

Archivace internetových zdrojů v prostředí knihoven

Ludmila Celbová, Petr Žabička

Digitální knihovna a digitální archiv

Knihovny jsou vedle archivů a muzeí označovány za paměťové instituce. V knihovnách současnosti jsou vedle klasických fondů knihovních materiálů (knihy, časopisy, zvukové záznamy, hudebniny, mapy aj.) zpřístupňovány uživatelům také elektronické dokumenty, a to nejen na fyzických nosičích, ale také elektronické dokumenty publikované v prostředí Internetu. V digitálních knihovnách nemusí tedy nutně být shromažďovány a zpřístupňovány pouze dokumenty vzniklé digitalizací klasických dokumentů (zatím se jedná většinou o digitalizaci za účelem ochrany fondů, tedy nejčastěji o digitalizaci starých knih a novin), ale také dokumenty vzniklé přímo v elektronické podobě. Na archivaci internetových zdrojů se zaměřují zejména národní a další depozitní knihovny, jejichž posláním je uchovávat národní publikační produkci.

ÚVOD

Síť Internet je informačním prostředím, v němž se dnes průběžně objevuje a mizí obrovské množství webových stránek přinášejících informace menšího i značného rozsahu, významného i méně důležitého obsahu, charakterizovaných různými formami i formáty.

S prudkým nárůstem objemu informací publikovaných výhradně na Internetu je úkolem moderní depozitní knihovny vedle péče o „tradiční“ dokumenty také 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 Moravskou zemskou knihovnou v Brně a s významným přispěním Ústavu výpočetní techniky MU buduje archiv českého webu.

ARCHIVACE INTERNETOVÝCH ZDROJŮ

Česká republika se jako první ze zemí bývalého východního bloku zařadila mezi nejvyspělejší země v oblasti řešení problémů trvalé ochrany a zpřístupnění internetových zdrojů v roce 2000, kdy Národní knihovna ČR získala grant Ministerstva kultury ČR pro řešení dvouletého pilotního projektu v rámci programu výzkumu a vývoje. Projekt zaměřený na registraci, ochranu a zpřístupnění domácích elektronických zdrojů v síti Internet dostal pracovní název WebArchiv. V rámci menších grantových podpor Ministerstva kultury ČR se daří projekt i nadále rozvíjet a postupně připravovat podmínky pro zavádění výsledků výzkumu do praxe.

Problém trvalého uchování národního bohatství v podobě elektronických publikací, zejména síťových, už přestává být experimentem „pokročilejších“ zemí a stává se obecně naléhavou výzvou pro knihovny i nakladatele. Mnohé z elektronických zdrojů, které neexistují souběžně v tradiční (tištěné nebo analogové) formě (tzv. digital born dokumenty), byly již trvale ztraceny, neboť jejich tvůrci nebo vydavatelé odstranili své elektronické publikace z webu, aniž by zajistili jejich trvalou archivaci. Problematikou ochrany digitálního dědictví se proto v rámci programu Paměť světa (Memory of the World) začala zabývat i organizace UNESCO.

Jedním z průkopníků na poli archivace webu je americká nezisková organizace Internet Archive (www.archive.org), jejíž archiv sahá až do roku 1996. Tato organizace se ve spolupráci s dalšími institucemi snaží (vcelku úspěšně) vybudovat co nejrozsáhlejší archiv světového webu. Takový záměr je však finančně vysoce nákladný a proto spolupracuje s největšími světovými národními knihovnami na vývoji nové generace volně dostupných nástrojů pro archivaci a zpřístupnění webových informačních zdrojů.

Je zřejmé, že každá knihovna nemá prostředky na to, aby si vytvářela vlastní archiv celého světového webu, zároveň ale není možné spoléhat se výhradně na vydavatele

elektronických informačních zdrojů, kteří mohou publikované dokumenty libovolně modifikovat nebo zcela odstranit. Je proto logické, že se každá vyspělá země snaží (většinou prostřednictvím národní knihovny daného státu) přednostně vybudovat národní archiv elektronických informačních zdrojů.

Přístup jednotlivých knihoven k řešení problému se ovšem velmi liší. Některé knihovny, jako například Australská národní knihovna, se snaží archivovat výběrově jen ty webové zdroje, jejichž kvalitu předem zhodnotí knihovník (pandora.nla.gov.au). Díky tomuto přístupu čítá sice nyní archiv australského webu po několika letech provozu jen cca 5800 webových sídel nebo jejich částí, nicméně jde o výběr toho „nejdůležitějšího“, co bylo na australském webu publikováno. Tento přístup je však velmi náročný na lidské kapacity a proto se většina knihoven vydala cestou automatizované plošné archivace všech dokumentů, které splňují automaticky vyhodnotitelná kritéria.

ARCHIVACE INTERNETOVÝCH ZDROJŮ V ČESKÉ REPUBLICE

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

Legislativní problematika

Kritéria výběru zdrojů a strategie jejich archivace

Bibliografická správa a zpřístupnění zdrojů

Legislativa

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

Pokud jde o zákon o povinném výtisku, resp. o dva zákony týkající se neperiodických publikací a vydávání periodického tisku, jejich znění není dostačující k tomu, aby české depozitní knihovny mohly tuto legislativu uplatňovat pro získávání elektronických publikací do svého fondu (digitálního archivu). Nicméně podle znění Autorského zákona „Do práva autorského nezasahuje knihovna, archiv a jiné nevýdělečné školské, vzdělávací a kulturní zařízení, zhotoví-li rozmnoženinu díla pro své archivní a konzervační účely.“ Vytváření archivu, tzn. stahování a ukládání elektronických online zdrojů v rámci WebArchivu je tedy z hlediska autorskoprávního legální.

Problematické z hlediska autorských práv je ovšem zpřístupnění uložených zdrojů z digitálního archivu, kdy ani z tohoto konzervačního fondu nemá knihovna právo zpřístupnit archivované zdroje bez souhlasu autora, resp. vydavatele. Národní knihovna ČR proto zatím uplatňuje náhradní řešení, které umožňuje zpřístupňování vybraných archivovaných zdrojů na základě smluvního souhlasu vydavatele. Ve smlouvě o poskytování elektronických online zdrojů souhlasí vydavatel mj. se zpřístupněním svých zdrojů uložených v archivu (varianta lokálního nebo „veřejného“ zpřístupnění) a současně s tím, že zajistí informování autorů o této skutečnosti. Není to však řešení ideální, poněvadž je časově a organizačně náročné.

Kritéria výběru zdrojů a strategie jejich archivace

Stanovení podmínek, které musí splňovat elektronické zdroje kandidující na včlenění do budovaného digitálního archivu, je jedním z nejkritičtějších okamžiků každého podobného projektu. Při stanovování těchto podmínek je nutné brát v úvahu jak objem finančních prostředků, které jsou pro tuto činnost k dispozici, tak i aktuální stav rozvoje celé oblasti informačních a komunikačních technologií. To představuje dvě souběžné linie:

shromažďování, registraci, archivaci a zpřístupňování vybraných domácích elektronických online dostupných zdrojů jako legitimní součásti národní publikační produkce podle stanovených kritérií výběru pro účely České národní bibliografie,

automatizovaný proces shromažďování a archivace domácích zdrojů z Internetu v relativní úplnosti (nyní servery v doméně .cz).

Výběr nejvýznamnějších zdrojů jako intelektuální činnost

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

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

Místo uložení zdroje

Primárně jsou brány v úvahu zdroje přístupné na serverech s doménou prvního stupně .cz. V této souvislosti však vyvstává problém, jak správně vymezit tzv. národní web (tj. zda uplatňovat pouze teritoriální hledisko nebo také jazykové hledisko podobně, jako je tomu u tradičních bohemikálních dokumentů). Faktem zůstává, že není možné výše uvedenou podmínku za všech okolností striktně dodržet, protože v některých případech čeští vydavatelé záměrně nebo nuceně (obvykle z důvodu předchozí registrace žádané domény ze strany spekulantů) využívají servery s doménami .com, .net, .org a další.

Obsah

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

Typ a forma

Repertoár typů zdrojů je poměrně pestrý a je do jisté míry ovlivněn předchozím kritériem. Při jejich výběru se vychází z běžných klasifikací dokumentů. Jde především o seriály, konferenční příspěvky, výzkumné a jiné zprávy, studie vzniklé např. jako výstupy vědeckých a výzkumných projektů, akademické práce, dokumenty veřejné správy. Je příznačné, že tyto zdroje spadají do kategorie tzv. šedé literatury. Jsou však brány v úvahu jen volně dostupné zdroje, které jsou publikovány pouze v elektronické formě (online), aby se zabránilo duplicitě zpracování webových zdrojů a tradičních (tištěných) dokumentů s identickým obsahem.

Formát

Z pragmatických důvodů jsou preferovány formáty, které jsou všeobecně podporovány producenty aplikačního softwaru (zejména webových prohlížečů), nikoliv tedy proprietární formáty, pro jejichž korektní zobrazení je třeba zvláštní aplikační software.

Automatizovaný proces shromažďování a archivace

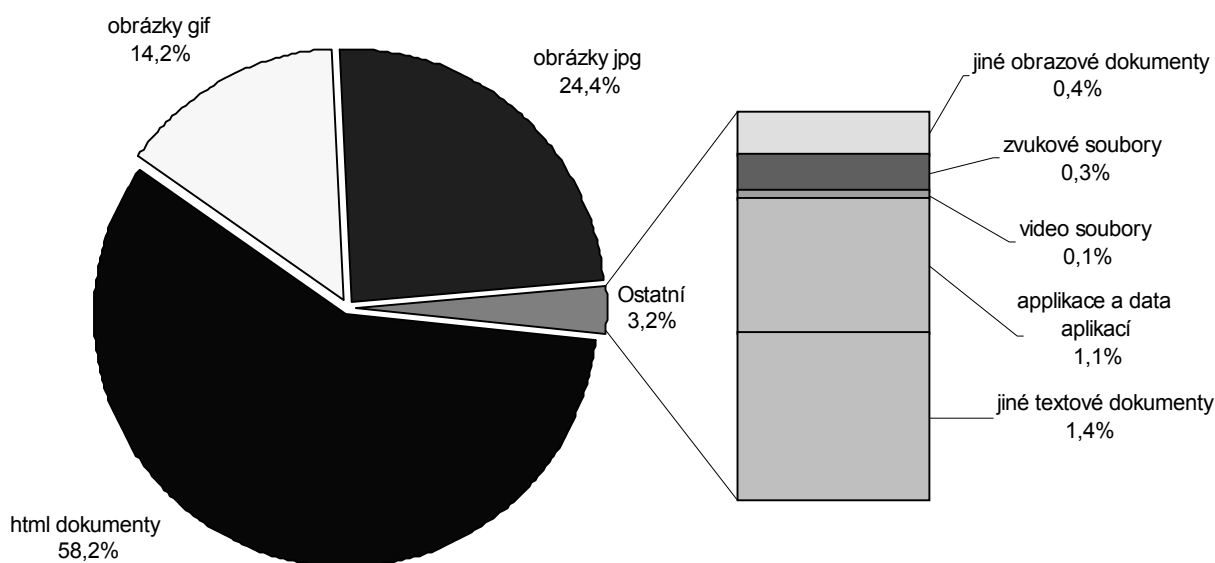
V případě automatizovaného sběru webových zdrojů se často hovoří o „plošném sklizení“ českého webu. Ani v tomto případě se neshromažďuje absolutně vše, co si lze představit pod českou produkcí online zdrojů, ale i zde jsou četné aspekty ovlivňující sběr a archivaci, tedy opět jakási kritéria výběru.

Protokoly, formáty

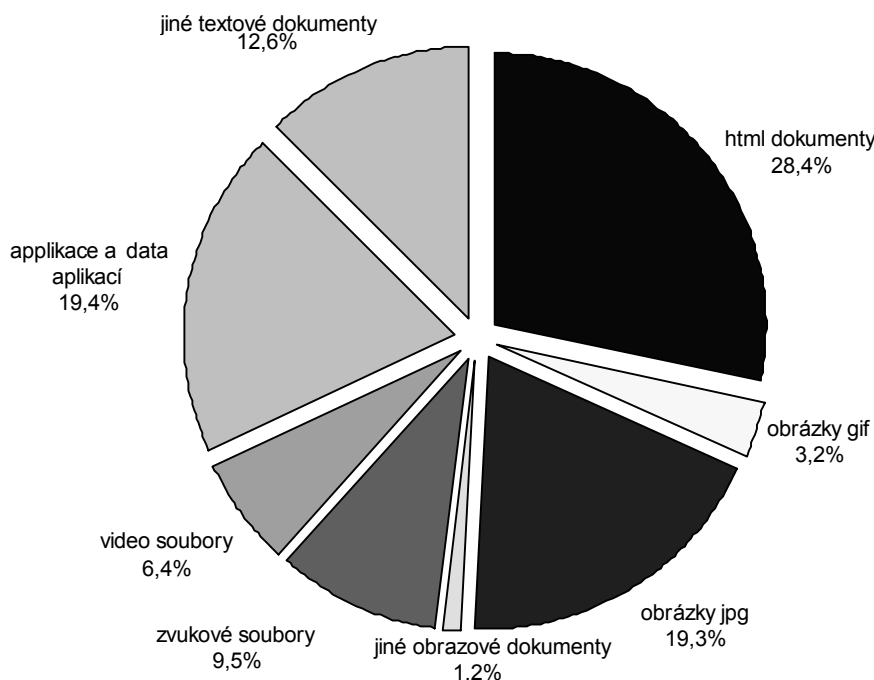
Je zřejmé, že pokus archivovat online elektronické zdroje, dostupné jinak než prostřednictvím Internetu, by byl velmi nákladný a jeho přínos by byl mizivý. Proto lze zatím stále obhájit názor, že většině populace je reálně přístupná jen ta část zdrojů, ke kterým se dostanou prostřednictvím běžného prohlížeče, tedy především prostřednictvím protokolu http.

Podobně jako v případě protokolů bychom mohli jednotlivé dokumenty hodnotit i co do použitého formátu. Grafy 1 a 2 ukazují, jak jsou v současnosti v našem archivu zastoupeny jednotlivé formáty souborů. Je vidět, že trojice formátů html, jpg a gif tvoří dohromady 97% počtu všech archivovaných souborů, ačkoli co do velikosti zauímají jen polovinu celkového objemu uložených dat. Pokud tedy dokážeme odpovědně určit, které z vzácně se vyskytujících formátů nemá smysl z různých důvodů archivovat, můžeme snadno ušetřit až třetinu objemu ukládacího prostoru, což může perspektivně představovat úsporu statisícových částek.

Graf 1: Relativní četnost souborů v archivu podle typů



Graf 2: Zastoupení hlavních typů souborů v archivu podle velikosti



Počáteční odkazy

Zkušenosti s celoplošným sklizením národních domén ukazují, že nejvhodnější strategií pro započítání sklizení je zajištění co největšího počtu adres webových stránek, ze kterých vedou další odkazy. Pro získání seznamu takových odkazů je nejvhodnější najít mezi českými doménami takové, na nichž je spuštěný www server a ty pak použít jako výchozí body pro sklizení. Tento seznam se dá dále zdokonalit vyhledáním běžících www serverů na dalších registrovaných serverech v doménách třetí úrovně.

Testováním v loňském roce bylo zjištěno, že ze 140.000 registrovaných českých domén je www server v provozu na 120.000 serverech typu *www.doména.cz* a na 76.000 serverech typu *doména.cz*. Port používaný protokolem http je ale aktivní na celkem 393.000 adresách 2. nebo 3. úrovně. Adresy všech těchto téměř 400.000 serverů byly použity jako výchozí body pro nyní probíhající plošné sklizení českého webu. To běží od 10.3.2004 a bylo při něm již staženo přibližně 13.1 milionu dokumentů z celkem 15 milionů různých adres (některé dokumenty byly přístupné na více adresách). Zatím tak bylo v tomto kole sklizeny 516 GB dat (před kompresí).

Vliv technického řešení na rozsah a průběh sklizení

U nás používaný produkt, NEDLIB Harvester, vyvinutý Helsinskou národní knihovnou, byl navržen pro potřeby archivace webu národními knihovnami a z dostupných nástrojů zatím vyhovuje našim požadavkům nejlépe. Tento nástroj má však i jisté nevýhody, například nepodporuje javascript ani flash technologie, nedokáže extrahovat odkazy z netypických formátů dokumentů a navíc byl jeho vývoj již v roce 2001 ukončen. Přestože je NEDLIB Harvester stále používán pro archivaci webu v mnoha zemích, jeho éra se tak nezadržitelně chýlí ke konci. Potvrzením tohoto faktu je nedávné zveřejnění ranných verzí volně dostupného produktu HERITRIX, vyvíjeného konsorciem velkých národních knihoven a organizace Internet Archive.

Celkový objem komprimovaných dat archivovaných v rámci projektu WebArchiv již přesahuje hranici 700 GB a bude dále růst. V pilotní fázi projektu bylo pro uložení archivu využito stávajícího páskového robota Národní knihovny ČR, jehož nevýhodou ovšem byla problematická dostupnost na něm uložených dat. Proto byl celý webový archiv později přesunut na vyhrazenou část diskového pole o kapacitě 2TB, ke kterému má řešitelský tým snazší přístup. I toto diskové pole má však omezenou kapacitu a je možné, že právě probíhající sklizeň je zaplní do té míry, že na ně již nebude možné další sklizeň podobného objemu uložit. Řešením tohoto problému by bylo pořízení samostatného diskového pole s větší kapacitou v příštím nebo nejpozději přespříštím roce, kdy se dá očekávat cenová dostupnost diskových polí o kapacitě řádově 5TB v cenách srovnatelných se současnými 2TB poli.

Jednou z cest, jak snížit objem archivu je omezení rozsahu archivovaných formátů. 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 vytvoření takového archivu je sice důležitým, ale zároveň jen prvním krokem na cestě k naplnění jeho smyslu, tedy ke zpřístupnění jeho obsahu.

Bibliografická správa a zpřístupnění zdrojů

Dokud nebude jasně stanoveno, kdy, komu, v jakém rozsahu a za jakých podmínek může být takový archiv zpřístupňován, není možné navrhnout optimální nástroj pro daný účel. Proto se zatím veškerá práce upírá na zpřístupnění těch dokumentů, které mají řešitelské deponitní instituce právo zpřístupnit, tj. jejichž volné poskytování je umožněno smluvně

Fulltext

Koncem roku 2001 byl na MFF UK vypsán ročníkový týmový vývojový projekt na vytvoření indexační a vyhledávací aplikace pro Webarchiv. V červnu roku 2003 byl tento program dokončen a nyní je provozován na serveru projektu WebArchiv. Tento program sice indexuje pouze html soubory, ale je otevřen i dalším formátům. Program zohledňuje českou gramatiku i diakritiku. Podporuje 9 vyhledávacích kategorií včetně fulltextu, zvýrazněného textu a metadat.

Nadějně se jeví i výsledek projektu Nordic Web Archive, volně dostupný nástroj NWA Toolset, který vyvinula v rámci mezinárodní spolupráce Norská národní knihovna. Tento systém je již také zprovozněn na serveru projektu.

Oba tyto systémy začnou v nejbližší době indexovat a veřejně zpřístupňovat stejnou množinu smluvně krytých dokumentů, aby bylo možné porovnat a vyhodnotit jejich vlastnosti. Již nyní je ale zřejmé, že každý z těchto systémů má své přednosti. Zatímco v případě norského systému je kladem elegantní časová osa, výhodou českého systému je například dobrá podpora češtiny.

Metadata

V procesu integrace elektronických zdrojů do stávající knihovní informační infrastruktury však není možné spoléhat se jen na fulltextové prohledávání. Řešením tohoto problému je spolupráce s vydavateli v oblasti pořizování a zpřístupňování metadat svázaných s jednotlivými dokumenty. Proto byla již v rámci pilotního projektu vybudována infrastruktura, zaměřená na podporu využívání metadat DC u nás. Tato infrastruktura by měla usnadnit zapojení autorů a vydavatelů do procesu tvorby a zveřejňování metadat již v okamžiku publikování dokumentu.

Nejdůležitější částí této infrastruktury je Dublin Core Metadata Generator, doplněný o generátor jednoznačného identifikátoru čísla národní bibliografie. Tento nástroj, veřejně přístupný na serveru projektu, umožňuje autorům webových stránek poloautomaticky nebo ručně vytvořit, editovat, konvertovat a ve zvolené syntaxi uložit metadata respektující pravidla kvalifikovaného Dublin Core (http://www.ics.muni.cz/dublin_core/).

Jakmile jsou webové stránky obsahující metadata archivovány, mohou být tato metadata využita jednak lokálně, pro usnadnění vyhledávání v archivu a jednak po zpřístupnění prostřednictvím protokolu OAI-PMH plánovaném na letošní rok i pro agregované prohledávání například prostřednictvím Jednotné informační brány (www.jib.cz).

Mimo popisu jednotlivých dokumentů již nyní existuje v rámci knihovního systému Aleph Národní knihovny ČR i databáze obsahující záznamy webových elektronických zdrojů vybraných na základě výše selekčních kritérií obdobných kritériím výběru tradičních druhů dokumentů pro konzervační fond. Výběr zdrojů je zaměřen na jejich trvalé uchování jako kulturního dědictví a současně k účelu jejich registrace v České národní bibliografii. V rámci silně limitovaných pracovních kapacit na tuto činnost je proto snahou řešitelů zařadit do souboru vybraných zdrojů pokud možno ty, které si z hlediska svého obsahu a/nebo formy zařazení mezi „kulturní dědictví národa“ skutečně zaslouží. Záznamy jsou samozřejmě ve formátu MARC (nyní UNIMARC, od léta 2004 MARC21).

Vedle funkce bibliografické slouží báze také zpřístupňování zdrojů popsanych v databázi bibliografickým záznamem. Prostřednictvím zapsané URL adresy je umožněn přístup ze záznamu přímo do zdroje přístupného na internetu. Ty zdroje, u nichž to dovoluje smlouva, budou zanedlouho prolinkovány i přímo do webového archivu.

ZÁVĚR

Jaká je perspektiva projektu?

To závisí i na úsilí, které bude této problematice věnovat širší komunita paměťových institucí, které nezávisle na svém typu musí řešit určité otázky spojené se získáváním, registrací, dlouhodobým uchováváním a zpřístupňováním elektronických informačních zdrojů.

Role jednotlivých institucí snad ukáže blízká budoucnost. Již nyní se ale rýsuje jasná role archivů jako důležitého partnera při stanovování budoucích strategií pro dlouhodobé uchovávání dokumentů (migrace, emulace, konverze), stejná problematika se ale dozajista týká i muzeí.

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 a zachování dlouhodobé paměti lidstva.

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