Budoucnost českého webového archivu

Jaroslav Kvasnica
Národní knihovna České republiky
jaroslav.kvasnica@nkp.cz

INFORUM 2015: 21. ročník konference o profesionálních informačních zdrojích Praha, 26. - 27. 5. 2015

Abstrakt

V prezentaci bude představen český webový archiv (webarchiv.cz), který se stará o dlouhodobou ochranu českých digitálních online zdrojů. Bude představeno, jakým způsobem dnes webový archiv funguje, jak probíhá akvizice zdrojů a jakým způsobem jsou data uchovávána. Prezentace ale bude primárně zaměřena na nově vznikající způsoby zpřístupnění dat v archivu pro potencionální uživatele. Jelikož se domníváme, že pouhé vyhledávání pomocí URL nebo klíčových slov není dostatečné, pro tak velký a specifický objem dat, kterým webový archiv je.

V prezentaci hodláme představit naši ideu zpřístupněni datových setů, které budou výstupem analýz nad velkými daty v archivu. Zaměříme se také na technologickou část věci a představíme clusterovou technologii Hadoop a HBase, které jsou nezbytnými nástroji pro práci s tzv. big daty. Hlavním cílem českého webového archivu je do budoucnosti motivovat výzkumné pracovníky z nejrůznějších oborů, poskytnutím dat, nástrojů a podpory a provádět výzkum nad unikátními velkými daty webového archivu.

Výrazné rozšíření internetu od jeho vzniku na počátku 90. let minulého století vedlo k enormnímu nárůstu elektronického publikování a mnohé dokumenty dnes vznikají již pouze v elektronické podobě. Nejen publikování, ale veškerá lidská kulturní produkce se alespoň částečně přesunula do prostředí internetu. Vzhledem k dynamické povaze webu každý den narůstá počet webových stránek a další obrovské množství stránek zaniká, mění svou podobu, obsah nebo adresu. Z toho důvodu může být současné kulturní, umělecké a historické hodnoty ztraceny pro další generace.

Archivací webu se zabývají především instituce zodpovědné za uchovávání kulturního dědictví, zejména národní knihovny. Cílem archivace webu je výběr, uchování a zpřístupnění webových dokumentů, tj. budování trvale přístupné kolekce digitálních zdrojů.

Webové archivy přispívají k zachování kulturního dědictví určitého regionu v době, kdy množství informací vzniká přímo v elektronické podobě. K této snaze se přidala i Národní knihovna ČR, která má poslání se podílet se na uchování a zpřístupňování kulturního dědictví současníkům i budoucím generacím.

Webarchiv Národní knihovny ČR je digitální knihovna českých elektronických online zdrojů. První stránky byly archivovány v roce 2001, pravidelná archivace pak probíhá od roku 2005. Od roku 2007 je Webarchiv členem mezinárodního konsorcia pro archivaci

webu IIPC (*International Internet Preservation Consortium*). Webarchiv je také součástí projektu *Národní digitální knihovna*.

Historie

Český webový archiv, který nese jméno Webarchiv začal jako projekt v roce 2000. První zachovaný dokument v archivu je z roku 2001. Dnes má webový archiv uloženo přes 200 TB komprimovaných dat. Tyto data pocházejí z pravidelného sklízení české internetu, které započalo v roce 2005.

Akvizice zdrojů probíhá pomocí automatizovaného softwaru tzv. sklízeče (ang. harvester), který prochází jednotlivé stránky a stahuje jejich obsah. Tato činnost se nazývá sklízení (harvesting) a výsledkem této činnosti jsou tzv. sklizně, které představují jeden časově ohraničený proces stahování a sběru dat. Český webový archiv rozlišuje tři typy těchto sklizní podle toho, jakým způsobem jsou sklízeči dodány URL odkazy.

Celoplošná sklizeň pokrývá webové zdroje s národní doménou .cz. Seznam těchto zdrojů je dodáván správcem domény, sdružením CZ.NIC. Tato celoplošná sklizeň je prováděna zpravidla jednou ročně a takto archivované stránky jsou z důvodu prostorových kapacit sklízeny pouze do určité úrovně. Cílem celoplošných sklizní je zachycení obrazu českého internetu v daném čase.

V*ýběrové sklizně* pokrývají pouze vybrané zdroje, ale na rozdíl od celoplošných sklizní je kladen důraz na zachycení zdroje a jeho změn v celém rozsahu. Vzhledem k omezené kapacitě úložného prostoru není možné sklízet veškerý český web dostatečně. Z tohoto důvodu je budována kolekce zdrojů s kulturní, historickou, výzkumnou, případně další hodnotou napříč všemi tématy. Cílem této kolekce je vytvořit reprezentativní vzorek českého kulturního dědictví, které vzniká elektronicky. Tato kolekce je budována pomocí soustavné práce kurátorů českého webového archivu.

Tematické sbírky jsou kolekce archivovaných zdrojů vztahujících se k určitému tématu. Obvykle se jedná o významné události jako jsou například volby, ale mohou být zaměřeny i na širší problematiku jako například návrh nové budovy Národní knihovny či české předsednictví EU. Sledovány jsou zejména události, které mají širší ohlas v prostředí internetu. Archivace zdrojů v rámci jedné tematické sbírky je prováděna jednorázově, případně několikrát po sobě v kratším časovém rozmezí v závislosti na určení a délce trvání události. Tematické sklizně jsou prováděny pro potřebu hlubšího zachycení otisku daného tématu v elektronických online zdrojích, které není možné zaznamenat prostřednictvím celoplošných sklizní.

Budoucnost a vize

Podstata existence

Proč archivovat web?

- Potřeba zachránit netištěné dokumenty kulturní, umělecké a historické hodnoty pro další generace

- Enormní nárůst elektronických online zdrojů zveřejněných pouze na internetu
- Prchavost elektronických zdrojů – cenné dokumenty mohou být nenávratně ztraceny Link rot - obsah na webu zaniká,
- konkrétní příklady: odkazy z Wikipedie směřují na neexistující obsah, zánik prezentací menších politických uskupení po volbách.

Vize

Kompletní archiv českého webu, který je veřejně přístupný pro své uživatele, s plnotextovým vyhledáváním a s rozhraním pro práci s obsahovými i popisnými metadaty. Volně stažitelné balíčky s archivovanými webovými daty a metadatovými sety pro použití vědeckou obcí. Spolupráce s výzkumníky při výzkumu nad archivovanými objekty.

V současné době je možné procházet celý český webovový archiv pouze na půdě Národní knihovny na specializovaných terminálech. Veřejně online dostupné jsou pouze výběrové sklizně. Kurátoři webového archivu se snaží uzavírat smlouvy s vydavateli jednotlivých webových stránek zahrnutých do výběrových kolekcí, tak aby bylo možné je zveřejnit online.

V realizaci této vize brání českému webovému archivu zejména problémy ze dvou oblastí:
1. legislativní
2. technické

1. legislativní

Největším problémem je ukotvení webové archivace v české legislativě. Jedná se zejména o absenci institutu povinného výtisku pro elektronické dokumenty.

Legal deposit

Legal deposit (česky povinný výtisk) pro online dokumenty české autorské právo nezná. Autorské právo ani knihovní licence nezahrnují elektronické zdroje. Díky knihovní licenci můžeme sklízet (vytvářet kopie), ale nesmíme je veřejně zpřístupňovat, proto jsme nuceni s vydavateli uzavírat smlouvy, popř. můžeme zpřístupňovat archivované verze na základě licence Creative Commons.

Některé země již v současnosti mají zákon upravující povinný výtisk elektronických zdrojů, který v národních knihovnách těchto zemí naplňují často webové archivy (např. Francie, VB, Dánsko).

2. technické

Vytvoření ideálního úložiště, tj. ekonomicky udržitelné, kapacitně a výkonnostně škálovatelné úložiště umožňující fulltextovou indexaci, metadatovou extrakci a analýzu.

Takové uložiště umožní efektivně pracovat s velkými daty v archivu:
- automatizované Quality Assurance tj. automatická kontrola kvality zpřístupnění archivovaných stránek. Automatizované vytváření bibliografických záznamů
- automatická katalogizace s využitím dat zadaných do kurátorského nástroje správy zdrojů
- nejen jednotlivé weby, ale interface pro průzkum metadat, který nejen zpřístupňuje “raw” metadata sety, ale i nástroj, který s těmito sety umožní pracovat přímo nad živým archivem
- možnost aplikování strategií dlouhodobé ochrany: migrace, emulace

Cíle

Ideální úložiště
Uložiště na bázi Hadoop umožňuje vysokou efektivitu zpracování dat v řádech stovek TB, snadno se škáluje jak objemem kapacit, tak potřebným výkonem pro jejich zpracování. Je stále více užíván mezi členy konsorcia International Internet Preservation Foundation (IIPC) a jsou pro něj vyvíjeny svobodné nástroje v komunitě Web archivů. Hadoop vytvoří podmínky pro efektivní zpracování a zpřístupnění fulltextového indexu a umožní další zpracování uložených dat dle potřeb uživatelské komunity.

Open Data
Open Data si je možné v kontextu webových archivů představit jako volný přístup k samotným archivním balíčkům a metadatovým setům. Jakmile extrahujeme první metadatové sety, můžeme je uvolnit volně ke stažení. Metadatové sety chceme vytvářet ve spolupráci s komunitou kolem Open a Linked Data.
Autorský zákon zatím neumožňuje veřejně zpřístupnit samotné archivní balíčky.

Kurátorství dat
V současnosti nám chybí kvalitní nástroje pro analýzu a vyhledávání zdrojů pro zařazení do WebArchivu, týká se to zejména zahraničních webů bohemikálního charakteru. Aktuálně je seznam webů s doménou .cz, který je podkladem pro provádění celoplošných sklizní, získáván od organizace CZ.NIC a zdroje pro výběrové sklizně vybírány podle stanovených kritérií a hodnoceny kurátory WebArchivu. Vzhledem k tomu, jak se web rozvíjí, a to nejen množstvím zdrojů, ale i proměnou jeho struktury (nárůst multimediálního obsahu) je taková činnost značně časově náročná, proto by bylo vhodné vytvořit nebo převzít některé z nástrojů automatického vyhledávání a výběru nejnavštěvovanějších či jinak významných zdrojů (viz Rozšíření záběru celoplošné sklizně).
Na základě analýzy požadavků uživatelů archivu by bylo také vhodné vytvářet konkrétní tematicky profilované kolekce webových zdrojů.
Aktuálně je pro každý webový zdroj zařazený do výběrové sklizně vytvářen kurátory vlastní katalogizační záznam, tyto zdroje jsou tak pro uživatele dostupné v katalogu NK. Nově bychom rádi integrovali nástroje pro automatické vytváření katalogizačních záznamů z již jednou vložených dat a automatizované provádění kontroly kvality zpřístupňování archivovaných dat.

Pro správu zdrojů zařazených do výběrových sklizní je v současnosti používán nástroj WA Admin vytvořený přímo pro WebArchiv. Do budoucna je však tento nástroj nedostačující, proto pracujeme na vývoji nového nástroje, který nám umožní efektivněji spravovat kolekce dat, které se ve WebArchivu nachází.

Rozšíření záběru celoplošné sklizně
Doposud se celoplošné sklizně realizovaly pomocí seznamu domén .cz dodané společností CZ.NIC. Takový výběr neregistruje weby na ostatních TLD doménách. Existují dvě strategie jak objevit další weby, jež by se měly zahrnout do celoplošné sklizně. Vytvořit tzv. explorační sklizeň, který bude pomocí specializovaného nástroje vyhledávat domény v českém jazyce či přímo rozpozná web bohemikálního charakteru. Tyto nástroje ale zatím nejsou standardní součástí webových archivů, kolegové z Portugalska v současnosti provádějí evaluaci dostupných nástrojů, jež bychom mohli v budoucnu použít. Druhá je explorační sklizeň, která vytváří seznam domén a IP adres, jež třídí podle GeoIP lokátoru, takto bychom objevili weby hostované na českých serverech. Ideálním řešením bude propojení obou metod.

Vytvoření fulltextového vyhledávání napříč celým webovým archivem

Fulltextové vyhledávání nad webovým archivem s sebou nese obtíže, kterým je nutné čelit. V prvé řadě se jedná o problém duplicity způsobené historickými verzemi stejných zdrojů. To znamená, že pokud uživatel zadá výraz, který chce hledat a hledaný výraz je obsažen na nějaké webové stránce, která je uložena v různých časových verzích, pak při standardním fulltextovém vyhledávání bude uživateli vrácena celá řada výsledků ze stejných webových stránek. Některé webové stránky mohou mít za téměř 15 let existence webového archivu uloženy desítky verzí. V případě obecnějšího dotazu uživatel získá několik stovek stejných výsledků.

Cílem fulltextového vyhledávání napříč celým webovým archivem by mělo být vytvoření unikátního vyhledávacího enginu, který si s touto eventualitou poradí. Tím se dostáváme k druhému problému úzce souvisejícímu s vyhledáváním, a to je prezentace výsledků vyhledávání. Pouhá agregace výsledků zde nebude dostačovat, neboť je potřeba uživateli dát možnost procházet výsledky hledání napříč časem. Vzniká tím poměrně specifický problém, který přináší nový rozměr do fulltextového vyhledávání a tím je čas.

Závěr

Webové archivy jakožto informační zdroje pro badatelskou činnost, nabývají v posledních pěti letech v zahraniční na významu. Vedle zájmu z tradičních společenskovědních oborů jako např. historie, vznikají nové transdisciplinární obory zkoumající vztahy mezi technologiemi a společností, tyto obory kombinují zavedené metodologie společenských věd s aplikovanou počítačovou vědou. Webové archivy v zahraničí již několik let spolupracují s širokou badatelskou obcí, jmenovitě např. webový archiv Britské knihovny, evropský Internet Memory Foundation či kalifornský Internet Archive. Existují národní a mezinárodní výzkumné iniciativy pracující primárně s daty webových archivů.

Český webový archiv by rád navázal na činnost kolegů ze zahraničí. Pomocí realizace cílů z článku dojde k otevření webového archivu a bude vytvořena možnost kvalitnějšího zpřístupnění webového archivu uživatelům. Toto přinese vyšší počet uživatelů a také možnost zapojení výzkumných pracovišť z celého světa, které budou moci provádět výzkumy nad těmito daty - tedy výzkumy nad historií českého internetu.