

MASARYKOVA UNIVERZITA
FAKULTA INFORMATIKY



OpenLibrary.org - Otevřená knihovna
www.OpenLibrary.org

ESEJ DO PŘEDMĚTU PV070

Stanislav Novák

324 610

2. ročník

16.12.2009

podzim 2009

1. Otevřená knihovna se představuje

Open Library, neboli česky Otevřená knihovna, je projekt neziskové organizace Internet Archive z části dotovaný knihovnou státu Kalifornie. Je to obrovský web postavený na systému wiki, který je otevřený, a tudíž do něj může přispívat úplně každý. Od obyčejného čtenáře, přes knihovníka, až po zkušeného programátora [1]. Celý projekt v dnešní době (2009) funguje pod taktovkou George Oatese. Tým zajišťující technické zázemí vede Americký programátor a aktivista Aaron Swartz [2].

2. Historie

Po několika letech usilovné práce se v roce 2007 na internet dostává první verze Open Library. Na této události se podílely světové firmy takového formátu, jako např. Open Content Alliance, Yahoo, Adobe, MSN nebo HP. V té době už knihovna umožňovala fulltextové vyhledávání ve všech dostupných zdigitalizovaných knihách a jejich stahování ve formátu PDF. [3]

V době spuštění jsme zde mohli nalézt pouze několik tisíc záznamů o knihách. První exempláře přineslo spojení projektu Open Library se skenováním knih, které v té době prováděla organizace Internet Archive. Vybírány byly pouze ty knihy, na které se nevztahoval autorský zákon. V současnosti, o dva roky později, jsou zde záznamy o více jak 30 milionech knih. Na stránkách Open Library se běžný uživatel dostane zhruba k 23 milionům, z čehož u milionu knih je umožněno fulltextové vyhledávání. [1]

3. Cíle projektu

„One web page for every book ever published. It's a lofty, but achievable, goal.“
Tolik hlásá věta uvozující sekci “O nás“ na webových stránkách Open Library. Cílem projektu je tedy vytvoření webové stránky pro každou knihu, která kdy byla vydána. A to včetně metadat, recenzí odborníků, čtenářských recenzí a diskuze o knize. Hned nato autoři dodávají, že jsou si vědomi, jak vysokého cíle chtějí dosáhnout, což ovšem není v době dnešních technologií nemožné. Jediným omezujícím faktorem je nedostatek lidské síly jak k rozšiřování databáze knih, tak k vylepšování obslužného softwaru a vzhledu webového rozhraní. K těmto činnostem proto vyzývají občany z celého světa. Prostor pro dobré nápady a smysluplnou práci, kterou ocení i nespočet dalších generací, není ničím omezen. [1]

Projekt nabízí možnost spolupráce i schopným programátorům. Na stránkách Open Library je vyvěšen seznam chyb, které uživatelé nahlásili a potřebují být opraveny. V současné době jsou zde například nedostatky typu „V Internet Exploreru nefunguje tlačítko plus“ nebo „Je potřeba importovat a zaarchivovat nový ONIX“ (viz níže). [6]

Knihovníci se mohou do projektu zapojit hned několika způsoby. Nejvyšší prioritu má v současnosti vývoj metadatového schématu s ohledem na uživatele, administrátory i dobrovolné spolupracovníky. Dále je to možnost diskuze s ostatními knihovníky o univerzálním číslování OLN. Oba zmíněné body zájmu jsou rozvedeny níže. [6]

Ani běžní internetoví uživatelé a milovníci kvalitního čtení nejsou o možnost přispívání do Otevřené knihovny ochuzeni. Ba právě naopak. Morální povinností všech uživatelů je nahlásit administrátorům každou chybu, nedostatek či návrh na zlepšení, kteří ji ihned předají k následnému zpracování.

I člověk, který knihovnické práci nerozumí a chyby nenachází, se může zapojit. A to sháněním kamenných knihoven, které by se chtěly na projektu podílet. Lidé obstarávající chod Open Library se snaží nabízet svou práci na internetu zdarma. Bohužel se většinou nejedná pouze o práci jejich, ale hlavně dalších osob - autorů knih a jiných osob, které mají s knihovnictvím co do činění. „*Pokud máte možnost získat od Vaší oblíbené knihovny výpis jejich záznamů, budeme Vám navždy zavázáni,*“ vyzývají tvůrci Open Library. [5]

4. Současný stav

4.1. Knihy existují, knihy vznikají

K počtu 23 mil. knih značně přispělo importování největších amerických digitálních knihoven. Jmenujme například Library of Congress (8 mil. záznamů), University of North Carolina (4 mil. záznamů) nebo University of Toronto (6 mil. záznamů). Největší kolekce knih, jež by v digitální podobě chtěla Open Library získat do svých sbírek, je knihovna University of California (12 mil. záznamů). Dále má velký zájem například o British Library, Bibliothèque Nationale de France nebo o Library and Archives Canada. [5]

Open Library ale nečerpá pouze z kamenných a již existujících digitálních knihoven. Stále více se zaměřuje přímo na vydavatele. Většina z nich poskytuje takzvané *ONIX feeds*, což jsou soubory informací o vydaných knihách ve formátu XML. Bohužel tyto informace nejsou veřejně dostupné a jejich opatření vyžaduje individuální přístup jednotlivce k vydavateli. I v tomto případě nás tvůrci žádají o pomoc. Z největších integrovaných vydavatelů jmenujme například *HarperCollins*, *Simon and Schuster* a *Random House*. [5]

Pokud si necháme vyhledat všechny knihy, které jsou kompletně zdigitalizované a mají umožněno fulltextové vyhledávání, bez ohledu na název, obdržíme jako výsledek 5000 záznamů. V následujících tabulkách jsou uvedeny počty knih vydaných v jednotlivých časových obdobích a jejich seřazení podle jazyka, v němž jsou napsány.

Časové období	Počet knih
Před rokem 1920	3298
1920 - 1939	198
1940 - 1959	149
1960 - 1979	57
1980 - 1999	1288
Po roce 2000	6

Jazyk	Počet knih
Angličtina	3820
Francouzština	362
Němčina	360
Italština	82
Španělština	45
Latina	40

4.2 futurelib

Vůbec poprvé se na svět dostává otevřená, univerzální a hlavně všem dostupná sbírka informací o všech knihách na světě. Z tohoto faktu ovšem plynou některá úskalí. To největší spočívá ve způsobu značkování knih pomocí metadat. Od doby, kdy byl zaveden formát *MARC* se mnohé změnilo a pro řadu případů je tento formát nevyhovující a zastaralý. V dnešní době již není zapotřebí tolik knihovnických informací, jako uchovával *MARC*. Mnohem větší důležitost zastávají informace vydavatelů. Lidé z Open Library si tedy kladou za cíl vytvořit formát, který by uspokojil všechny, a zároveň byl srozumitelný pro běžné, v této oblasti neškolené, uživatele. Nazývá se *futurelib*. K jeho dalšímu vývoji plánují tvůrci

sérii setkání společných rozprav technických nadšenců, knihovníků i běžných milovníků knih. [7]

Schéma futurelib obsahuje následující atributy: [17]

Název atributu	Popis
Name of Property	název díla bez interpunkčních znamének kromě tečky a spojovníku
Property URI	URI pro identifikaci díla, popřípadě číslo podle standardu ISO
Preferred label(s)	označení díla v různých jazycích ve tvaru "xx-label", kde "xx" označuje zkratku jazyka
Defined By	odkaz na dokument, ve kterém je definováno schéma metadat; ideálně URI, popřípadě umístění či bibliografický identifikátor
Refines	širší označení; popisované označení blíže popisuje odkazované označení; bližší popis specifikuje význam elementu
Refined By	užší označení; popisované označení je blíže popsáno odkazovaným označením
Source Definition	definice označení, označení nese přiřazenou definici jako zdroj svého původu
Source Comments	komentáře označení ze jmenného prostoru, ve kterém je označení původně uchováno
[Domain] Comments	doménová specifikace komentářů označení
Value type	dílo je schopno nést jednu nebo více hodnot uvedených níže, které nesmí být v rozporu s definičním dokumentem odkazovaným výše
Value type: uncontrolled	nekontrolovaná hodnota je řetězec, který není dále specifikován; nemůže být použita pro odkazování na jiné hodnoty nebo metadata; tento atribut nabývá hodnot YES a NO
Value type: typed string	řetězec, který má danou strukturu
Value type: controlled list	obsahuje identifikátor seznamu a samotný seznam hodnot
Value type: other	hodnotou tohoto atributu může být komplexní struktura, jiné dílo, odkaz na další metadata; hodnota musí být identifikovatelná, nejlépe pomocí URI
Obligation	určuje, zda je element povinný vždy (M), povinný, pokud je k něčemu použitelný (MA), silně doporučený (R) nebo nepovinný (O)
Occurence	hodnota indikující omezení na opakování elementu

4.3 OLN

Kromě jednotného formátu metadat je také zapotřebí vytvořit univerzální číslování jednotlivých knih. Nejlépe takové, které by se co nejvíce podobalo číslování ISBN či ISTN, ale aby se s ním daly označit všechny knihy, nikoli pouze ty, jež už očíslovány jsou.

Prozatímní název tohoto jednoznačného identifikátoru je Open Library Number, zkráceně *OLN*.

4.4 Jak to celé funguje

Počáteční idea byla následující:

- vytvořit databázi, která bude obsahovat desítky milionů záznamů
- každý běžný uživatel bude mít možnost měnit její obsah
- aby se zabránilo případným záškodníkům, bude existovat kompletní seznam provedených změn

Všechny tyto body již byly v minulosti vyřešeny. Háček je v tom, že pouze jednotlivě, nikdy všechny tři naráz. Proto bylo vytvořeno databázové schéma pod názvem *Infobase*. Každá položka v této databázi je uložena jako objekt typu *Thing*, včetně webových stránek, knih a autorů. Implementace je zajištěna jazykem *PostgreSQL*. [8]

Samotná databáze ale ke spuštění projektu nestačí. Bylo zapotřebí vytvořit novou wiki, na které by se dalo stavět. A tak vzniklo *Infogami*. Od ostatních wiki systémů se liší svou jednoduchostí. Na druhou stranu je schopna pracovat s rozličnými typy dat a umí ukládat i data strukturovaná. [8]

4.5 Architektura, aneb stavíme knihovní web

Infogami je napsáno v Pythonu a používá kostru *web.py* [10]. Běží na webovém *lighttpd* serveru přes rozhraní *FastCGI* za pomoci *Flupu* [9]. Základní šablony jednotlivých stránek jsou statické HTML, ale při vytváření konečné podoby jsou editovány pomocí rozhraní wiki. K vyhledávání je použit *Solr engine* [11], který stojí na vyhledávací knihovně *Lucene* [12]. Obě knihovny jsou napsány v Javě. Samotný *Solr* je postaven na *Apache Ant* [13]. Katalogizovaná data a fulltext jsou získány z různých zdrojů. Co se týče managementu projektu, pro řízení zdrojů je použit systém *Mercurial* [14] a *Launchpad* [15] pro sledování jednotlivých vydání. [16]

4.6 Role uživatelů

Z důvodu utvoření alespoň základní hierarchie pravomocí, mají uživatelé v systému Open Library následující role: [18]

Administrátor	spravuje metadata, modifikuje xml, vytváří podpůrné meta soubory
Komentátor	čte knihy, kupuje knihy, mluví o knihách, vkládá odkazy
Spisovatel	píše knihy
Recenzent / vkladatel metadat	čte knihy, kupuje knihy, píše o knihách, přidává ke knihám metadata
Příznivec / podporovatel	vyzdvihuje kvalitní knihy a jejich autory
Lovec knih	kupuje na internetu knihy, o kterých našel informace na Open Library
Literární vědec	čte knihy a píše z nich výtahy

Vyžírka	do knihovny přistupuje přistupuje přes Google, zaměřuje se na knihy v plných textech, které následně zveřejňuje na blozích
Kurátor	sdužuje knihy, organizuje informace, vytváří značky a taxonomie
Korektor	opravuje informace, vytváří značkování
Milovník knih	mluví, kupuje, sdužuje, značkuje, organizuje a stahuje knihy po více kusech najednou a následně o nich píše
Knihovna	sdílí knihy
Vydavatel	publikuje knihy
Uživatel API, vývojář	vytváří programy, navrhuje šablony, zlepšuje datovou výbavu knihovny

5. Můj názor

Už na střední škole jsem proklínal literaturu a nutnost chodit kvůli dvěma básním za měsíc do kamenné knihovny, neboť na internetu žádná taková možnost nebyla. Proto jsem velice rád, že se našla parta nadšenců, kteří se snaží nashromáždit všechny knihy světa a nabídnout je široké veřejnosti na jedné webové adrese. Design stránek je příjemný na pohled, ovládání je intuitivní a dokonce je i část přeložena do češtiny. Volně dostupné knihy jsou zobrazitelné přímo, knihy, na které se vztahují autorská práva, mají napravo nabídku internetových obchodů, kde je právě tato kniha skladem.

Autorům projektu držím palce a těším se, až moje děti budou za několik let vesele brouzdat internetem, stahovat knihy v digitální podobě a nebudou se muset obtěžovat cestou do kamenné knihovny přes celé město.

6. Použité zdroje

- [1] Open Library - About Us
< <http://openlibrary.org/about> >
- [2] Open Library na wikipedii
< http://en.wikipedia.org/wiki/Open_Library >
- [3] Emag.cz - článek o spuštění Open Library
< <http://www.emag.cz/open-library-spustena/> >
- [4] Radar O'Reilly - článek o spuštění demoverze Open Library
< <http://radar.oreilly.com/archives/2007/07/open-library-la.html> >
- [5] Open Library - How You Can Help
< <http://openlibrary.org/about/help> >
- [6] Bugs in Open Library
< <https://bugs.launchpad.net/openlibrary/> >
- [7] Open Library - About the Librarianship
< <http://openlibrary.org/about/lib> >
- [8] Open Library - About the Technology
< <http://openlibrary.org/about/tech> >
- [9] Flup: Random Python WSGI Stuff
< <http://trac.saddi.com/flup> >
- [10] Web.py
< <http://webpy.org/> >
- [11] Apache Solr search engine
< <http://lucene.apache.org/solr/> >
- [12] Lucene search software
< <http://lucene.apache.org/> >
- [13] The Apache Ant Project
< <http://ant.apache.org/> >
- [14] Systém Mercurial
< <http://www.selenic.com/mercurial/> >
- [15] Software Launchpad
< <http://launchpad.net/> >

[16] Open Library - Open Library Architecture
< <http://openlibrary.org/about/architecture> >

[17] Futurelib - dokumentace metadatového formátu
< <http://futurelib.pbworks.com/AP+template+examples> >

[18] The State of the UI
< <http://openlibrary.org/dev/docs/ui> >

Všechny odkazy byly naposledy zkontrolovány 28.11.2009.

7. Metadata

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title"
      content="OpenLibrary.org - Otevřená knihovna" />
<meta name="DC.Creator" content="Stanislav Novák" />
<meta name="DC.Subject" content="Open Library" />
<meta name="DC.Subject" content="Otevřená knihovna" />
<meta name="DC.Subject" content="books" />
<meta name="DC.Description" content="Esej popisuje stav projektu
OpenLibrary.org, jeho principy, cíle, možnosti dalšího
rozšíření a náměty pro různé skupiny obyvatelstva, jak se
aktivně zapojit do vývoje digitální knihovny, která chce
shromáždit všechny knihy světa." />
<meta name="DC.Date" content="27.11.2009" />
<meta name="DC.Type" content="esej" />
<meta name="DC.Type" content="text" />
<meta name="DC.Format" content="application/pdf" />
<meta name="DC.Format" content="computerFile" />
<meta name="DC.Identifier" content=
      "http://is.muni.cz/www/324610/9839986/
OpenLibrary.pdf" />
<meta name="DC.Source" scheme="URL"
      content="http://openlibrary.org/about/" />
<meta name="DC.Source" scheme="URL"
      content="http://en.wikipedia.org/
wiki/Open_Library/" />
<meta name="DC.Source" scheme="URL"
      content="http://www.emag.cz/
open-library-spustena/" />
<meta name="DC.Source" scheme="URL"
      content="http://radar.oreilly.com/
archives/2007/07/open-library-la.html" />
<meta name="DC.Source" scheme="URL"
      content="http://openlibrary.org/about/help/" />
<meta name="DC.Source" scheme="URL"
      content="https://bugs.launchpad.net/
openlibrary/" />
<meta name="DC.Source" scheme="URL"
      content="http://openlibrary.org/about/lib/" />
<meta name="DC.Source" scheme="URL"
      content="http://openlibrary.org/about/tech/" />
<meta name="DC.Source" scheme="URL"
      content="http://trac.saddi.com/flup/" />
<meta name="DC.Source" scheme="URL"
      content="http://webpy.org/" />
<meta name="DC.Source" scheme="URL"
      content="http://lucene.apache.org/solr/" />
<meta name="DC.Source" scheme="URL"
      content="http://lucene.apache.org/" />
<meta name="DC.Source" scheme="URL"
      content="http://ant.apache.org/" />
<meta name="DC.Source" scheme="URL"
```

```
        content="http://www.selenic.com/mercurial/" />
<meta name="DC.Source" scheme="URL"
        content="http://launchpad.net/" />
<meta name="DC.Source" scheme="URL"
        content="http://openlibrary.org/
        about/architecture/" />
<meta name="DC.Coverage" content="FI:PV070 Digitální knihovny" />
<meta name="DC.Language" content="cze" />
```