

WebArchiv – digitální knihovna českého webu

Petr Žabička – Ludmila Celbová

S prudkým nárůstem objemu informací publikovaných výhradně na Internetu by se úkolem moderní depozitní knihovny mělo stát i shromažďování, ochrana a zpřístupnění online dostupných elektronických informačních zdrojů. V souladu se svým posláním se touto cestou vydala i Národní knihovna ČR, která ve spolupráci s Ústavem výpočetní techniky MU buduje archiv českého webu.

1. Archivace webu - situace ve světě

Jedním z průkopníků na poli archivace webu je americká nezisková organizace *Internet Archive* (www.archive.org), jejíž archiv sahá až do roku 1996. Tato organizace se ve spolupráci s dalšími institucemi snaží (vcelku úspěšně) vybudovat co nejrozsáhlejší archiv světového webu. Takový záměr je však finančně vysoce nákladný, v letošním roce proto zahájil Internet Archive spolupráci s největšími světovými národními knihovnami s cílem vyvinout novou generaci nástrojů pro archivaci a zpřístupnění webových informačních zdrojů. V průběhu tří let bude na vývoj těchto nástrojů a na archivaci webů zemí zúčastněných knihoven vynaloženo přibližně 3 miliony dolarů. Předpokládá se, že softwarové nástroje vyvinuté v rámci tohoto projektu budou dány k dispozici pod nějakým typem licence zajišťující volný přístup ke zdrojovým kódům.

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

Přístup jednotlivých knihoven k řešení problému se ovšem velmi liší. Některé knihovny, jako například Australská národní knihovna, se snaží *archivovat výběrově* jen ty webové zdroje, jejichž kvalitu předem zhodnotí knihovník (pandora.nla.gov.au). Díky tomuto přístupu čítá sice nyní archiv australského webu po několika letech provozu pouhých 3395 webových sídel nebo jejich částí, nicméně jde o výběr toho „nejdůležitějšího“, co bylo v dané době na webu publikováno. Tento přístup je však velmi náročný na lidské kapacity a proto se většina knihoven vydala cestou automatizované *plošné archivace* všech dokumentů, které splňují automaticky vyhodnitelná kritéria. K tomu využívají nejčastěji softwarových nástrojů vyvinutých v minulých letech v rámci projektů evropské unie nebo projektů evropských severovýchodních zemí. Vznikají však i další iniciativy, například ve výše zmíněném konsorciu Internet Archive se po několikaletém zkoumání problematiky rozhodly spojit své síly americká Kongresová knihovna, Britská knihovna, Francouzská národní knihovna a některé severovýchodní národní knihovny. V současné době tak problematiku archivace webu zkoumají národní knihovny nejméně 15 evropských zemí. Pozadu nezůstává ani Japonská národní knihovna a zahájen byl i projekt na archivaci webových zdrojů v čínštině.

Národní knihovna ČR zahájila své aktivity na poli archivace webu v roce 2000, když ve dvouletém pilotním projektu „*Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet*“ zprovoznila ve spolupráci s Ústavem výpočetní techniky Masarykovy univerzity infrastrukturu pro tvorbu digitálního archivu českého webu (webarchiv.nkp.cz). Díky navázané spolupráci a později i díky získání omezené grantové podpory Ministerstva kultury se práce na projektu nezastavily a pokračovat budou i v roce letošním.

2. Projekt WebArchiv

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

2.1 Výběr zdrojů k archivaci

Pokud padla v úvodu tohoto článku zmínka o „online“ publikovaných zdrojích, je nutné upozornit na to, že již rozhodnutí zaměřit se primárně na „webové“ zdroje znamená, že se zaměřujeme jen na jistou část množiny všech online elektronických zdrojů. Je zřejmé, že pokus archivovat online elektronické zdroje dostupné jinak než prostřednictvím Internetu, by byl velmi nákladný a jeho přínos pro archiv zanedbatelný. Takové kategorické tvrzení již však nelze pronést o ne-webových Internetových zdrojích. Většinou totiž dopředu nelze určit, která technologie začne mít v budoucnosti význam a která je jen krátkou epizodou v dějinách Internetu. Přesto lze zatím stále obhájit názor, že většině populace je reálně přístupná jen ta část zdrojů, ke kterým se dostanou prostřednictvím běžného www-prohlížeče a proto právě tato část zdrojů by měla být primárním předmětem zájmu Národní knihovny. 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 je určen pro šifrovaný přenos dat, lze jej proto považovat za protokol určený především k přenosu důvěrných informací, které nejsou předmětem veřejného zájmu).

Pokud dosavadní zkušenosti ukazují, že z hlediska dlouhodobé konzervace opravdu nejvýznamnější část dokumentů je dostupná přes protokoly http a ftp, je nutné dodat, že prostřednictvím protokolu ftp jsou zpřístupněny také obrovské objemy dat zrcadlených ze zahraničních archivů. Proto je vhodné sběr dokumentů v případě protokolu ftp zaměřit jen na ty relevantní, tedy na dokumenty přímo odkazované ze stránek přístupných přes protokol http. V případě již zmiňovaných diskusních skupin je nutno vzít v úvahu, že archivy mnoha z nich jsou zároveň přístupné ve formě html archivů dostupných i přes protokol http. Pokud by se ukázalo, že je důležité vytvářet jejich samostatný 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. Výzkumy ve světě, potvrzené i během dosavadní archivace českého webu ukazují, že cca 97% počtu všech archivovaných souborů tvoří trojice formátů html, jpg a gif, ačkoli co do velikosti zaujímají jen asi polovinu celkového objemu dostupných dat. Pokud tedy dokážeme odpovědně určit, které ze vzácněji se vyskytujících formátů nemá smysl z různých důvodů archivovat, můžeme snadno ušetřit třeba i třetinu kapacity ukládacího prostoru, což může představovat úsporu značných částek. Nesmíme totiž zapomínat na to, že dlouhodobé zachování dostupnosti informačního obsahu dokumentu (po desetiletí až staletí) je nutné zajistit buď prostřednictvím jeho konverze nebo prostřednictvím emulace, přičemž oba přístupy jsou technicky a tedy i finančně velmi náročné.

2.2 Český web

Jak již bylo uvedeno, předmětem zájmu projektu Webarchiv je archivace online publikované části české produkce, tedy, zjednodušeně řečeno, český web. V ideálním případě by tedy měl být výsledkem projektu archiv obsahující pokud možno vše, co kdy bylo v rámci českého webu publikováno. Proto se provádí archivace dvěma cestami: *plošnou archivací*, kdy se s delším časovým odstupem vytváří co nejuplněnější snímky celého českého webu (například

2krát ročně), a *výběrovou archivací*, kdy se naopak velmi často (v případě potřeby i každý den) doplňuje archiv zrcadlící jen omezenou vybranou skupinu nejvýznamnějších českých zdrojů.

Aby bylo možné tyto postupy realizovat, je nutné nejprve stanovit, jaký je vlastně *rozsah českého webu*. Ačkoli jej můžeme zjednodušeně definovat jako „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 doplňujících kategorií: dokumenty v doménách druhé úrovně registrovaných subjektem sídlícím v České republice; dokumenty publikované na serverech fyzicky umístěných v ČR; dokumenty v českém jazyce; dokumenty českých autorů; dokumenty se vztahem k Česku.

V doméně .cz je nyní registrováno téměř 130.000 domén 2. úrovně. Přidáváním dalších podmínek stoupá jak náročnost nalezení všech dokumentů podmínky splňujících, tak i náročnost prokázání, že nalezený dokument některou podmínku opravdu splňuje.

Stanovili-li jsme si tedy alespoň přibližně rozsah českého webu, můžeme v jeho rámci začít hledat takovou podmnožinu zdrojů, kterou by bylo vhodné archivovat výběrově v co největší úplnosti. V současné době se nabízí několik způsobů, jak tuto činnost zajišťovat; nejperspektivnějším z nich by mohlo být využití potenciálu projektu Jednotné informační brány CASLIN (www.jib.cz). Jedním z jejích výstupů bude totiž průběžně aktualizovaný předmětově členěný informační portál online elektronických zdrojů. Správa jednotlivých oborů tohoto portálu bude svěřena vždy té knihovně, která má v daném oboru největší zkušenosti. Díky tomu lze očekávat, že každý obor bude v portálu reprezentován i nejvýznamnějšími národními informačními zdroji, které se tak stanou i předmětem zájmu projektu Webarchiv.

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

2.3 Nástroje pro plošnou archivaci

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, vyvinutý Helsinskou národní knihovnou, ve srovnávacích testech rozhodně nezaostává. Díky tomu, že byl navržen pro potřeby archivace webu národními knihovnami, vyhovuje nejlépe i našim požadavkům. Mezi možnosti jeho nastavení patří volba seznamu výchozích webových stránek, omezení rozsahu sklizně pomocí URL nebo jejich částí, povolení nebo zakázání podpory protokolu ftp, logování zamítnutých URL, akceptování omezení pro roboty na jednotlivých serverech (robots.txt), podpora URL s parametrem, nebo maximální hloubka zanoření v rámci jednoho serveru. Zvláště poslední dva parametry pak mohou velmi významně ovlivnit rozsah a kvalitu sklizně.

Podpora URL s parametry umožňuje omezit sklizení jen na ta URL, která neobsahují znak ‘?’ uvozující seznam parametrů. Díky tomu lze sice do značné míry zabránit problémům spojeným s nekonečnými smyčkami při procházení serverů, na druhou stranu se tak nepříjemně omezuje rozsah sklizně. Jako typický příklad lze uvést server 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 nejcenější.

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

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

3. Dlouhodobé uchování a zpřístupnění zdrojů

Problematika archivace webu zahrnuje dvě oblasti: jednou z nich je záležitost automatizovaného (plošného či výběrového) sklizení informací nacházejících se na definovaném výseku web a jejich uložení do archivu. Druhou oblast pak představuje zpřístupnění informací uložených v takto vytvořených (a objemem dat velmi rozsáhlých) archivech.

3.1 Sklizeň českého webu

V loňském roce probíhala po několik měsíců již druhá testovací sklizeň celé domény .cz, která bude po přestávce spojené s přechodem na nový server v letošním roce pokračovat. Tato sklizeň 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 různé problémy se již podařilo stáhnout z 10.490.000 URL celkem 10.090.000 souborů o celkové velikosti přes 240 GB. Alespoň jednou přitom bylo navštíveno přibližně 30.000 domén 2. úrovně (tj. čtvrtina domén v doméně .cz).

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

3.2 Provoz archivu

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

V pilotní fázi projektu bylo s výhodou využito stávajícího páskového robota Národní knihovny ČR, jehož nevýhodou ovšem je problematická dostupnost na něm uložených dat v okamžiku, kdy by bylo nutné tato data zpřístupnit veřejnosti. Protože stažené dokumenty jsou společně s příslušnými metadaty ukládány jako tar+gzip komprimované soubory přímo do souborového systému, neměl by být problém ani s migrací dat na nová úložiště.

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

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

3.3 Zpřístupnění informací v archivu

Pro zpřístupnění archivu se nabízejí technologie fulltextového indexování a automatizované extrakce autorem vytvořených metadat. Na naši zakázku byl koncem roku 2001 vypsán na MFF UK 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 – jeden z takových nástrojů, Convera Retrievalware, je v NK již zkušebně provozován. Jedním z budoucích cílů projektu bude proto pokus o jeho využití pro indexování některých netradičních typů souborů obsažených v archivu.

Nadějně se jeví též kontakty s týmem Norské národní knihovny, která vyvinula a v letošním roce se chystá dát volně k dispozici vlastní systém pro indexaci a zpřístupnění webového archivu založený na indexovacím enginu Apache Jakarta Lucene.

4. Perspektiva projektu

Zda bude některá z dosud popisovaných technologií nasazena také v ostrém reálném provozu, bude samozřejmě záviset i na vyřešení autorskoprávní problematiky související s tvorbou a provozem takového archivu. Nedotaženost zákona o povinném výtisku u nás otevírá cestu různým výkladům omezení daných zákonem o autorském právu. Automatickou identifikaci a archivaci online publikovaných dokumentů lze srovnávat s běžně používanou technologií indexování webu, jak ji provádějí Internetové prohledávače. Přesto ale není jisté, zda bude bez opory v zákoně možné využívat stávající strategii plošné archivace. Existující infrastruktura je však nastavitelná tak, že bude možné zachovat alespoň omezený rozsah sklizení i v případě, že by bylo nutné se podřídit určitým zákonným omezením. Jediným důsledkem takových omezení by pak bylo velmi výrazné zmenšení rozsahu sbírky, tvořené pak víceméně na základě dobrovolně dodávaných dokumentů. Na druhou stranu by se díky takovému zásahu výrazně zmenšila i finanční náročnost provozování celého archivu. Je

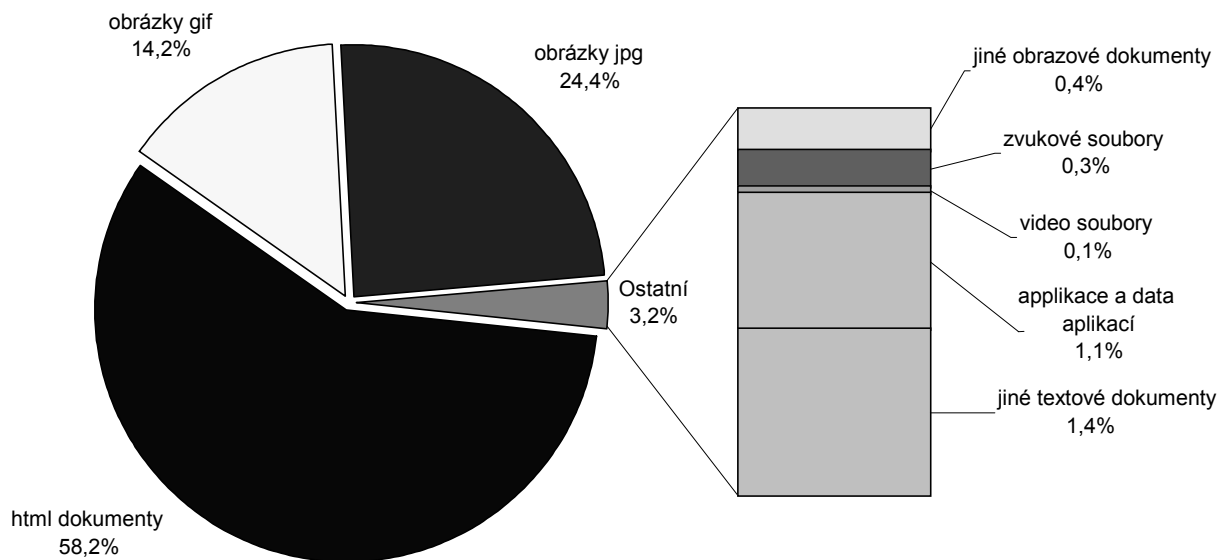
možné prohlásit, že právo občana na informace by mělo být naplněno i existencí digitální knihovny obsahující elektronicky publikované dokumenty v nezměněné podobě. Zajištění integrity takové knihovny musí být proto jedním z prioritních úkolů jejího provozovatele.

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

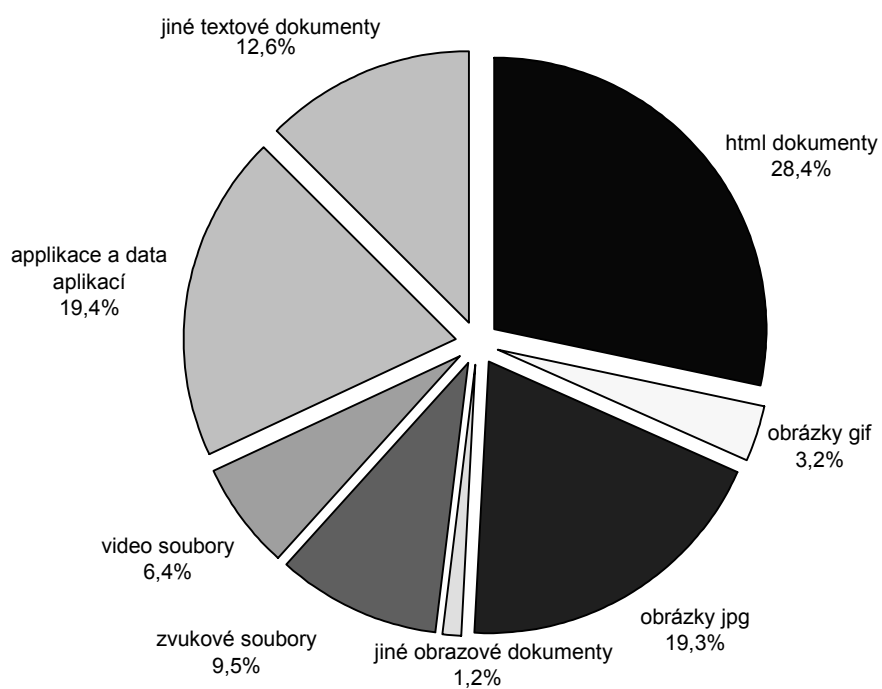
Ačkoli je díky vytvořené infrastruktuře již nyní možné udělat mnohé pro zachování dnešních informačních zdrojů pro budoucí generace, další rozvoj této infrastruktury, stejně jako vývoj v podstatě všech softwarových produktů, nemůže být nikdy zcela ukončen. Zde nejde jen o hledisko potřeb uživatele nebo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

Počet zemí, které se vážně zabývají archivací webu a elektronicky publikovaných informací stále stoupá. Úměrně tomu se zvětšují i znalosti a zkušenosti všech zúčastněných institucí. Protože většina dosažených výsledků je volně přístupná dalším zájemcům, jsou položeny i základy pro postupné vytvoření celoevropské sítě národních archivů elektronicky publikovaných dokumentů. Věřím, že jedním z nich bude brzy i archiv slovenský.

Graf 1: Relativní četnost souborů v archivu podle typů



Graf 2: Zastoupení hlavních typů souborů v archivu podle velikosti



Literatura

1. *WebArchiv* [online]. Praha : Národní knihovna ČR, posl. aktual. 16. ledna 2003 [cit. 2003-01-31]. Dostupné na World Wide Web: <<http://webarchiv.nkp.cz>>.
2. CELBOVÁ, Ludmila. *WebArchiv – vytvoření podmínek pro zpřístupnění českých webových zdrojů (knihovnické, legislativní a technické aspekty) : zpráva o plnění cílů projektu VISK3* [online]. Praha : Národní knihovna ČR, leden 2003, [cit. 2003-01-31]. Dostupné na World Wide Web: <<http://webarchiv.nkp.cz/zprava2002/zprava2002.pdf>>.
3. ŽABIČKA, Petr. Konference ECDL 2002. *Ikaros* [online]. 2002, č. 10 [cit. 2003-01-31]. ISSN 1212-5075. Dostupné na World Wide Web: <<http://www.ikaros.cz/Clanek.asp?ID=200209068>>.
4. ŽABIČKA, Petr. Webarchiv – digitální knihovna českého webu. In *RUFIS 2002*. Brno : ApS Brno, 2002, s. 121-129. ISBN 80-86510-40-9. Dostupné na World Wide Web: <http://webarchiv.nkp.cz/rufis2002_pz.pdf>.
5. ŽABIČKA, Petr. Archiv českého webu v roce 3. *Národní knihovna*. 2002, roč. 13, č. 3, s. 168-176. ISSN 0862-7487.
6. CELBOVÁ, Ludmila; ŽABIČKA, Petr. Internetové zdroje jako součást digitálních knihoven i jako součást kulturního dědictví. In *Knihovny současnosti 2002*. Brno : Sdružení knihoven ČR, 2002, s. 294-308. ISBN 80-86249-18-2.
7. ŽABIČKA, Petr. Infrastruktura Webarchivu v roce 2002. In *Inforum 2002* [online]. Praha : Albertina icome Praha, 2002 [cit. 2003-01-31]. Dostupné na World Wide Web: <http://www.inforum.cz/inforum2002/prednaska8.htm>.
8. CELBOVÁ, Ludmila. *Registrace, ochrana a zpřístupnění domácích elektronických zdrojů v síti Internet : závěrečná zpráva za léta 2000-2001* [online]. Praha : Národní knihovna ČR, leden 2002, [cit. 2003-01-31]. Dostupné na World Wide Web: <<http://webarchiv.nkp.cz/zprava2001/zprava2001.pdf>>.