Masarykova univerzita
Fakulta informatiky

HathiTrust

http://www.hathitrust.org/

Barbora Bajtošová
3.ročník (5.semester)

Brno, december 2012

1

HathiTrust

Úvod

HathiTrust na jednej strane predstavuje medzinárodnú spoločnosť výskumných inštitúcií a knižníc, ktorej cieľom je zachovanie kultúrneho dedičstva pre budúce generácie, ale na strane druhej ide o veľkokapacitný repozitár (úschovňu) digitálnych kolekcií (knihy, periodiká, multimediálne súbory, apod.) jednotlivých členov, spolu s materiálmi, ktoré boli zdigitalizované prostredníctvom projektu Google Books či Internet Archive.

História a správa

HathiTrust vznikol v októbri 2008 ako spoločné dielo Committee on Institutional Cooperation (13 amerických univerzít), University of California a University of Virginia za účelom vytvorenia repozitára na archiváciu a zdieľanie ich digitálnych kolekcií. HathiTrust sa začal veľmi rýchlo rozširovať a v súčasnosti má už 60 členov. Pridať sa môžu nielen inštitúcie s rozsiahlymi kolekciami, ale aj záujemcovia o dlhodobejšiu spoluprácu a pomoc pri spravovaní dát, výmenou za pokročilejšie služby pri prístupe a spracovaní dát z digitálnych knižníc.

Keďže projekt takýchto rozmerov vyžaduje nemalé finančné prostriedky či služby a tieto potreby zabezpečujú v rôznej miere jednotliví partneri, bolo potrebné ustanoviť určitú formu správcovstva. V súčasnosti spravuje HathiTrust 12-členná rada „Board of Governors“, ktorú tvorí 6 volených členov a 6 členov, ktorých ustanovujú zakladajúce inštitúcie.

Ciele

Hlavným cieľom HathiTrust je prispieť k „všeobecnému dobru“ zbieraním, katalogizáciou, zachovávaním a zdieľaním záznamov ľudských znalostí.

Ďalšie ciele:

- Vybudovať spoľahlivý archív pre digitalizovaný materiál, ktorý bude vlastnený a spravovaný akademickými inštitúciami.
- Vylepšiť prístup k týmto materiálom takým spôsobom, aby v prvom rade spĺňal potreby partnerských inštitúcií.
- Pomôcť zachovať dôležité záznamy vytvorením spoľahlivých a prístupných elektronických reprezentácií.
- Z dlhodobého hľadiska zredukovať náklady knižníc spojené s úschovou a starostlivosťou o diela v tlačenej podobe.
- Vytvoriť prostredie, ktoré je nielen vnímavé k potrebám užívatelov, ale aj dostatočne otvorené na vytváranie nových nástrojov a služieb.

Technológia

Projekt takýchto rozmerov si vyžaduje nielen finančné prostriedky, ale aj efektívne, spoľahlivé a zároveň bezpečné technologické spracovanie dát, ktorých množstvo sa momentálne pohybuje okolo 474 terabytov a stále sa zvyšuje. Pre zvýšenie bezpečnosti (aby nedošlo k strate/poškodeniu dát) sa využívajú dve dátové centrá, ktoré od seba delí značná geografická vzdialenosť (Ann Arbor (Michigane) a Indianapolise (Indiana)), ale ktoré sú stále synchronizované. Samozrejmosťou sú zálohy – ide o zakódované pásky obsahujúce dáta za posledných šesť mesiacov uložené v samostatnom priestore niekoľko kilometrov od Ann Arbor, a ku ktorým má prístup len autorizovaný technický personál. Primárnymi médiami na ukladanie dát sú rotačné disky, keďže je potrebná neustála kontrola integrity dát na disku. Pre vnútornú štruktúru ukladania dát na disk nevyužívajú v súčasnosti populárnu schému RAID5/6, ale dali prednosť analogickému riešeniu *N+3 Reed-Solomon parity redundancy check*, ktoré oplýva väčšou chybovou odolnosťou ako klasické RAID5 vďaka dodatočnej redundancii. Ukladací systém je virtualizovaný, so súbormi rozdelenými do blokov, ktoré sú rozdistribuované do jednotlivých uzlov clustru a automaticky predistribuované podľa potreby. Táto štruktúra umožňuje nielen jednoduchšiu správu systému, ale aj výmenu (odstránenie/pridanie) uzlov clustru iba administrátorským príkazom, pričom presun dát prebieha počas priebežnej kontroly integrity. Okrem spomínaných priebežných (interných) kontrol, HathiTrust taktiež vykonáva kontroly mimo samotný ukladací priestor pomocou uložených kontrolných súčtov (checksum), aby sa uistili, že dáta neboli pozmenené a či nové dáta boli korektne spracované.

Obsah

V súčasnosti (údaje z 8.12.2012) HathiTrust obsahuje 10,588,871 zväzkov (volumes) a z toho približne 31% je prístupných v sekcii *public domain*. Pojmom public domain sa označuje digitálny obsah, ktorý nepodlieha autorskému zákonu, a ktorý je teda možné voľne šíriť, kopírovať a využívať, alebo obsah publikovaný pod licenciou, ktorá umožňuje jeho prehliadanie. Veľkú časť tvoria diela publikované v USA pred rokom 1923 a vládne dokumenty, ktoré taktiež spadajú do kategórie public domain.

HathiTrust je prístupný komukoľvek s prístupom na internet, ale primárne určený pre študentov (prevažne humanitných oborov), na akademickú/výskumnú prácu a keďže veľkú časť tvoria historické diela, poskytuje dobré zázemie pre ľudí so záujmom o genealógiu či históriu. Spomeniem napríklad kolekciu vybraných kníh o psychológii z 19. a 20. storočia, či diela Sigmunda Freuda, ktoré aj v súčasnosti zohrávajú významnú rolu pri štúdiu psychológie. V public domain nenájdeme len knihy, ale aj rôzne vedecké časopisy alebo periodiká, ktoré už síce dávno prestali vychádzať (The Gentleman's Magazine), ale ich články stále dokážu zaujať, alebo aj napríklad časopis *Cosmopolitan*, ktorý v modernejšej podobe vychádza dodnes.

Prehliadanie dokumentov

Každý titul (či už ide o knihu alebo časopis) je zobrazený v prehliadači prostredníctvom aplikácie PageTurner, ktorá umožňuje komfortné prehliadanie podľa užívateľových preferencií ako sú napríklad spôsob zobrazenia, full-screen, zoom, vyhľadávanie či prechod na určitú stranu. Okrem samotného zobrazenia sú vždy v ľavej časti stránky uvedené ďalšie informácie ako copyright – pod akou licenciou bolo dielo vydané a tým pádom aké sú možnosti využitia, permanentný link, odkaz na podrobný katalógový záznam o tomto diele. V prípade, že si necháte zobraziť katalógový záznam, zistíte nielen informácie o titule, ale môžete si tiež nechať zobraziť citáciu (čo je veľmi užitočné hlavne pre študentov, ktorí nemajú veľké skúsenosti s citovaním). Hoci je možné stiahnuť konkrétnu stranu alebo v prípade, že na to máte právo (sekcia prístup) aj celú knihu, stále je väčšia časť titulov prístupná iba v režime *search-only*, ktorý síce neumožňuje si titul prezerať priamo, ale umožňuje vyhľadávať frázy, a následne vráti všetky strany, na ktorých sa hľadaný výraz vyskytol. Týmto spôsobom si môžete aspoň čiastočne uľahčiť rozhodovanie, či sa oplatí si tento titul kúpiť alebo si nechať priamo vyhľadať, v ktorej knižnici je možné si ho požičať. Túto, dalo by sa povedať „nadštandardnú“ funkcionalitu (vyhľadanie knižníc) neposkytuje priamo Ha-

thiTrust, ale ste presmerovaný na stránky najväčšej siete knižníc *WorldCat*, ktorej je HathiTrust členom. Stačí zadať vašu polohu a zobrazia sa najbližšie knižnice s hľadaným titulom. Členom je napríklad aj *Národní knihovna ČR*.

V obmedzenom režime (bez full-text vyhľadávania a prehliadania kolekcií) funguje aj prístup prostredníctvom mobilných telefónov.

Vyhľadávanie

Užívateľ má k dispozícií štyri rôzne spôsoby vyhľadávania, pričom by sa dalo povedať, že každá z možností je založená na miere užívateľových znalostí o tom, čo hľadá.
Katalógové vyhľadávanie: Vyhľadávanie na základe bibliografických údajov (autor, titul, dátum vydania, ...). Na vyhľadávanie presného citátu/frázy sa používajú zátvorky, na vyhľadanie všetkých možných foriem sa používa * pre viacero znakov a ? pre jeden znak. Taktiež je možné využiť AND/OR medzi slovami na boolovské vyhľadávanie – napr. (heart OR cardio) AND surgery vráti súbory zaoberajúce sa heart surgery a cardio surgery.
Full-Text vyhľadávanie: Vyhľadávanie kľúčových slov vo všetkých súboroch obsiahnutých v HathiTrust. Rovnaké podmienky ako pri katalógovom vyhľadávaní.
Collection Builder: Vyhľadávanie v rámci kolekcií.
Single-volume vyhľadávanie: Ide o vyhľadávanie v zvolenom dokumente, ktorý je zobrazený prostredníctvom aplikácie PageTurner.

Prístup

Rozlišujú sa štyri základné kategórie užívateľov s rôznymi obmedzeniami, či výhodami.

Základnou kategóriu sú *obyčajní užívatelia* (Ordinary Users – ORD), ktorí nie sú prihlásení do systému, nepoužívajú počítač s autorizovanou IP adresou, alebo užívatelia s tzv. „friend“ kontom, ktoré autentizuje University of Michigan (UM). Títo užívatelia nemajú možnosť prehliadať alebo sťahovať celé dokumenty zo sekcie public domain (uvedené nižšie), ak sa nejedná o diela pod Creative Commons licenciou. Pre užívateľov s „friend“ kontom platia rovnaké obmedzenia, ale môžu vytvárať vlastné permanentné kolekcie.
Užívateľ v knižnici: Ide o užívateľov, ktorí na prístup využívajú počítač so špecifickou IP adresou nachádzajúci sa v knižnici. Momentálne ide len o knižnicu UM.
„Priateľ“ University of Michigan: Túto skupinu tvoria študenti, personál a fakulty University of Michigan. Užívatelia majú špeciálne HathiTrust privilégia prístupu.
Partneri HathiTrust: Skupinu tvoria užívatelia z partnerských inštitúcií HathiTrust, ktorí sa musia autentizovať prostredníctvom Shibboleth (ide o mechanizmus na autentizáciu užívateľov medzi spolupracujúcimi inštitúciami).
Špecifickú skupinu tvoria *znevýhodnení užívatelia* (Print-disabled users).

Zhodnotenie

Už pri zbežnom pohľade, je zrejmé, že ide o projekt, do ktorého bol investovaný čas, peniaze, ale aj ľudské zdroje. Webové rozhranie bolo vytvorené s cieľom poskytnúť užívateľovi čo najjednoduchší prístup, o čom svedčí aj úvodná stránka, ktorá poskytuje priamy prístup k vyhľadávačom. Využila som všetky dostupné formy vyhľadávania a musím konštatovať, že aj pri zložitejších požiadavkách (nielen bibliografické údaje) prebehlo vyhľadávanie rýchlo a výsledky boli (väčšinou) relevantné. Pozitívne hodnotím aj množstvo informácií – takmer všetko, čo by vás mohlo o HathiTrust zaujímať (technické informácie, cena, správa,...), nájdete na stránkach (alebo je uvedená adresa). Za nevýhodu však považujem obmedzený prístup a funkcionalitu poskytovanú užívateľom mimo partnerské inštitúcie (v súčasnosti sa jedná prevažne o americké univerzity/inštitúcie).

Veľkým prínosom a podľa mňa najväčším pozitívom tohto projektu je, že sa nielen snaží o zachovanie nášho kultúrneho dedičstva, ale že nám ho aj sprístupňuje spôsobom, ktorý je dostupný

každému. Vďaka HathiTrust má človek možnosť študovať texty z celého sveta, ktoré sú inde neprístupné alebo sú vzácne či krehké.

Z môjho pohľadu sa jedná o odvážny projekt, ktorý však (na rozdiel od množstva iných) spĺňa stanovené ciele a hľadá nové možnosti ako ich dosiahnuť, čím sa posúva vpred. Nepochybujem o tom, že členov bude iba pribúdať a dúfam, že sa v blízkej budúcnosti dostane aj do povedomia európskych študentov.

Štatistické informácie

Štatistické informácie zverejnené k 8.12.2012

- 10,588,871 total volumes
- 5,526,795 book titles
- 274,657 serial titles
- 3,706,104,850 pages
- 475 terabytes
- 125 miles
- 8,603 tons

Metadata

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="HathiTrust" />
<meta name="DC.Creator" content="Barbora Bajtošová" />
<meta name="DC.Subject" content="HathiTrust" />
<meta name="DC.Subject" content="Digitálne knižnice" />
<meta name="DC.Subject" content="archív" />
<meta name="DC.Date.available" content="4.12.2012" />
<meta name="DC.Type" scheme="DCMIType" content="Text" />
<meta name="DC.Format" scheme="IMT" content="application/pdf" />
<meta name="DC.Format.medium" content="computerFile" />
<meta name="DC.Source" scheme="URL" content="http://www.hathitrust.org/" />
<meta name="DC.Language" scheme="RFC3066" content="slo" />
```

Zdroje

http://www.hathitrust.org/
http://www.hathitrust.org/mission_goals
http://www.hathitrust.org/help
http://www.hathitrust.org/technology http://en.wikipedia.org/wiki/HathiTrust
http://www.itcs.umich.edu/itcsdocs/s4316/
http://www.worldcat.org/
http://www.libraryjournal.com/lj/communityacademiclibraries/890917-419/unlocking_hathitrust_inside_the_librarians.html.csp