

Archiv českého webu jako prostředek zajištění svobodného přístupu občanů k informacím

Petr Žabička, Moravská zemská knihovna a Fakulta informatiky Masarykovy Univerzity Brno
Ludmila Celbová, Národní knihovna ČR, Praha

S prudkým nárůstem objemu informací publikovaných výhradně na Internetu by se úkolem moderní depozitní knihovny mělo stát 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. V tomto konsorciu se nyní rozhodly spojit své síly americká Kongresová knihovna, Britská knihovna, Francouzská národní knihovna a některé severské národní knihovny.

Z celosvětového hlediska je však přístup jednotlivých národních knihoven k řešení problému archivace webu velmi různorodý. 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 vyhodnotitelná kritéria. K tomu využívají nejčastěji softwarových nástrojů vyvinutých v minulých letech v rámci evropských projektů nebo projektů evropských severských zemí. Pozadu nezůstává ani Japonská národní knihovna a zahájen byl i projekt na archivaci webových zdrojů v čínštině.

Podobným směrem se v roce 2000 vydala i Národní knihovna ČR, když v 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).

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.

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.

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.

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.

3. Dlouhodobé uchování a zpřístupnění zdrojů

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).

Z hlediska dlouhodobé dostupnosti informací budou největším oříškem samotné archivované dokumenty. 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.

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. 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 prohlížeče. S tím, jak bude stoupat podíl čistě elektronické produkce, bude růst i význam její dlouhodobé archivace z hlediska ochrany národního kulturního dědictví. Přesto ale není jisté, zda bude bez opory v zákoně možné využívat stávající strategii plošné archivace.

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 účelně vynaložit, je spolupráce na meziřesortní i mezinárodní úrovni, která se velmi osvědčila již během řešení pilotního projektu.

Záležitosti archivace digitálních dokumentů se však netýkají pouze knihoven v souvislosti se zajištěním trvalého přístupu k národnímu kulturnímu bohatství. Problematiku jinou obsahem, ale obdobnou technicky (příp. i legislativně) budou muset řešit např. archivy, muzea, ale i vládní a správní orgány. Zdá se, že je nejvyšší čas, aby se problematika archivace internetových zdrojů dostala k řešení na vyšší instanci, tedy na úroveň státních orgánů odpovídajících za informační politiku a za zajištění svobodného přístupu občanů k informacím.