Velká data v knihovnách

Open source tools and their use in Czech libraries

Petr Žabička
Moravská zemská knihovna v Brně
www.mzk.cz

Obsah

Informace a knihovny

- Knihovny mají tisícilenou tradici uchovávání, zpracování, vyhledávání a zpřístupňování informací na analogových nosičích...
- ... a relativně krátkou dobu i informací v podobě digitální.

Bibliografický záznam

- Bibliogafický standard MARC vznikl v 60. letech...
  - Z39.2 -> ISO 2709: Documentation – Format for bibliographic information interchange on magnetic tape
- ...a stále se nemá k odchodu do důchodu
  - http://**MARC-must-die**.info/

Bibliografický záznam

- Bibliogafický standard MARC vznikl v 60. letech...
  - Z39.2 -> ISO 2709: Documentation – Format for bibliographic information interchange on magnetic tape
- ...a stále se nemá k odchodu do důchodu
  - http://**MARC-must-die**.info/
- Problém: setrvačnost
  - Cca 1 milion knihoven
  - Miliardy bibliografických záznamů (převážně v MARCu)

Bibliografikcý záznam

- ČR: téměř 10 tis. knihoven
- Souborný katalog ČR:
  - 384 knihoven
  - 12,8 mil. svazků = 5,8 mil. titulů
- Největší knihovny mají do databázové podoby převedeno jen cca 50% svých lístkových katalogů
- Kvalita záznamů odpovídá kvalitě katalogizačních záznamů...

Antonius a Resurrectione Domini, Augustiner [illegible] Prior des h. Convents bei Maria Loretto in Wien 1732.

Das neu [illegible] glückselige Bruderschafts-Buch: oder eine schuldige Lob- u. Ehrenrede von der Bruderschaft U. L. F. [illegible] ein glückseliges [illegible], da selbe [illegible] hundertjährigen [illegible] im [illegible] zum [illegible] gemeinet, durch 3 Tage [illegible]. Am 3. Tage [illegible] [illegible] vorgetragen durch P. Antonium a Resurrectione Dni. Vid. Bruderschaft Maria Wien zum [illegible] 100jährigen Alter. B.Nr. 14533. p. 86–106.

Bazali, M... 7080

Die Berechnung der Betonkanäle. Vereinfachte Formeln zur Berechnung der Kanal[illegible] [illegible]digkeit und des Gefäll[illegible] [illegible] Leschke, 1909. 8°

R. 1/3 [illegible] Mk 2.0[illegible]

Bau, Alexander | 33153

Handbuch für Käfer-Sammler. Beschreibung der in Deutschland, Oesterreich-Ungarn und der Schweiz vorkommenden Coleopteren von Alexander Bau. Mit 144 naturgetreuen Zeichnungen im Text.

Magdeburg, Creutz'sche Verlagsbuchhandlung, 1888.

[illegible] 49456.

58547/50

52. [illegible]

Souborný katalog ČR

- Roční přírůstek: zasláno cca 1,3 mil. záznamů
  - 200 tis. nepřijato kvůli chybám
  - 745 tis. úspěšně automaticky deduplikováno (SK nedrží jednotlivé dodané záznamy, jen
    - 640 tis. připsáno
    - 85 tis. přepsáno
  - 400 tis. přidáno jako nové
  - ruční práce:
    - 25 tis. deduplikováno,
    - 50 tis. Smazáno

Souborný katalog ČR

- Problém: jak identifikovat záznamy stejných knih vzniklé v různých knihovnách?
  - Absence identifikátorů (ISBN v ČR až od 1989)
  - Rozdíly ve způsobu zápisu (+překlepy)
  - Rozdíly v přístupu ke katalogizaci (vícesvazková díla, přívazky apod.)
  - Chyby v identifikátorech uvedených v záznamu
  - Stávající deduplikační procedury nelze vyladit lépe
    - chybovost vs. úspěšnost propojení
    - významný podíl ruční práce

ObalkyKnih.cz

- Zdroj pro obohacování záznamů v katalozích
- Poskytují náhledy obálek a obsahy knih knihovnám v celé ČR
- Nově i sdílené komentáře a hodnocení
- **917 719 obálek a 115 143 obsahů** (25.11.)
  českých a zahraničních publikací.
- Zdroje: nakladatelé, knihkupci, **knihovny**
- cca 1 TB dat
- Provozuje Jihočeská vědecká knihovna

OBÁLKY KNIH

ObalkyKnih.cz

- Převažuje kooperativní skenování (skenovací klient vyvinutý MZK, komunikace přes API)
  - 200-300 nahraných titulů / den
  - 4 GB dat obálek / den
  - OCR obsahů zajišťuje server
- Využíváno více než 180 knihovnami
  - Datový tok 20 Mbit/s
  - **1,5 mil. požadavků za den** (17 za sekundu)
- Open source systém, otevřený vývoj

ObalkyKnih.cz

- Problém: jak propojit naskenované či stažené obálky se záznamy v knihovním katalogu
- Role identifikátorů:
  - ISBN, ISSN, EAN
  - OCLC number
  - číslo České národní bibliografie (čČNB) – přiděluje Národní knihovna ČR – nutná zpětná synchronizace katalogů přes souborný katalog ČR

Digitalizace

- V ČR knihovnami naskenováno přes 34 mil. stran dokumentů
  - přes 120 tis. svazků
  - cca 10% celkové produkce vydané u nás
  - koordinace prostřednictvím Registru digitalizace
- digitální knihovna Kramerius (free, open source)
    - Jpeg 2000 + IIPImage; OCR: ALTO XML (ABBYY)
    - Solr (Lucene) index, Fedora Repository
  - digitální produkce (dobrovolně poskytovaná) – převažuje pdf, jinak problém s DRM
  - MZK vyvíjí open source klient pro Android

Digitalizace

- Národní digitální knihovna (NK + MZK)
  - Od konce 2012 skenováno cca 50 tis. stran **denně**
  - Ukládání v lossless jpeg2000 (LTO5 robot)
  - Zpřístupnění v lossy jpeg2000 (1:8 – 1:20)
    - 25 mil. stran v Krameriu = 220 GB Solr fulltext index
    - 25 mil. stran v Krameriu = 1,8 TB Fedora (OCR+metadata)
    - 25 mil. stran v Krameriu = 725 GB Postgres (triplet vazby)
  - V současnosti: 94000 monografií, 1025 periodik
  - Stále roste
  - Konce projektu: 2014 + 5 let udržitelnost (financování?)

Digitalizace

- Problémy:
  - jak dokumenty organizovat, zpřístupňovat
  - jak provádět aktualizace (např. nové verze OCR z lossless obrazových dat)
  - jak dokumenty třídit z věcného hlediska (nedostatečná metadata)
  - jak dokumenty dlouhodobě uchovávat (digital preservation)
  - jak nacházet/opravovat chyby v OCR
  - automatická konverze do epub apod.
  - autorský zákon

Digital born

- Problémy:
  - Digitální dokumenty na fyzických nosičích ve fondu knihoven (min. desítky tisíc nosičů, převážně CD, DVD)
  - Jak je spolehlivě přenést do digital preservation systému (včetně např. CD audio, DVD apod.)
  - Velká roztříštěnost formátů
  - Jak uchovávat software?
  - Co ebooky s DRM?
  - Elektronický „povinný výtisk“ ...?

WebArchiv

- Archiv českého webu od roku 2001
  (Internet Archive od 1996)
- Open source nástroje pro sklízení, indexaci, zpřístupnění
- Archivní formát arc, od 2012 warc (ISO 28500)
  - arc cca 100 MB, warc cca 1 GB
- Smlouvy o zpřístupnění (4200 webů)
- Primárně doména .cz (přes 1 mil. domén)

WebArchiv

- Sklízení českého webu
  - Heritrix 3, distribuovaně, deduplikace v rámci 1 roku
  - 87 TB v archivu, 626000 arc + 39000 warc souborů
  - Celkem 1,2 mld. URL
  - Uloženo na GPFS, úvahy + menší testy Hadoop
  - Není fulltextová indexace
  - 10-15 domén na 1 celoplošnou sklizeň domény .cz
  - Cca 5000 dotazů na doménu
  - 9 crawlerů vytvoří 10 TB archiv během 5,5 dne

WebArchiv

- Problémy:
  - Sklizně umírají na nedostatek místa nebo málo RAM
  - Sklízení mimo doménu .cz
    - nutnost dokončení vývoje WebAnalyzeru
  - Tvorba fulltextu při ukládání do stávajícího LTP systému = desítky let pro uložení stávajícího archivu
  - Pro zpřístupňování archivu nutné diskové úložiště
  - Jak archiv otevřít pro výzkum
  - Hledání pilotních záměrů pro jeho využití

Centrální portál knihoven

- Plánovaný portál pro zpřístupnění zdrojů knihoven
- Integrace metadat a ideálně i plných textů
- Předpokládané zahájení vývoje 2015 v MZK
  - Na bázi open source systému VuFind (jádro Solr index)
  - Obdoba finna.fi
  - Integrace zdrojů zejména velkých knihoven
  - Praktické testy:
    - VuFind.mzk.cz
    - CistBrno.cz
    - NarodniFonoteka.cz
    - HistorickeFondy.cz

Centrální portál knihoven

- Očekávané výzvy:
  - Správa sklízení značného množství zdrojů dat a metadat
  - Integrace různorodých zdrojů a jejich specifik
    - časová osa (vydání, platnost, o době)
    - plné texty vs. Metadata
    - geografické hledání (mapy, místa vydání, o místě)
    - smysluplné fasety (filtry)
    - obohacení záznamů o relevantní služby (přístup k dokumentu, digitalizace na vyžádání apod.)
  - Deduplikace (i na úrovni díla (?))
  - Jak legálně vytvářet a prohledávat fulltextový index zdroje, jehož plný text není možné získat (např. normy)?

…atd. atd.

Děkuji za pozornost!

Petr Žabička
petr.zabicka@mzk.cz

Moravská zemská knihovna v Brně
www.mzk.cz