http://www.kb.se/Info/Pressmed/Arkiv/2002/020605_eng.htm>.

[5] ŽABIČKA, Petr. Nástroje pro tvorbu metadat Dublin Core. In *Automatizace knihovnických procesů - 8*. Praha : ČVUT - Výpočetní a informační centrum, 2001, s. 86-91. Dostupný též na WWW: <http://platan.vc.cvut.cz/akp/clanky/09.pdf>. ISBN 80-01-02-366-4

[6] ŽABIČKA, Petr. NEDLIB Harvester - technika „sklizně" informací. *Ikaros*. [online]. 2000, roč. 4, č. 10 [cit. 2002-06-15]. Dostupný na WWW: <http://ikaros.ff.cuni.cz/2000/c10/harvest.htm>. ISSN 1212-5075.

[7] ROTHENBERG, Jeff. *Using emulation to preserve digital documents.* Hague : Koninklijke Bibliotheek, July 2000. 69 s. Dostupné též na WWW: <http://www.konbib.nl/kb/pr/fonds/emulation/usingemulation.pdf>. ISBN 90-6259145-0.