Esej do předmětu PV070 Digitální knihovny

Universal Library
Million Book Project

Carnegie Mellon University
http://tera-3.ul.cs.cmu.edu/

Vítězslav Číp (2.semestr) FI-MU, 19.12.2005

Charakteristika projektu

Univerzální knihovna si klade za cíl zdigitalizovat všechny významné literární, umělecké a vědecké práce. Poučeni z historie různými nehodami, jako např. požárem Alexandrijské knihovny, chtějí autoři uchovat tato díla a uchránit je zubu času a dalších nehod.

Dále autoři chtějí odstranit nevýhodu dnešních knihoven a některých jejich svazků, které jsou velmi vzácné a dostupné jen malému okruhu zasvěcenců. Digitální technologie umožní neustálý přístup k těmto pracím milionům lidí po celém světě. Celý projekt má poskytnout tuto službu zdarma pro všechny lidi a tím rozšířit jejich vzdělání. Hlavní část tohoto projektu se jmenuje Million Book Projekt a má za úkol vybrat, naskenovat a zpřístupnit přes internet milion knih.

Aktuální stav

Projekt chce naskenovat milion knih do roku 2007. V současné době běží 22 skenovacích center v Indii a 18 center v Číně. Dále jsou materiály skenovány v Egyptě, Havaji a Carnegie Mellon. Do listopadu 2005 bylo naskenováno přes 600,000 knih: 170,000 v Indii, 420,000 v Číně a 20,000 v Egyptě. Zhruba 135,000 knih je v angličtině, zbytek je v indštině, čínštině, arabštině, francouzštině a dalších jazycích. Většina knih byla volně přístupná a asi u 60,000 kusů bylo třeba vyjednat svolení vlastníků práv. Knihy budou dostupné na serverech v Indii, Číně, Mellon Carnegie, Internet Archive a na daších místech. Žádná ze zatím naskenovaných knih není dostupná.

Cíle projektu

Primárním úkolem je převedení a zachycení všech knih do digitální podoby. Mnozí to považují za nemožné, myslí, že to zabere stovky let a nikdy to nebude dokončeno. Prvním krokem je proto zdigitalizování jednoho milionu knih (což je méně než jedno procento knih, které byly kdy vydány) do roku 2007. Cílem projektu není pouze naskenování a vystavení knih, ale také řešení problému s tím spojených. Jde o podporu dalších, kteří se zabývají vylepšováním skenovacích technik, lepším OCR a indexováním.

Popis projektu a jeho výsledky

Projekt je vedený universitou Carnegie Mellon. Spolupracuje na něm s vládou a partnery z Číny a Indie. Financování projektu zajišťuje The National Science Foundation (NSF), která přispívá 3,6 milionu dolarů po čtyři roky na různá zařízení a přepravu knih. Indie přispívá 25 milionů dolarů ročně na výzkum a překlady. Ministerstvo školství Číny přispívá 8,46 milionu dolarů. Mezi další sponzory patří Internet Archive a The University of California Libraries v Merced, která se stará o zajištění autorských práv.

Nyní bych rád popsal, jak celý projekt funguje a co všechno se muselo řešit. Začnu nejdříve s technickými detaily. Jako první je problém vytvoření databáze. Plánuje se centrální virtuální databáze s mnoha mirrory v několika zemích. Měla by tak být dostatečně zajištěna dostupnost a bezpečnost. Databáze by měla obsahovat textový i obrazový záznam knihy a počítá se s obsazením asi 50-60 MB na knížku. Jelikož stále klesá cena za uskladnění dat, neměl by v tomto bodě být výraznější problém.

Jako další následuje skenování. V dřívějších fázích se používal levný duplxní skener, který byl složitý na zacházení a vyžadoval jisté úpravy knihy, jako je ořezávání stránek. Často se také díky prachu z knih zasekával. Nyní se používá Minolta PS7000, která je na tři směny schopná naskenovat 10,000 stránek. Je sice 5 krát dražší než původní skener, přesto poskytuje tolik výhod, že se její používání vyplatí. Stručně shrnuto jsou to:

- Není třeba rozkládat knihy a ořezávat stránky
- Knihy mohou být po naskenování normálně používány

- Dobře pracuje i s prašnými knihami a knihami různých rozměrů
- Dobře se obsluhuje
- Je spolehlivá

Data se skenují do archivní kvality 600 bodů na palec jako binární obraz 1 bit na pixel. Obraz je uložen ve formátu TIFF.

Po naskenování se řešil problém s OCR (Optical Charakter Recognition). Projekt si neklade za cíl vylepšení výstupu. Používá OCR pro hledání uvnitř textu. S jeho pomocí bude vytvořen index pro vyhledávání. OCR funguje na 98 procent, což při častém opakování slov umožní vyhledat odpovídající pasáž. Zajímavé byly také problémy s OCR pro jiné jazyky a abecedy (Jen v Indii je na 17 různých abeced).

Abychom mohli s digitální knihou efektivně pracovat, musíme ji nastavit metadata. Většina metadat se dá stáhnout z již existujících popisků z knihoven. Knihovny v Carnegie Mellon již vyvinuly software, který používá standardní protokol Z39.50 k vyhledávání a získání metadat z katalogů. Ruční zadávání by bylo velmi časově náročné.

Celý proces naskenování knihy zahrnuje vyhledání knihovny, která vlastní danou knihu, zapůjčení knihy, doprava na konkrétní skenovací místo, naskenování a vrácení knihy zpět do knihovny. Otázka vhodné dopravy byla také předmětem mnoha diskusí a nakonec byla zvolená přeprava letadly, neboť tím dochází k menšímu poškození knih. Jelikož na tomto projektu spolupracuje i Čína a Indie, jsou některé dopravní vzdáleností obrovské a čím déle je v kniha v transportu, tím více se poškodí. Na závěr se ještě zmíním o výběru knih. Měly by to být jedinečné historické záznamy, vládní dokumenty, technické dokumentace a práce a různé výběry z univerzitních knihoven. Bohužel se mi nepodařilo vypátrat, podle jakého klíče se konkrétní výběry dělají.

Zhodnocení a přínos

Celá myšlenka Universal Library je velmi dobrá a chvályhodná. Bohužel kromě technických problémů musí řešit, asi jako každý podobný projekt, i problémy s autorskými právy, což v mnoha případech zpomaluje nebo komplikuje práci. Já osobně se velmi těším až budu mít možnost podívat se na webu na např. naše staré kroniky nebo jiné historické knihy, ke kterým se jinak normální smrtelník nedostane. Řadě lidí se tím i zrychlí a usnadní práce, neboť i vyhledávání v elektronické podobě je nesrovnatelné. Tím se dostávám i k zlepšení studia díky zvýšené dostupnosti materiálů. Další výhodou je, že digitální knihovna může mít otevřeno 24 hodin denně, čehož normální knihovna dosáhne asi těžko.

Zdroje a odkazy

http://www.library.cmu.edu/Libraries/MBP_FAQ.html (Frequently Asked Questions)

http://tera-3.ul.cs.cmu.edu/ (the Universal Library)

http://www.ulib.org.cn/ (Universal Library, China site)

http://dli.iiit.ac.in/ (Digital Library of India)

http://www.archive.org/details/millionbooks (the archived pilot)

http://en.wikipedia.org/wiki/Million_Book_Project

Metadata v Dublin Core

Dublin Core Atribut	Schéma	Hodnota
DC.Title		Universal Library
DC.Creator		Číp, Vítězslav
DC.Creator.Address		60910@mail.muni.cz
DC.Subject		Universal Library
DC.Subject		The Million Book Project
DC.Subject		Digital Library
DC.Data.Created	W3CDTF	2005-12-19
DC.Type		Text
DC.Format	IMT	aplication/pdf
DC.Language	RFC3066	cze
DC.Source	URL	http://www.library.cmu.edu/Libraries/MBP_FAQ.html
DC.Source	URL	http://tera-3.ul.cs.cmu.edu/
DC.Source	URL	http://www.ulib.org.cn/
DC.Source	URL	http://dli.iiit.ac.in/
DC.Source	URL	http://www.archive.org/details/millionbooks
DC.Source	URL	http://en.wikipedia.org/wiki/Million_Book_Project