Národní knihovna České republiky

Webarchiv

Strategie budování sbírky Webarchivu

aktualizované znění

Autoři: Mgr. Jaroslav Kvasnica, Mgr. Barbora Rudišinová, Mgr. Marie Haškovcová, Mgr. Monika Holoubková, Mgr. Markéta Hrdličková

Datum: červenec 2019

Verze: 3.0

Úvod

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 digitální podobě. Vzhledem k jeho dynamické povaze 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, jedná se zejména o 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 současným i budoucím generacím. Pro tištěnou produkci existuje institut povinného výtisku, u elektronických zdrojů však chybí.

Český webový archiv Národní knihovny ČR (Webarchiv) je digitální knihovna českých elektronických online zdrojů. Národní knihovna dle zřizovací listiny buduje specializované knihovní fondy, které sestávají z knihovních dokumentů určitého zaměření. Jedním z nich je i fond online dokumentů ve formě webových stránek. V souladu s § 2 písm. c) zákon č. 257/2001 Sb., je Webarchiv organizovaný, soustavně doplňovaný, zpracovávaný, ochraňovaný a uchovávaný soubor knihovních dokumentů, přičemž jednotlivé knihovní dokumenty, tedy jednotlivé uchovávané webové stránky, jsou evidované jako samostatné jednotky knihovního fondu. 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[1] (International Internet Preservation Consortium).

Cíle

Hlavními cíli Webarchivu jsou:

- pravidelné sklízení webových zdrojů
- 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

[1] https://netpreserve.org

Webarchiv Národní knihovny zabezpečuje jak vytváření komplexního archivu českého webu a jeho dlouhodobou ochranu, tak i jeho výběrovou, kurátorsky zpracovanou část a její zpřístupnění široké veřejnosti prostřednictvím online přístupu. V širších souvislostech tak jde o součást naplňování poslání Národní knihovny, budování sbírek českého kulturního dědictví, jehož částí jsou také elektronicky publikované dokumenty.

Typy sklizní

Klíčovým prvkem pro tvorbu webového archivu je volba postupů, které jsou použity pro zařazení zdrojů do archivu. Webarchiv využívá tří přístupů k akvizici zdrojů: celoplošnou, tematickou a výběrovou sklizeň. Sklizní se rozumí časově ohraničený proces automatizovaného stahování a sběru dat z vybraných webových zdrojů na základě definovaných parametrů.

Celoplošné sklizně

Celoplošná sklizeň je zaměřena na archivaci všech webových stránek zveřejněných na .cz doméně. Jejich kompletní seznam má Webarchiv k dispozici díky podpoře sdružení CZ.NIC[2]. Sklizeň je prováděna zpravidla jednou až dvakrát ročně a takto archivované stránky jsou z důvodu prostorových kapacit obvykle sklízeny na nižší úrovni než sklizně výběrové a tematické. Cílem celoplošných sklizní je zachycení obrazu českého webu v daném čase.

Výběrové sklizně

Výběrová sklizeň pokrývá pouze vybrané zdroje, ale na rozdíl od celoplošných sklizní je kladen důraz na zachycení zdroje a jeho změn v nejširším rozsahu. Vzhledem k omezené kapacitě úložného prostoru není možné sklízet veškerý český web dostatečně. Z tohoto důvodu je budována kolekce hodnotných zdrojů napříč všemi tématy. Cílem této kolekce je vytvořit relevantní vzorek českého kulturního dědictví, které vzniká elektronicky.
Kolekce je budována pomocí výběrových sklizní, tj. archivací vybraných hodnotných zdrojů. Je tvořena v souladu se strategií tvorby fondu NK ČR a využívá metody Konspektu, tj. rozdělení fondu do předmětových kategorií a skupin. Zdroje jsou v rámci těchto předmětových kategorií navrhovány kurátory webového archivu nebo mohou být navrženy veřejností prostřednictvím webového formuláře. Tyto zdroje jsou dále individuálně posuzovány kurátory dle stanovených kritérií (viz Kritéria výběru).

[2] https://www.nic.cz

Tematické kolekce

Tematické kolekce jsou sbírky archivovaných zdrojů vztahující se k určitému tématu nebo události. Mohou být vytvářeny za účelem zachycení významných aktuálních událostí, které mají širší ohlas v prostředí internetu - ať už plánovaných (volby, výročí) nebo nenadálých (povodně, vládní krize). Dalším typem jsou dlouhodobě budované kolekce, které nejsou podmíněné aktuálností nebo časovým ohraničením tématu. Může se jednat například o archivaci určitého typu dokumentů (například kolekce elektronických periodik, zdrojů vystavených pod licencí Creative Commons) nebo zdrojů významné instituce (Univerzita Karlova, Národní archiv).
Tyto kolekce jsou tvořeny jednotlivými sklizněmi. Tematické sklizně jsou prováděny pro potřebu hlubšího zachycení otisku daného tématu v elektronických online zdrojích, které není možné zaznamenat prostřednictvím celoplošných sklizní. V případě aktuálních událostí probíhá monitorování a sklizeň zdrojů obvykle v několika fázích - pokud možno před samotnou událostí, v jejím průběhu a po ukončení. Dlouhodobě budované kolekce jsou sklízeny průběžně.
Český webový archiv se také zapojuje do tvorby tematických kolekcí, které jsou koordinovány mezinárodním konsorciem IIPC.

Kritéria výběru

Nejvýznamnějším kritériem pro výběr zdrojů do výběrových sklizní a tematických kolekcí Webarchivu je bohemikální charakter zdroje. Toto kritérium se řídí pravidlem výběru dokumentů registrovaných v národní bibliografii, které zahrnuje dokumenty splňující alespoň jedno z následujících kritérií:

- vydané na území dnešní České republiky (územní bohemikum)
- napsané autory původem z Česka (autorské bohemikum)
- napsané v českém jazyce (jazykové bohemikum)
- pojednávající o České republice (obsahové bohemikum)

Zdroje jsou do výběrových sklizní a tematických kolekcí 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 vybírány pouze volně přístupné/zveřejněné zdroje, případně je nutné, aby byla přístupná jejich obsahově podstatná část (zdroj může obsahovat např. sekci pro registrované uživatele). Zdroje jsou také zařazovány s přihlédnutím k jejich technické povaze. Archivovány jsou ty,

které je z technického hlediska možné sklidit alespoň v přibližné podobě, v jaké se nacházejí na webu.

Zpřístupnění

Webarchiv se coby specializovaný knihovní fond může opřít o tzv. knihovní licenci. Archivaci webu v České republice, zejména zpřístupnění archivovaných elektronických zdrojů, vymezuje Autorský zákon (č. 121/2000 Sb.). Tento zákon umožňuje Národní knihovně prostřednictvím 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 zařazené do výběrových sklizní, které nebyly ošetřeny smlouvou nebo licencí Creative Commons.
Aby bylo možné zdroje v rámci výběrových sklizní zpřístupňovat online prostřednictvím webových stránek (https://webarchiv.cz), uzavírá NK ČR s vydavateli licenční smlouvu o užití díla nebo tyto zdroje archivuje a zpřístupňuje na základě licence Creative Commons, pod níž jsou stránky vystaveny. Záznamy všech veřejně přístupných zdrojů v rámci výběrových sklizní jsou dostupné v katalogu Národní knihovny.

Pravidla pro mazání a blokaci archivních kopií webových stránek

Aby Webarchiv mohl garantovat autenticitu svých archivních kopií, musí mít jasně nastavená pravidla nejen pro jejich akvizici, ale také pro jejich další správu. Prvním pravidlem je, že Webarchiv za žádných okolností nemaže archivní kopie nebo jejich součásti. Hlavním důvodem je právě zachování autenticity archivních dat, aby tak mohlo být garantováno, že s nimi nebylo nijak manipulováno, nebyla porušena a jsou tím, za co se vydávají. Právě povaha digitálních archivních dat s sebou nese velké riziko narušení autenticity.
Dalším důvodem je, že webové stránky jsou navzájem propojeny hypertextovými odkazy, a velmi často se stává, že samotná stránka nemusí bez dalších stránek, na které odkazuje, obsahovat žádnou informaci, nebo obsahuje jen informaci nekompletní. To znamená, že webové dokumenty na úrovni stránky, ale také na úrovni webu téměř nikdy samotné nedávají smysl, protože jsou vmíchané do větší sítě dokumentů.

Nicméně Webarchiv si uvědomuje, že v některých případech se archivní kopie mohou dotýkat práv některých uživatelů. Tyto problémy nebere na lehkou váhu a vždy k nim přistupuje individuálně se snahou najít s uživateli vhodné řešení. Pokud je Webarchiv upozorněn na závadný obsah, přistupuje k jeho blokaci. Blokace má nastavená jasná pravidla a může probíhat na několika úrovních:

1. Zamezení zobrazování webové stránky uživatelům veřejně online, přičemž stránka je stále přístupná z referenčního centra NK ČR a je i nadále archivována. Tento případ může nastat např. při vypovězení licenční smlouvy ze strany vydavatele, při změně vlastníka domény apod. Jde tedy o problémy s licencí, která umožňuje zpřístupnit webové stránky veřejnosti i mimo budovu NK.
2. Zamezení vytváření archivních kopií webové stránky. Stránka již není nadále archivována, ale její archivní kopie jsou stále dostupné z referenčního centra NK ČR a uživatelé si je mohou přijít i nadále prohlédnout. Velmi často se jedná o případy, kdy sklízecí robot omylem pronikne do špatně zabezpečené stránky, např. administrační části periodika nebo internetového obchodu.
3. Zařazení stránky na černou listinu zobrazovací aplikace. Toto je krajní řešení, protože takové stránky již nejsou přístupné pro uživatele ani v referenčním centru a nejsou běžně dostupné ani pracovníkům Webarchivu. Ve většině případů je stránce zamezena i další archivace. K tomuto kroku se přistupuje právě při porušení zákona, nejčastěji jde o nahlášení poškození autorských práv ze strany autora.

Uživatelé

Webarchiv, archiv českého webu, jehož část je volně dostupná online, je určen široké veřejnosti. Vzhledem k regionálnímu vymezení jeho sbírek je 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 přicházejí do webového archivu s vlastní informační potřebou. Zájmem těchto uživatelů je zejména procházení jednotlivých historických dat. Touto skupinou se rozumí veřejnost s přístupem k internetu a webovému prohlížeči.

Institucionálními uživateli jsou takové instituce, které potřebují a využívají data z webového archivu pro svou činnost. Takovými institucemi mohou 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 využívá různých vizualizací, textových analýz, zkoumání trendů a jiných metod. Požadavky skupiny výzkumníků zabývajících se těmito souhrnnými soubory dat se odlišují od požadavků individuálních uživatelů zaměřených na konkrétní informace z archivu.

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, větší než například pokrytí organizací Internet Archive[3], která se zabývá archivací webu na mezinárodní úrovni.

Cílem Webarchivu NK ČR je vytvoření webového archivu, který je veřejně přístupný pro své uživatele, umožňuje plnotextové vyhledávání a má rozhraní 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 badateli při výzkumu archivovaných objektů.

Další praktické informace jsou k dispozici na webových stránkách v sekci FAQ - Často kladené dotazy (https://webarchiv.cz/cs/faq).

[3] https://archive.org

Použitá terminologie

Webová archivace

Archivace webu je proces, který zahrnuje získávání webových zdrojů, jejich ukládání, trvalé uchování, ochranu a v neposlední řadě i jejich zpřístupnění.

Sklízení webu

Jde o proces sběru dat z webu, který spočívá v automatizovaném mapování, vyhledávání a stahování určitých webových stránek pomocí crawlerů na základě definovaných parametrů. Crawler je speciální počítačový program, který dokáže automaticky procházet a stahovat webové stránky. Používají je nejen internetové vyhledávače, ale i jednotlivé webové archivy. Webarchiv používá crawler Heritrix, který vytvořil Internet Archive. Jedná se o open software.

Sklizeň

Jeden časově ohraničený proces automatizovaného stahování a sběru dat z vybraných webových zdrojů na základě definovaných parametrů.

Born digital

Dokument, který vznikl elektronicky bez analogového ekvivalentu (např. webová stránka).

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.

Creative Commons

Licence Creative Commons (CC) jsou souborem licencí, které slouží k legálnímu sdílení a využívání autorských děl. CC vznikly z potřeby právně upravit vztah mezi uživateli díla a jeho autorem v případě, že autor chce, aby jeho dílo mohla za vyznačených podmínek používat široká veřejnost. Licence CC tedy umožňuje autorovi nabízet jeho dílo prostřednictvím licenční smlouvy, na základě které poskytuje veřejnosti některá práva k dílu a jiná si vyhrazuje.