Archivace elektronických zdrojů v ČR a jejich registrace v české národní bibliografii (Srovnání s výsledky přehledu IFLA)

Ludmila Celbová
Národní knihovna ČR
Ludmila.Celbova@nkp.cz

Jak vyplývá z výše uvedené analýzy, průzkumu IFLA se zúčastnila i Česká republika, respektive Národní knihovna ČR, která ve spolupráci s Moravskou zemskou knihovnou a Ústavem výpočetní techniky Masarykovy univerzity v Brně tuto problematiku řeší.

Co se v České republice daří a co je brzdou pro archivaci „českého webu"?

Projekt svým způsobem bojuje rok od roku o přežití. Přestože jde o obrovské objemy dat k uložení do archivu a o náročné činnosti související s vývojem a aplikací softwaru a další intelektuální práce, je každým rokem téměř absolutně závislý na skrovném objemu grantových prostředků, tzn. po celou dobu řešení je financován mimo rozpočet NK ČR. I přes tvrdé podmínky finanční, personální i prostorové, které dosud měli řešitelé k dispozici, se Česká republika v oblasti archivace svého „národního" webu, v problematice trvalého uchování této dnes již podstatné části publikační produkce jako svého kulturního a historického dědictví, zařadila k nejvyspělejším zemím. Tím je ovšem míněna stránka znalostí, kvality práce v oblasti souvisejících informačních technologií (sklízení webu, indexace, archivace, zpřístupnění) a v oblasti výběru a popisu zdrojů. Projekt je na solidní úrovni experimentální (výzkum a testování). Bohužel kvantitativní stránka se odvíjí od nedostatku financí – na výpočetní techniku, na personální zabezpečení a v této souvislosti též na řešení legislativy. Tím nelze přejít z etapy pilotního projektu do praktického provozu.

Současná situace

V oblasti *informačních technologií* se od počátku řešení problematiky archivace webu průběžně testují a přizpůsobují volně dostupné SW nástroje s otevřeným zdrojovým kódem. Pokud jde o HW, začínali jsme de facto s jedním větším PC v roli serveru (stahování) a robotickými páskami pro ukládání dat. V současné době pracujeme se dvěma zastaralými servery, které jsou téměř na hranici použitelnosti – jeden je využíván pro archivaci, druhý pro zpřístupnění. Letos přibylo diskové pole pro ukládání dat s využitelnou kapacitou cca 5 TB. Objem dat uložených v archivu je cca 2 TB, což představuje asi 26 mil. unikátních dokumentů. O celou oblast IT – vývoj, programování, správa systémů – se převážně starají studenti.

Výběr a popis dokumentů jsou intelektuální práce a jejich stávající objem je maximem, jakého lze při současném personálním obsazení dosáhnout. Výběr se provádí v souladu se stanovenými kritérii výběru (http://www.webarchiv.cz/kriteria.html). Kvůli neexistující legislativě pro povinný výtisk on-line zdrojů a stávajícímu autorskému zákonu se zvolilo – obdobně jako v jiných zemích – náhradní řešení: oslovování jednotlivých vydavatelů a uzavírání smluv o poskytování elektronických on-line zdrojů významných z hlediska historického, kulturního či

odborného. Smlouvy umožňují Národní knihovně ČR uchovávat a zpřístupňovat „konzervační fond“ on-line zdrojů přes webové rozhraní projektu. Jde však o zdlouhavý a ne příliš efektivní proces. „Nasmlouvané“ zdroje obsahují metadatový záznam ve formátu kvalifikovaného Dublin Core. Bibliograficky jsou popisovány v samostatné bázi WEB elektronického katalogu NK ČR ve formě úplných záznamů ve formátu MARC21. Ze záznamů v katalogu NK je umožněn přímý přístup do zdroje na aktuálním umístění na webu, eventuálně do zdroje uloženého v digitálním archivu. V současné době má NK ČR uzavřeno na 50 smluv o poskytování elektronických on-line zdrojů. Seznam spolupracujících vydavatelů a jejich zdrojů lze najít na webových stránkách projektu: http://www.webarchiv.cz/partneri.html.

Legislativa je bohužel silnou brzdou archivace webu, zejména pokud jde o přístup veřejnosti k archivovaným datům. Pokud jde o povinný výtisk monografií (zákon č. 37/1995 Sb., o neperiodických publikacích), je zákon natolik obecný, že jej lze aplikovat i pro on-line zdroje – monografických publikací ale na internetu bohužel v dnešní době mnoho nenajdeme. Naprostá většina sklízených webových zdrojů jsou z hlediska typu publikací seriály či integrační zdroje (průběžně aktualizované); pro tyto zdroje je tiskový zákon (zákon č. 46/2000 Sb.), jehož součástí je ustanovení týkající se povinnosti vydavatelů odevzdávat povinný výtisk periodických tiskovin, zcela nepoužitelný. Legislativní zázemí sběru dat z webu (vytváření digitálního archivu) nacházíme ve stávající verzi autorského zákona (zákon č. 121/2000 Sb.), umožňující knihovně zhotovit rozmnoženinu díla pro své archivní a konzervační účely; pokud jde o právo zpřístupnit data z digitálního archivu (konzervačního fondu), čekáme na novelu autorského zákona (předpoklad přijetí v polovině roku 2006), v rámci níž by mělo být schváleno ustanovení umožňující lokální zpřístupnění. Toto ustanovení je v souladu s evropskou Směrnicí o harmonizaci některých aspektů autorského práva a práv s ním souvisejících v informační společnosti (2001/29/ES). Tato směrnice v jednom ze svých článků doporučuje vládám členských států, aby umožnily zpřístupňování autorských děl (včetně jejich on-line podoby), která má knihovna ve svých sbírkách, na vyčleněných terminálech ve svých prostorách jednotlivým členům veřejnosti za účelem výzkumu nebo soukromého studia. Projednání v Parlamentu by tedy mělo být bezproblémové. Na veřejný přístup on-line není ovšem v dohledné době z legislativních důvodů šance. K potřebné „modernizaci“ legislativy týkající se povinného výtisku sbíráme v současné době podklady.

Srovnání a perspektiva

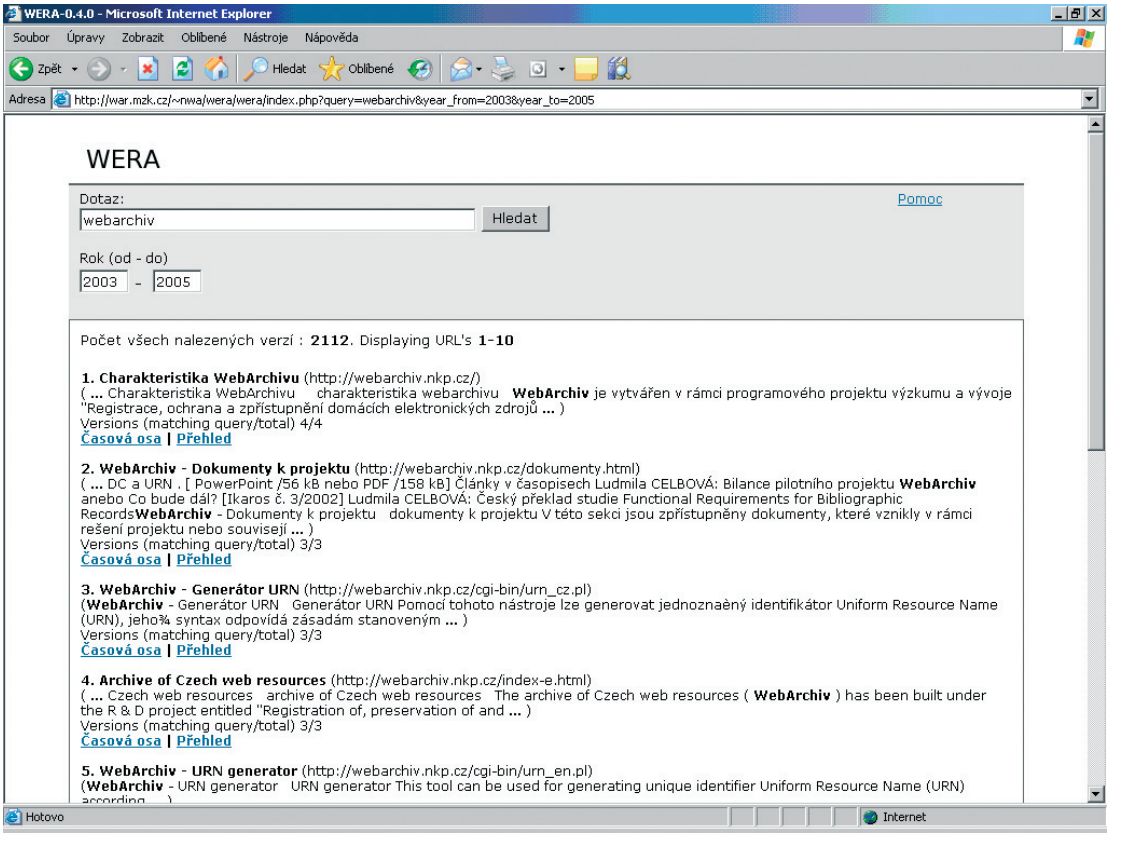
Vzhledem k tomu, že jde o relativně nové úkoly národních knihoven a současně velmi náročné na financování a lidské zdroje, není situace růžová nikde ve světě. Nicméně, zejména v zemích, které byly průkopníky řešení archivace internetových zdrojů, si již vlády uvědomují důležitost této problematiky. Příklady: Americký Kongres vyčlenil v prosinci 2000 pro Kongresovou knihovnu jako koordinátora Programu na ochranu digitálních dokumentů 100 miliónů dolarů. V Německu byla v loňském roce v rámci nového knihovnického zákona samostatně řešena problematika ochrany elektronických on-line dokumentů – vyčleněno 1,9 mil. eur do roku 2007, dalších 2,9 mil. eur do roku 2011; současně řešeno personální nasazení (nárůst do roku 2011 na 28 osob) a legislativa k povinnému výtisku. V Dánsku je projekt netarchive.dk (spolupráce The Royal Library, Kopenhagen a The State & University Library, Aarhus) finančně zajištěn částkou 400 000 euro ročně; od poloviny roku 2005 vstoupil v platnost nový zákon o povinném výtisku, který povoluje oběma kooperujícím knihovnám sběr kompletního obsahu dánského webu. Také Litva uvádí sumu blížící se 100 000 eur na rozvojový program a každoroční investici téměř 30 000 eur na technické vybavení. V členských zemích konsorcia International Internet

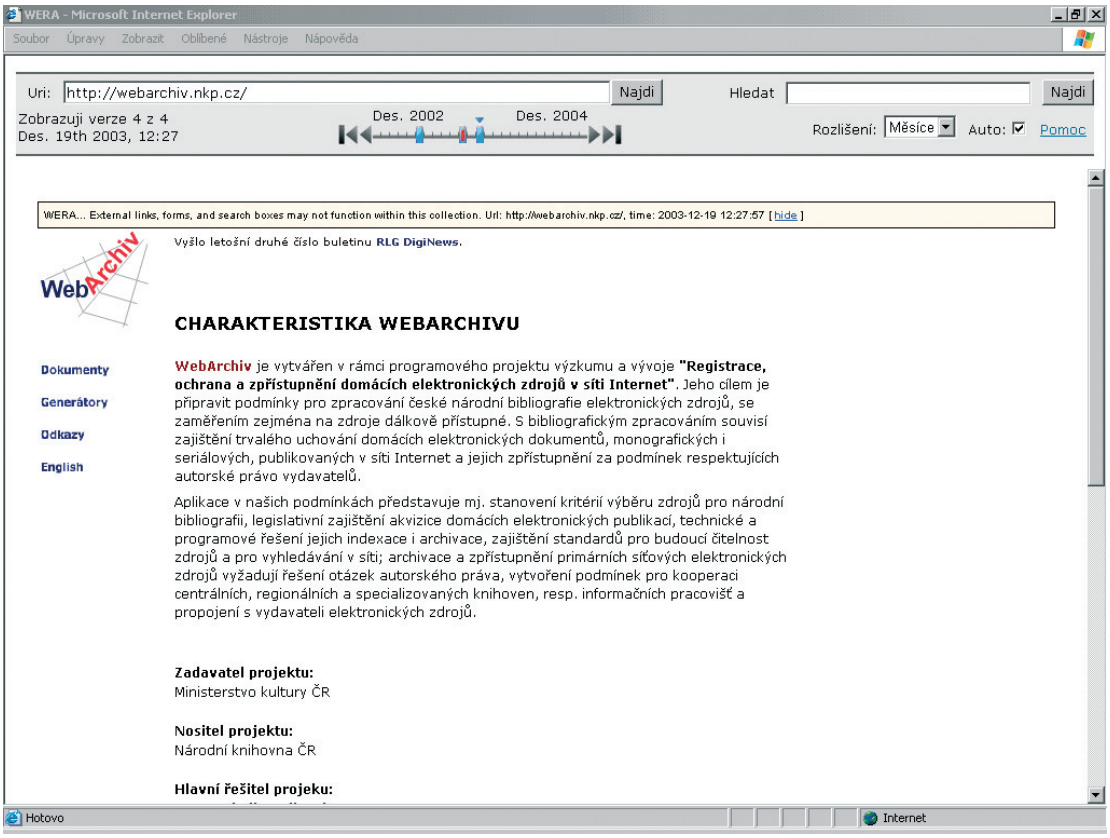
reservation Consortium (Francie, Norsko, Austrálie aj.) je silně podporován vývoj softwarových nástrojů, které jsou (zatím) poskytovány jako volně dostupné SW nástroje s otevřeným zdrojovým kódem. Jak ukazuje výše zmíněný průzkum, už v jedenácti evropských zemích je legislativně řešen povinný výtisk pro elektronické on-line zdroje.

Národní knihovna ČR připravuje v současné době Koncepci rozvoje trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010, jejímž úkolem má být: „Vytvořit legislativní, organizační a technické předpoklady pro shromažďování, trvalé uchování a zpřístupnění publikovaných digitálních a digitalizovaných dokumentů jako důležité složky kulturního dědictví.“ Koncepce této digitální knihovny, samozřejmě s vyčíslením potřeby finančního zabezpečení úkolů, bude předána koncem tohoto roku na Ministerstvo kultury ČR. Podaří se také v České republice přesvědčit vládu a parlament o nezbytnosti řešení ochrany digitálního dědictví?

Informace k archivaci českého webu jsou dostupné na serveru http://www.webarchiv.cz. Z této stránky se uživatel dostane také přímo do digitálního archivu – vyhledávat může ovšem pouze v té části archivu, jež obsahuje zdroje, na které má NK ČR uzavřenu s vydavateli smlouvu o poskytování elektronických on-line zdrojů. Přístup do báze WEB elektronického katalogu NK ČR je na URL: http://sigma.nkp.cz/cze/web.

Ukázka fulltextového vyhledávače WERA (Web aRchive Access), který je nyní využíván pro zpřístupnění archivovaných dokumentů. Systém umožňuje mimo jiné fulltextové vyhledávání a zohledňuje změny (verze) dokumentů v čase.





Závěrem

Česká republika se v oblasti archivace svého „národního“ webu, v problematice trvalého uchování této dnes již podstatné části publikační produkce jako svého kulturního a historického dědictví, zařadila k vyspělejším zemím, které se větší či menší měrou začaly touto problematikou zabývat. Důležité je, že Česká republika nezaspala, že uchování významných webových dokumentů pro (nejen) budoucí generace je alespoň v podobě jakýchsi vzorků zajištěno.

V uplynulých pěti letech řešitelé na základě mnoha analýz stanovili metodiku práce s využitím mezinárodních standardů, otestovali různé SW nástroje pro sběr, indexaci, ukládání i zpřístupňování těchto dokumentů, položili základ pro tvorbu národní bibliografie on-line zdrojů, navázali kontakty se zahraničními partnery aj. Nyní nastává již čas pro vytvoření takových podmínek, které by umožnily praktický provoz jak v oblasti IT, tj. nekomplikovaný sběr a ukládání dat, tak souběžně výběrové zpracování webových zdrojů a vzájemné propojení těchto činností. A následně zpřístupnění v rámci podmínek legislativních. Pro úspěšné fungování projektu WebArchiv je nyní v první řadě nutné zajistit takové personální a finanční zabezpečení projektu, aby bylo možné přejít postupně z fáze testování do praktického a rutinního provozu.