FAKULTA INFORMATIKY

MASARYKOVEJ UNIVERZITY

Projekt Gutenberg

Brno 2013

Autor: Juraj Duráni
Ročník: 5.
Dátum spracovania: 28.11.2013
Projekt: Projekt Gutenberg, Michael Stern Hart, http://www.gutenberg.org/

Úvod

Projekt Gutenberg (PG) je najstaršia digitálna knižnica poskytujúca voľne prístupné elektronické knihy (eKnihy) na svete. Založil ho Michael Stern Hart (8. 3. 1947 – 6. 9. 2011) v roku 1971. Jeho cieľom bolo digitalizovať knihy a poskytnúť ich širokej verejnosti. Predpokladá sa, že na svete je 20 až 30 miliónov voľne šíriteľných publikácií, z nich je približne 5 miliónov na internete.

PG sa zameriava na širokú oblasť čitateľov. Od toho sa odvíja aj spôsob vyberania diel, ktoré budú digitalizované a spôsob použitia týchto kníh. Všetky texty sú čo najjednoduchšie, aby mohli byť použité kdekoľvek. Preto je ako kódovanie použité ASCII, ktoré dokáže prečítať akékoľvek zariadenie.

Vytvárané knihy nie sú autoritatívne. PG sa zameriava na používateľov, ktorým záleží viac na obsahu diela ako na forme a doslovnom prepise.

História

V roku 1971 Michael Hart dostal na Xerox Sigma V mainframe v Materials Research Lab na University of Illinois účet, kde mohol využiť 100 000 000 $ počítačového času. Rozhodol sa, že tento čas nepoužije na počítanie, ale na ukladanie a prehľadávanie dát. Prepísal americkú Deklaráciu nezávislosti a poslal odkaz, kde je voľne prístupná 100 používateľom na internete, z toho 6 ľudí si ju stiahlo. To sa považuje za vznik PG a jeho prvé digitálne dielo.

Potom sa už projekt začal rozrastať. Svoju 10. knihu pridal v auguste 1989 (The King James Bible), 100. knihu v januári 1994 (The Complete Works of William Shakespeare), 1 000. v auguste 1997 (La Divina Commedia di Dante) atď. V apríli 2002, 37 rokov od vzniku projektu, zahŕňa PG približne 5 000 kníh. Od tohto okamihu sa trend pribúdania kníh ustálil na zhruba 5 000 kníh každých 18 mesiacov.

V roku 1997 Michael Hart vyjadril záujem aj o iné jazyky. Až doteraz boli všetky diela v angličtine (a aj v súčasnosti je drvivá väčšina diel v anglickom jazyku) a už v roku 1998 PG zahŕňal 10 kníh vo francúzskom jazyku. Do roku 2008 PG začlenil diela v 55 jazykoch, z toho napríklad aj 45 v esperante či 6 v bulharčine[1].

V roku 1998 sa Michal Hart vyjadril, že chce vytvoriť 10 000 eKníh, čo sa mu podarilo v októbri 2003. Toto číslo samozrejme nie je končené a vo svojom výroku sa ani Michael Hart pri ňom nezastavil[2].

V októbri 2000 štartuje pridružený projekt Distributed Proofreaders, ktorého účel je zapojiť dobrovoľníkov aj do oblasti kontroly textov a zvýšiť tak kvalitu diel. K projektu sa môže pripojiť ktokoľvek a pomôcť tak pri skvalitňovaní digitalizovaných diel. V januári 2008 projekt zahŕňa 52 000 dobrovoľníkov.

December 2006 je štartovným dátumom portálu PG News, ktorý sumarizuje týždenné a mesačné novinky PG. Taktiež poskytuje štatistiky pre PG zahŕňajúce týždenné, mesačné a ročné pribúdanie kníh do archívu.

---

[1] Súčasne 9 diel v českom jazyku. Slovenské diela zatiaľ bohužiaľ chýbajú.

[2] "My own personal goal is to put 10,000 eTexts on the Net and if I can get some major support, I would like to expand that to 1,000,000 and to also expand our potential audience for the average eText from 1.x% of the world population to over 10%, thus changing our goal from giving away 1,000,000,000,000 eTexts to 1,000 times as many, a trillion and a quadrillion in US terminology."

Filozofia

PG sa snaží poskytnúť čo možno najširšej verejnosti čo možno najviac voľne šíriteľnej literatúry. Základný predpoklad, na ktorom Michael Hart postavil PG je, že čokoľvek, čo je možné napísať do počítača (alebo sa už v digitálnej podobne nachádza), môže byť ľubovoľný počet krát kopírované ("Replicator Technology"). Rovnako tak môžu byť tieto kópie poskytnuté komukoľvek na svete, alebo aj mimo neho.

1. PG by mal stáť tak málo, že sa nik nebude zaujímať, koľko stojí. Mal by mať dostatočne malú veľkosť, aby sa zmestil na štandardné médium.

Dokumenty dostupné cez PG sú zdarma pre kohokoľvek[3]. Problém ale môže nastať pri pokuse o získanie tohto diela. Keď PG začínal, Michael Hart prepísal Deklaráciu nezávislosti a odkaz poslal 100 užívateľom. Súbor mal veľkosť 5 kB. Keby sa v tej dobe pokúšal poslať tento súbor namiesto odkazu, pravdepodobne by to sieť nevydržala.

Keď sa rozšírili osobné počítače, veľkosť úložného priestoru nebola príliš veľká. Snažiť sa digitalizovať rozsiahle diela, ktoré by užívatelia neboli schopní uložiť na svoj disk preto nemá význam. V roku 1991 Michael prepísal Alicu v krajine zázrakov od Lewisa Carrolla a Petra Pana od Jamesa M. Barrie. Obe tieto detské knižky bolo možné uložiť na disk. Ako sa pos-tupne zvyšovala kapacita úložných médií, bolo možné digitalizovať postupne väčšie a väčšie diela.

Aj keď v súčasnosti už nie je problém s veľkosťou disku, dôraz na veľkosť súborov neprešiel do úzadia. Štandardom pre PG je stále ukladanie textov v tzv. „Plain Vanilla ASCII“, teda čistom ASCII kódovaní.

2. PG by mal byť natoľko jednoduchý, aby sa nik nemusel strachovať, ako ho použiť.

Keďže sa PG zameriava na čo najširšiu vrstvu ľudí, jeho použitie musí byť čo najjednoduchšie. To je dôvod, prečo je každý text v archíve PG uložený v jednoduchom 7-bitovom ASCII formáte. Tento formát je čitateľný pre akýkoľvek počítač. Preto je jednoduché ho použiť na klasickom PC, tablete, alebo čítačke kníh.

Vzhľadom k tomu, že PG je multilinguálny, podporuje aj iné kódovania pre možnosť zapisovať diakritiku ako napríklad 8-bitové ASCII. Každý takýto text má ale svoju variantu bez diakritiky v 7-ASCII. Výnimkou sú texty, ktoré nemôžu byť prepísané do tejto podoby, ako napríklad čínske texty, ktoré sú uložené v Big-5.

Základné formáty súborov sú *.txt ako textový súbor a *.zip ako štandardná komprimácia. PG však nijako neobmedzuje prispievateľov a eKnihy môžu byť v ľubovoľnom formáte, aký vyhovuje autorovi. Rovnako tak nekladie žiadne prekážky pre užívateľov a snaží sa poskytnúť ľubovoľný formát textov, aký je pre nich vyhovujúci.

PG sa snaží povzbudzovať autorov, aby vytvárali eKnihy a aby ich šírili voľne medzi všetkých užívateľov internetu. Podporuje ľudí s novými myšlienkami a snaží sa nehovoriť „NIE“ ich nápadom. Poskytuje toľko voľnosti autorom a dobrovoľníkom v projekte, koľko je len možné s jediným obmedzením na autorský zákon. PG je organizácia, ktorá nie je podporovaná finančnou, alebo politickou mocou. Celá organizácia je postavená na dobrovoľníctve. Svedčí o tom aj citát:

„Having money is fine ... becoming dependent on it should be avoided.“

[3] S ohľadom na autorský zákon. Bližšie viď kapitolu Copyright.

Súčasnosť

Dnes PG archív obsahuje viac ako 44 200 eKníh a ďalších 100 000 je dostupných cez partnerov, pobočky a pridružené organizácie. Prístupných je aj takmer 950 audio kníh, či už čítaných človekom, alebo automaticky generovaných počítačom. Archív obsahuje aj hudbu, obrázky alebo jednoduché dáta[4].

Denne sú sťahované desaťtisíce eKníh. Napríklad 15. novembra 2010 bolo stiahnutých 130 254 eKníh, mesačne následne 3 582 153. Od štartu „podpory viacjazyčnosti" v roku 1997 sú do decembra 2010 prístupné knihy v 60 jazykoch, aj keď značná väčšina diel je stále v anglickom jazyku (36 821[5]).

S postupom projektu a pribúdaní stále viac a viac diel, boli všetky rozdelené do troch kategórií. Ľahká, ťažká a referenčná literatúra. Neskôr sa upustilo od tohto jednoduchého rozdelenia a PG prešiel ku podrobnejšiemu deleniu.

Vyhľadávanie v archíve zahŕňa aj fulltext v prvých 100 kB každého textu cez Google a prehľadávanie metadát. Formát metadát je RDF/Dublin Core. Dostupný je aj formát MARC21[6].

Hlavná stránka projektu je http://www.gutenberg.org, vyhľadávať knihy je možné jednoducho na http://www.gutenberg.org/catalog/world/search, alebo online katalógu http://www.gutenberg.org/catalog/, kde je možné hľadať podľa autorov alebo titulu. Ďalší archív, kde je možné hľadať a sťahovať knihy v pdf formáte pre iPad, Kindle, Nook, Sony Reader, Kobo a mnoho ďalších zariadení je na adrese http://self.gutenberg.org/.

Motto projektu je:

„Break Down the Bars of Ignorance and Illiteracy."

„Less is more."

Digitalizácia

Na začiatku prebiehala digitalizácia ručným prepisovaním vybraných kníh. Dnes sa robí pomocou štandardných nástrojov, ako je skenovanie a OCR. Niektorí dobrovoľníci ale preferujú ešte stále ručný prepis diela, čo je samozrejme taktiež podporované PG. Staršie knihy, ktoré by sa mohli pri skenovaní poškodiť sa prepisujú ručne.

Naskenované texty po OCR sú vo formáte 7-ASCII, poprípade 8-ASCII pre jazyky používajúce diakritiku alebo Big-5 pre jazyky nekonvertovateľné do týchto formátov ako je napríklad Čínština.

Každá takáto kniha následne podstúpi tzv. proofreading, teda čitateľskú kontrolu. Tú robia dobrovoľníci, ktorí sa môžu prihlásiť skrze projekt Distributed Proofreaders. Proofreader si na začiatku zvolí knihu, ktorú by rád kontroloval a dostane jednu stránku knihy. Na jednej strane je naskenovaný originál na strane druhej je text po OCR. Proofreader skontroluje text a opraví prípadné chyby. Potom môže podľa chuti pokračovať ďalšou stranou, alebo skončiť.

V závislosti na kvalite skeneru a od úspešnosti OCR je počet chýb na stránku približne 10, v prípade nekvalitnej niektorej (alebo oboch) zložky aj viac. Každá stránka podstúpi ešte druhý proofreading od niektorého zo skúsených užívateľov.

---

[4] Medzi inými napríklad aj druhú odmocninu čísla 2 na 5 miliónov desatinných miest, alebo popis ľudských chromozómov.

[5] Údaj z 28. novembra 2013.

[6] http://dbpearsonmlis.com/ProjectGutenbergMarcRecords.html [28.11.2013]. Zber metadát v auguste 2012.

Takto upravená (e)kniha už dosahuje presnosť prepisu 99.95%, čo je štandardom Kongresovej Knižnice. Pri proofreadingu sa nekontroluje iba text, ale čiastočne aj formát. Teda proofreader môže upraviť aj text podľa štandardov používaných PG, ktoré dostane v e-maily spolu s ostatnými informáciami pri registrácii.

Spracovanie jednej knihy, ktoré zahŕňa výber, kontrolu autorských práv, skenovanie, proofreading a formátovanie zaberie približne 50 hodín. Napriek tomu v októbri 2010 bolo po proofreadingu 18 848 eKníh.

Základný text knihy je v už spomínanom Plain Vanilla ASCII. Mnoho organizácií ale poskytuje konverziu do iných formátov. Autorom projektu nie je dokonca cudzia ani myšlienka prekladu do Barilovho písma či hlasu. S predpokladom, že do 10 rokov bude strojový preklad na dostatočnej úrovni a schopný prekladať medzi dostatočným počtom jazykov, majú autori víziu skutočnej multilinguality[7] (na rozdiel od súčasných zanedbateľných počtov diel v niektorých jazykoch).

Copyright

PG poskytuje diela všetkým zdarma. Súčasná politika to ale nedovoľuje pre všetky knihy, preto PG pozorne vyberá diela, ktoré bude digitalizovať a ponúkať svetu. Neposkytuje žiadne diela, pri ktorých by mohol byť problém s autorským zákonom. Pred stiahnutím diela tiež doporučuje užívateľom skontrolovať, či vyžadované dielo je voľne šíriteľné pre krajinu, v ktorej sa nachádzajú.

Autorský zákon v USA bol za posledné desaťročia často menený. Od niekdajších 14 rokov od vydania (s možnosťou predĺženia o ďalších 14 rokov), až po 50 rokov od smrti autora[8]. To prakticky znamená, že PG môže v USA voľne digitalizovať a šíriť iba diela vydané pred rokom 1923. Niektoré knihy, na ktoré sa ešte nevzťahuje úprava zákona medzi rokmi 1923 – 1964 ešte voľne šíriteľné byť môžu. Odhaduje sa, že ich počet je približne jeden milión. Tie sa snaží PG vyhľadávať a digitalizovať skôr, než bude neskoro.
Jeden z výrokov Michaela Harta v júli 1999 znie:

„No one has said more against copyright extensions than I have, but Hollywood and the big publishers have seen to it that our Congress won't even mention it in public. The kind of copyright debate going on is totally impractical. It is run by and for the 'Landed Gentry of the Information Age.' 'Information Age'? For whom?“

Autorské zákony ale samozrejme neplatia iba v USA, ale aj v iných krajinách. S prihliadnutím k nim musia pracovať ako sesterské projekty mimo USA, tak aj dobrovoľníci prispievajúci svojou činnosťou a užívatelia „profitujúci“ zo snahy PG.

Partnerské projekty

Existuje niekoľko partnerských projektov PG. Prvým bol PG Austrália (2001), ďalej napríklad PG Europe , PG Canada, PG Portugal, PG Philippines, Self publishing portal, nordická

[7] Myšlienka prekladu do Barilovho písma alebo iných jazykov je z práce z roku 2008. V inej práci z roku 2010 autorka uviedla víziu plne funkčného strojového prekladu v roku 2020 (v nadpise uviedla *„may be“*)

[8] Teda ak napríklad autor vydá knihu v 25 rokoch, a zomrie vo veku 75 rokov, kniha vstúpi do verejnej sféry až 100 rokov od jej vydania.

literatúra v podobne Projekt Runneberg a mnohé iné[9]. Taktiež existujú projekty spojené s proofreadingom ako napríklad Distributed Proofreaders Canada alebo Europe.

Dobrovoľníctvo

Celý projekt je založený na dobrovoľníkoch. K tisícom dobrovoľníkov sa môže zapojiť každý, každý môže pomôcť a nik nie je obmedzovaný v tom, čo, koľko a ako to bude robiť. PG si uvedomuje, že dobrovoľníci sú skutočne dobrovoľníci, a preto všetko, čo im ponúkajú sú doporučenia v ich práci[10].

Forma pomoci je rôzna. Pre užívateľov, ktorý nemajú prístup k vysokorýchlostnému internetu PG ponúka možnosť zaslať CD alebo DVD s výberom kníh. Zasielajú ich dobrovoľníci pre príslušnú oblasť. Stačí prejaviť záujem a počkať na odpoveď, či pre vašu oblasť potrebujú autori ešte dobrovoľníka. V prípade nutnosti je k dispozícii aj možnosť preplatenia nákladov na CD/DVD, ktoré napálite.

Ďalšou možnosťou je napríklad proofreading. Stačí sa registrovať v niektorom z Distributed Proofreaders projekte a po zaslaní e-mailu s inštrukciami sa môžete pustiť do kontroly skenovaných kníh. PG doporučuje svojim dobrovoľníkom, aby overili aspoň jednu stranu denne. Nijako to však od nich nevyžaduje. Napriek tomu to funguje celkom spoľahlivo a denne pribúdajú tisíce strán.

Projekt je možné podporiť aj finančne, či už priamou cestou a to poukázaním nejakého finančného obnosu na účet projektu, alebo nepriamo ako napríklad nepreplácaním nákladov spojených s rozširovaním diel.

Záver

Základná myšlienka a filozofia PG je podľa môjho názoru výborná. Každý by mal mať prístup k literatúre. Voľne prístupné vedecké články by značne prispeli k rozvoju vedy. Voľne prístupné knihy zas k sčítanosti a gramotnosti ľudí. Jednoduchosť použitia a prístupnosť týchto zdrojov je jedným zo základných rysov, ktoré by mala digitálna knižnica spĺňať. PG ako najstaršia DL má všetko z toho.

Napriek prekážkam, ktoré projektu v súčasnosti kladie autorský zákon si myslím, že výsledky sú výborné. V projekte existuje široká sieť dobrovoľníkov, základné fungujúce postupy aj užívatelia využívajúci poskytované služby. PG poskytuje voľné knihy pre kohokoľvek a postupne začleňuje rôzne jazyky. Jeho filozofia nemá dopad iba na knihovnícky svet, ale aj na zmýšľanie ľudí v spoločnosti.

Nesúhlasím ale so všetkými myšlienkami, ktoré v prácach boli vyslovené. Napríklad názor, že tlačené knihy nemôžu predstavovať konkurenciu pre elektronické texty, pokiaľ niekto objaví ich pohodlnosť. Predpoklad, že tlačené knihy už nebudú existovať sa už vyskytol mnohokrát, no zatiaľ sa to nepotvrdilo a ani to nevyzerá, že by sa k tomu schyľovalo. Tlačené knihy majú svoje čaro a mnohí na nich nedajú dopustiť.

[9] Existuje aj PG DE, ktorý môže niesť oficiálne meno projektu, avšak niektorí ho nepovažujú za sesterský projekt, nakoľko diela nie sú voľne šíriteľné a prístup k nim je limitovaný.
[10] S výnimkou niektorých základných vecí ako napríklad formátovanie textov pre čitateľov.

Literatúra

Partners, Affiliates and Resources. [cit. 28.11.2013]. Dostupné na: http://www.gutenberg.org/wiki/Gutenberg:Partners,_Affiliates_and_Resources.

Statistics, 2011. [cit. 28.11.2913]. Dostupné na: http://www.gutenbergnews.org/statistics/.

Michael S. Hart and Gregory B. Newby. Project Gutenberg Principle of Minimal Regulation / Administration, 2004. [cit. 28.11.2013]. Dostupné na: http://www.gutenberg.org/wiki/ Gutenberg:Project_Gutenberg_Principle_of_Minimal_Regulation_/_Administration_by_Michael_Hart_and_Greg_Newby.

Michael S. Hart. The History and Philosophy of Project Gutenberg, 1992. [cit. 28.11.2013]. Dostupné na: http://www.gutenberg.org/wiki/Gutenberg:The_History_and_Philosophy_of_Project_Gutenberg_by_Michael_Hart.

Michael S. Hart. Project Gutenberg Mission Statement, 2004. [cit. 28.11.2013]. Dostupné na: http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Mission_Statement_by_Michael_Hart.

Michael S. Hart. Administrivia 2004. [cit. 28.11.2013]. Dostupné na: http://www.Gutenberg.org/wiki/Gutenberg:Administrivia_by_Michael_Hart.

Marie Leberet. 40 years / 40 años / 40 ans, 2010. [cit. 28.11.2013]. Dostupné na: http://www.gutenberg.org/ebooks/34731.

Marie Leberet. Project Gutenberg (1971-2008), 2008. [cit. 28.11.2013]. Dostupné na: http://www.gutenberg.org/ebooks/27045.

Marie Leberet. Project Gutenberg (1971-2009), 2010. [cit. 28.11.2013]. Dostupné na: http://www.gutenberg.org/ebooks/31632.

Marie Leberet. From the Print Media to the Internet, 2008. [cit. 28.11.2013]. Dostupné na: http:// www.gutenberg.org/ebooks/27030.

Wikipedia contributors. Project Gutenberg. Wikipedia, The Free Encyclopedia. 2013 Nov 26, 12:20 UTC [cit. 28.11.2013]. Dostupné na: http://en.wikipedia.org/w/index.php? title=Project_Gutenberg&oldid=583375687.