Národní knihovna
České republiky
National Library
of the Czech Republic

Strategie budování sbírky Webarchivu

Datum vytvoření: 4. 2. 2015

Terminologie:

archivace webu – proces sběru, ukládání, trvalého uchovávání a zpřístupňování webových zdrojů

Webarchiv – oficiální název digitálního webového archivu Národní knihovny ČR

born digital – dokument, který vznikl elektronicky, bez analogového ekvivalentu (např. webová stránka)

sklízení (harvesting) - proces automatického stahování a sběru dat z vybraných webových zdrojů (vytváření kopií),

sklizeň (harvest) – jeden časově ohraničený proces stahování a sběru dat

český web - webové stránky v českém jazyce, vytvořených na území ČR, českým autorem nebo weby obsahově se vztahujících k Česku

zpřístupnění - užití díla, které zahrnuje umožnění vnímání díla jiné osobě, např. rozšiřování, pronájem, půjčování, vystavování, sdělování díla apod., u elektronických zdrojů se jedná zejména o rozšiřování díla prostřednictvím sítě internet

1. Úvod

Archivace webu

Výrazné rozšíření webu 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. Vzhledem k dynamické povaze webu každý den narůstá počet webových stránek a další obrovské množství stránek zaniká, mění svou podobu, obsah nebo adresu. Mnoho cenných dokumentů může být ztraceno a tak je třeba zachránit i netištěné dokumenty kulturní, umělecké a historické hodnoty 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 (*born digital*).

Posláním Národní knihovny ČR je podílet se na uchování a zpřístupňování kulturního dědictví současníkům i budoucím generacím. Pro tištěnou produkci existuje institut povinného výtisku, u elektronických zdrojů však tento institut chybí.

Historie

Webarchiv Národní knihovny ČR je digitální knihovna českých elektronických online zdrojů. První stránky byly archivovány v roce 2001, pravidelná archivace pak probíhá od roku 2006. Od roku 2007 je Webarchiv členem mezinárodního konsorcia pro archivaci webu IIPC (*International Internet Preservation Consortium*). Webarchiv je také součástí projektu *Národní digitální knihovna*.
V současné době množství dat uložených ve Webarchivu přesahují miliardu soborů, což do rozsahu zabírá téměř 100 TB dat (2014).

2. Cíle

Hlavními cíli Webarchivu jsou

- pravidelné sklízení webových zdrojů (viz kap. 3)
- zpřístupnění sbírky na terminálech v budově Národní knihovny ČR a online zpřístupňování vybraných archivovaných dokumentů
- zajištění dlouhodobého uchování a trvalého přístupu ke všem archivovaným dokumentům
- kontinuální vytváření sbírky archivovaných webů a její organizace za účelem zajištění vyhledávání uvnitř sbírky

Webarchiv Národní knihovny zabezpečuje jak vytváření komplexního archivu českého webu a jeho dlouhodobé uchovávání v LTP systému Národní digitální knihovny, tak i výběr reprezentativního vzorku českého webu a jeho zpřístupnění široké veřejnosti prostřednictvím online přístupu. V širších souvislostech tak jde o součást naplňování poslání Národní knihovny[1], budování sbírek českého kulturního dědictví, jehož částí jsou také elektronicky publikované dokumenty.

[1] http://www.nkp.cz/soubory/ostatni/zrizovaci-listina-nk.pdf

3. Typy sklizní

Typy sklizní

Národní knihovna ČR provádí tři typy archivace:

1) Celoplošná sklizeň
2) Výběrová sklizeň
3) Tematická sklizeň

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.

Výběrová sklizeň pokrývá 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 *výběrových sklizní*, tj. archivací vybraných hodnotných zdrojů. Kolekce je vytvářena v souladu se strategií tvorby fondu NK ČR a využívá metody konspektu[2], tj. rozdělení fondu do předmětových kategorií a skupin. Zdroje jsou v rámci těchto předmětových kategorií navrhovány kurátory webového archivu nebo mohou být navrženy prostřednictvím webového formuláře[3] (viz role navrhovatelů). Tyto zdroje jsou dále individuálně posuzovány kurátory dle kritérií (viz kap. 5).

Tematické sbírky jsou kolekce archivovaných zdrojů vztahující 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.

Role navrhovatelů zdrojů do výběrových sklizní

[2] http://konspekt.nkp.cz/
[3] http://webarchiv.cz/formular-url/

Kurátor – knihovník, který vybírá, hodnotí a testuje zdroje a komunikuje s vydavateli stránky ohledně jejich souhlasu s archivací

Agentura ISSN – přiděluje mezinárodní standardní číslo ISSN seriálovým publikacím, zasílá seznam vydavatelů elektronických seriálů, kteří v žádosti o přidělení čísla ISSN projevili zájem o archivaci jejich publikací

Návštěvník – jakákoliv osoba, která může navrhnout stránku k archivaci

Vydavatel – osoba zodpovědná za vydávání obsahu webových stránek

4. Přístup

Aut. Zákon

Archivaci webu v České republice, zejména zpřístupnění archivovaných elektronických zdrojů vymezuje *Autorský zákon* (č. 121/2000 Sb.). Tento zákon umožňuje prostřednictvím tzv. knihovní licence vytvářet rozmnoženiny díla pro své archivní a konzervační účely. Vzhledem ke znění zákona však není možné tyto rozmnoženiny díla zpřístupnit veřejnosti online.

Na základě autorského zákona jsou kompletní data z Webarchivu zpřístupňována pouze na terminálech v budově Národní knihovny ČR. Takto jsou přístupné zejména zdroje z celoplošných a tematických sklizní, ale i zdroje vybrané v rámci výběrových sklizní, které nebyly ošetřeny smlouvou nebo licencí Creative Commons.

Aby bylo možné zdroje v rámci výběrových sklizní zpřístupňovat online prostřednictvím webových stránek (http://webarchiv.cz) uzavírá NK ČR s vydavateli Smlouvu o poskytování elektronických online zdrojů[4] nebo tyto zdroje archivuje a zpřístupňuje na základě licence Creative Commons[5]. Záznamy všech zdrojů v rámci výběrových sklizní jsou dostupné v katalogu Národní knihovny[6].

[4] Smlouva je k dispozici online na adrese http://webarchiv.cz/files/vydavatele/smlouvaWebarchiv.doc.
[5] http://creativecommons.org/
[6] http://aleph.nkp.cz/F/

5. Kritéria výběru

Kritéria výběru zdrojů pro celoplošné sklizně

Zdroje pro celoplošné sklizně jsou sklízeny na základě seznamu URL adres s doménou .cz poskytovaného správcem domény, sdružením CZ.NIC. Zahrnuty jsou i další webové zdroje bohemikálního charakteru s jinými doménami, které doplňují kurátoři.

Kritéria výběru zdrojů pro výběrové sklizně

Nejvýznamnějším kritériem pro výběr zdrojů do výběrových sklizní Webarchivu je *bohemikální charakter zdroje*. Toto kritérium se řídí pravidlem výběru dokumentů registrovaných v národní bibliografii, které zahrnuje:

- Území – všechny dokumenty (zdroje) publikované na území České republiky
- Jazyk – všechny zdroje v češtině (bez ohledu na místo vydání)
- Autorství – všechny zdroje českých autorů (bez ohledu na místo vydání)
- Předmět/obsah – všechny zdroje, jejichž obsah se týká České republiky nebo českého národa (bez ohledu na místo vydání)

Zdroje jsou do výběrových sklizní zařazovány zejména na základě jejich *obsahu*. Preferovány jsou zdroje s kulturní, vědeckou či historickou hodnotou, které mají originální a unikátní obsah a dlouhodobou badatelskou hodnotu.

K archivaci jsou zařazovány pouze *volně přístupné/zveřejněné zdroje*. Případně je nutné, aby byla přístupná obsahově podstatná část zdroje (zdroj může obsahovat např. sekci pro registrované uživatele).

Zdroje jsou také zařazovány s přihlédnutím k jejich *technické povaze*, jelikož není možné z technického hlediska sklidit všechny zdroje v takové podobě, v jaké se nacházejí na webu.

6. Uživatelé

Webarchiv je jako archiv českého webu, jehož část je volně dostupná online, určený široké veřejnosti. Vzhledem k regionálnímu vymezení jeho sbírek je Webarchiv určen zejména pro uživatele se vztahem k České republice. Uživatele Webarchivu je možné rozdělit do skupin na základě jejich informačních potřeb:

a) Individuální uživatelé
b) Institucionální uživatelé

c) Výzkumníci a vědci

Největší skupinu tvoří individuální uživatelé, kteří přichází do webového archivu s vlastním informační potřebou. Zájmem těchto uživatelů je zejména jednotlivé procházení historických dat. Touto skupinou se rozumí obecná veřejnost s přístupem k internetu a webovému prohlížeči.

Institucionálními uživateli se rozumí takové instituce, které potřebují a využívají data z webového archivu pro svou činnost. Takovými institucemi může být například policie, soudy, výzkumné ústavy atd. Specifikem těchto uživatelů je možnost získání dat z archivu na základě odůvodněného písemného požadavku. Mezi institucionální uživatele mohou také patřit provozovatelé počítačových či internetových služeb.

Současným trendem v oblasti archivace webu je rostoucí význam a využití rozsáhlých souborů dat získaných z webových archivů. Tato tzv. big data mohou sloužit pro zkoumání jazyka, technologie, historie nebo dalších oblastí. Pro výzkum těchto dat se používá různých vizualizací, textových analýz, zkoumání trendů a jiných metod. Požadavky skupiny výzkumníků zabývajících se těmito souhrnnými se odlišují od požadavků individuálních uživatelů zaměřených na konkrétní informace z archivu.

7. Závěr

Vzhledem k proměnlivé povaze internetu bude potřeba zachovávat jeho historii a kulturní dědictví publikované online stále narůstat. Do budoucna můžeme také očekávat požadavek na uchování většího rozpětí formátů dostupných na internetu, jako jsou například sociální sítě nebo hry.

Posláním institucí jako jsou národní knihovny je získávání, uchovávání a zpřístupňování kulturního dědictví dané země nebo regionu ve všech jeho podobách, včetně elektronické. Webarchiv Národní knihovny vykazuje nejvyšší pokrytí českého webu z hlediska národní domény, vetší než například pokrytí organizací Internet Archive, která se zabývá archivací na mezinárodní úrovni.

Cílem Webarchivu NK ČR tak je vytvoření kompletního webového archivu který je veřejně přístupný pro své uživatele, s plnotextovým vyhledáváním a s rozhraním pro práci s obsahovými i popisnými metadaty. Cílem do budoucnosti je také zveřejnění volně stažitelných balíčků s archivovanými webovými daty a metadatovými sety pro použití vědeckou obcí a spolupráce s výzkumníky při výzkumu nad archivovanými objekty.