Vyhledávání duplicit v bibliografických metadatech

PV070 Digitální knihovny

Michal Merta
FAKULTA INFORMATIKY, MASARYKOVA UNIVERZITA

5. prosince 2014

Abstrakt

Tato esej popisuje problematiku vyhledávání duplicit bibliografických metadat v centrálních indexech. Zaměřuje se především na rozdíly v přístupu Souborného katalogu a dvou projektů Moravské zemské knihovny (Číst Brno a Národní fonotéka). Text je doplněn o autorovy postřehy z praxe a příklady reálných problémů.

Klíčová slova: Bibliografický záznam, deduplikace, metadata, Souborný katalog, Číst Brno, Národní fonotéka

Motivace

Při vytváření digitálních knihoven hrají bibliografická metadata podstatnou roli. Pokud ale v rámci stavby takové knihovny dochází k agregaci metadat z různých na sobě nezávislých zdrojů, s velkou pravděpodobností bude výsledek obsahovat některé záznamy několikrát. Existují různé způsoby pro nakládání s duplicitami, nicméně jejich základem je vždy úspěšné rozpoznávání potenciálně duplicitních záznamů mezi všemi ostatními. Ve větším množství se jedná o zajímavou algoritmickou úlohu.

Nejdříve je ale potřeba definovat, co v našem kontextu myslíme duplicitou. Intuitivně se jedná o dva záznamy z různých zdrojů, které popisují tu samou věc. Formálněji lze použít definici: „Dva záznamy lze označit za duplicitní, pokud se oba vztahují k totožnému informačnímu zdroji“. [3] V praxi ale nelze záznamy proti sobě porovnávat čistě jen bajt po bajtu. Tento přístup by nebyl funkční minimálně z toho důvodu, že každý z nich může obsahovat jiné údaje, jiné identifikátory a jednotky. Je tedy potřeba nadefinovat části záznamů, které budou pro porovnávání využity. Ovšem vzhledem k jejich různorodosti a (ne)úplnosti je automatizace hledání duplicit obvykle kompromis mezi mezi kvalitou a kvantitou - pokud zvolíme příliš benevolentní přístup, ve výsledku se budou vyskytovat spárované nesouvisející záznamy. Při opačné strategii naopak zůstanou některé duplicity nespárované. Spárování záznamů znamená jejich sloučení do jednoho

společného záznamu, který obsahuje části ze všech původních. Tento proces je také obecně známý jako *deduplikace*. [8]

Na přednáškách jsme se seznámili s různými metadatovými formáty, ve zbytku textu budu uvažovat záznamy v MARCu 21. Důvod je čistě pragmatický, z mých zkušeností je to nejvíce používaný formát, který podporují všechny významné u nás používané knihovní systémy.

Aktuální situace

Hlavním hráčem na poli sběru bibliografických metadat je v České republice Souborný katalog, který aktuálně agreguje data z téměř čtyř set knihoven[1] na území České republiky a vytváří z nich centrální index. Tato data sbírá dávkově v nepravidelných intervalech[2]. Dalším významným projektem v této oblasti je u nás Jednotná informační brána, která oproti Soubornému katalogu využívá odlišný přístup. Místo budování centrálního indexu pokládá dotazy do vybraných katalogů a duplicity vyhledává až v získaných výsledcích na základě parametrů definovaných jednotlivými katalogy. [6]

Druhým rokem se agregaci metadat věnuje také Moravská zemská knihovna. V nedávné době byly veřejnosti zpřístupněny výsledky dvou vzájemně podobných projektů. Prvním z nich je portál Číst Brno[3], agregující bibliografická metadata z brněnských knihoven, které byly ochotné se do projektu zapojit. Druhým je projekt Národní fonotéky[4], který prezentuje metadata popisující zvukové záznamy. Svým záběrem navazuje na původní neúspěšné pokusy v této oblasti z konce devadesátých let minulého století. [1]

V následujících odstavcích se budu věnovat vyhledávání duplicit v centrálních indexech a dvěma odlišným přístupům. Filozofií Souborného katalogu je především korektnost. Jakýkoliv záznam prochází před zařazením do katalogu kvalitativní selekcí, která má za cíl odfiltrovat záznamy nesplňující dané požadavky (např. převeditelnost do formátu, který používá systém Aleph). [4] Oproti tomu Národní fonotéka má cíl přesně opačný a to získávat metadata o vzácných zvukových dílech (především šelakových deskách). V jejím centrálním indexu jsou agregována nejen metadata z knihovních fondů, ale také různé soukromé databáze (např. vydavatelství Supraphon). [7] Unifikace nejrůznějších proprietárních formátů je sama o sobě velmi zajímavou a obtížnou činností. Projekt Číst Brno v tomto kontextu tvoří střední cestu - agreguje pouze bibliografická metadata z velkých knihoven (které většinou poskytují záznamy v „relativně“ dobré kvalitě), ale proto Soubornému katalogu přináší hledání duplicit ve vícesvazkových dílech (např. dvousvazkový slovník A-L a M-Z je považován za duplicitní a uživateli zobrazen jako jeden sloučený záznam). [2]

[1] Podle statistik aktuálních k 3. listopadu 2014 je těchto knihoven 384.

[2] Pravidelné denní aktualizace probíhají pouze z Národní knihovny a Jihočeské vědecké knihovny.

[3] www.cistbrno.cz

[4] www.narodnifonoteka.cz

Základní prvky deduplikace

Využití identifikátorů

Základním přístupem při hledání duplicit je porovnání různých přidělených identifikátorů (ISBN, ISSN, NBN...). Myšlenka schovaná za tímto přístupem je prostá - záznamy se stejným identifikátorem lze považovat za duplicitní. V praxi ovšem vyšlo najevo, že jediným, alespoň částečně spolehlivým, identifikátorem je ISBN. V rámci Národní fonotéky jsme experimentovali s dalšími identifikátory (např. matriční a vydavatelská čísla). Ovšem ukázalo se, že jejich chybovost v záznamech je natolik zásadní, že se na ně nedá v žádném případě spoléhat a značně zvedají procento chybně deduplikovaných záznamů.

Souborný katalog využívá pro hledání duplicit ISBN a ISRC[5]. Při shodě na těchto identifikátorech ještě porovnává počet stran, rok vydání a několik znaků z názvu. [5] Zbylé dva portály považují ISBN za dostatečnou shodu, pokud mají záznamy stejný fyzický formát, viz dále. Číst Brno původně spoléhalo pouze na shodu formátu a ISBN. Za zmínku stojí, že tento přístup přinesl několik komplikací, když se mimo jiné na základě chybného ISBN deduplikovaly knihy *Žofka ředitelkou ZOO* a *Pán prstenů: Návrat krále*. Záznamů s problematickým ISBN se ve vstupních datech podařilo nalézt více než 10 000. Vzhledem k chybovosti záznamů tedy obecně není možné identifikátorům stoprocentně důvěřovat.

Detekce formátu

Nepostradatelnou součástí vyhledávání duplicit je identifikace fyzického formátu. Samozřejmě je také podstatná i pro samotnou prezentaci uživateli. Myšlenka při ověřování totožnosti fyzických médií je poměrně přímočará - chceme deduplikovat knihy s knihami, elektronické zdroje s elektronickými zdroji a podobně. Je ovšem otázka, jakým způsobem se tato situace bude vyvíjet v budoucnu. Je možné, že se objeví pokusy o deduplikaci relevantních záznamů napříč fyzickými médii. Například kniha může být klasicky v papírové formě, v elektronické verzi, v audio verzi, v Brailově písmu atd.

Souborný katalog detekci formátu ignoruje, pokud záznam obsahuje ISBN nebo ISRC, v opačném případě používá vnitřní pole systému Aleph FMT. [5] Obecně se v MARC21 používá pro detekci formátu pole 007, 008 a tzv. *leader*. [9] Pro Číst Brno ani fonotéku ale tyto údaje nejsou dostačující. Kvůli nižší kvalitě metadat oproti souhrnému katalogu je detekce často problematická a také chybná. Některé zdroje tato pole vůbec neobsahují nebo v nich je velké množství chyb. Navíc různé instituce informace o formátech uchovávají ve vlastních definovaných polích nebo je přidávají do identifikátoru záznamu. Ve výsledku jsme v rámci Národní fonotéky museli sáhnout k probabilistickým metodám založeným na regulárních výrazech. S každou další zapojenou institucí zpravidla přichází také potřeba úpravy stávajícího systému pro detekování formátů.

[5] v MARC21 se jedná o pole 020 a 024

Když identifikátory chybí

V praxi často nastává situace, kdy záznamy jednoznačné identifikátory jednoduše neobsahují. V takovém případě deduplikace probíhá na základě dostatečné podobnosti dvou záznamů.

V této situaci jsou přístupy Souborného katalogu i zmíněných dalších portálů velmi podobné. Srovnává se rok vydání, počet stran a normalizovaný[6] název. Souborný katalog vyžaduje jeho úplnou shodu. Číst Brno toleruje drobné odchylky kvůli potenciální chybovosti a povoluje desetiprocentní odchylku jejich Levensteinových vzdáleností.

Zvláštní podmínkou je, že duplicitní záznamy nepocházejí ze stejné instituce. Volně přeloženo, duplicity jsou vyhledávány jen v cizích záznamech. Na základě experimentování s deduplikačními algoritmy jsme se v rámci Číst Brno rozhodli od této podmínky upustit. To má za následek dvojí efekt. Prvně je to již zmíněná deduplikace vícesvazkových děl[7]. Druhým efektem je slučování duplicitních děl i v rámci jedné instituce. Pěkným příkladem je v tomto ohledu například na naší fakultě známá publikace *Linux: Praktický průvodce* od doc. Brandejse. V Číst Brno jsou deduplikovány záznamy čistě podle roků vydání. Oproti tomu portál *discovery.muni.cz* při vyhledávání vrátí pro rok vydání 1996 několik záznamů.

Přístup k deduplikaci použitý v Číst Brno je do značné míry kontroverzní, zejména mezi knihovníky. Můj osobní pocit je, že převládají spíše pozitivní ohlasy, ale možná jen kritici nejsou tolik slyšet. Jeho implementace je velmi závislá na kvalitě (nebo alespoň rozumné uniformnosti) agregovaných metadat. Pro korektní prezentaci vícesvazkových děl je důležité, aby záznamy obsahovaly korektní informace o částech díla v polích 245n a 245p. Toto se ale neděje vždy a výsledek může být pro uživatele matoucí.

Shrnutí

Souborný katalog navíc k nalezeným duplicitám přidává ještě ruční kontrolu knihovníkem. V kombinaci s nároky na kvalitu záznamu a dodatečnou kontrolou je pravděpodobnost chybně deduplikovaného záznamu velmi malá. Na druhou stranu Číst Brno přináší v oblasti deduplikace něco nového a snaží se uživatelům zpřehlednit vyhledávání i za cenu občasných chyb. Je také nutné podotknout, že tento projekt má za cíl mj. experimentovat s hledáním duplicit a testovat různé přístupy, které budou v budoucnu použity při vytváření centrálního portálu knihoven. Projekt Národní fonotéky je v kontextu deduplikace velmi specifický, právě kvůli různorodosti zdrojů metadat.

Na závěr bych rád zdůraznil, že drtivá většina problémů s deduplikací bibliografických metadat souvisí s nedostatečnou kvalitou metadat a velmi vysokou mírou chybovosti. Ta je z velké části způsobena lidským faktorem, např. v jisté významné české knihovně jistý katalogizátor dlouhodobě zaměňoval pole 505 a 520. Na tento problém jsme přišli v Zemské knihovně v rámci deduplikace.

[6] Řetězec zbavený diakritiky, interpunkce a velikosti písmen.

[7] např. https://www.cistbrno.cz/Search/Results?lookfor=Akademick%C3%BD+slovn%C3%ADk+ciz%C3%ADch+slov+&type=AllFields&limit=20&sort=relevance

Reference

[1] ŽABIČKA Petr, ŠÍR Filip. Virtuální národní fonotéka jako projekt moravské zemské knihovny v brně. *[ONLINE] http://duha.mzk.cz/clanky/virtualni-narodni-fonoteka-jako-projekt-moravske-zemske-knihovny-v-brne*, Listopad 2014.

[2] ŽABIČKA Petr, ŽABIČKOVÁ Petra. Číst brno - virtuální katalog brněnských knihoven. *[ONLINE] http://duha.mzk.cz/clanky/cist-brno-virtualni-katalog-brnenskych-knihoven*, Listopad 2014.

[3] SITAS Anestis, KAPIDAKIS Sarantos. Duplicate detection algorithms of bibliographic descriptions. *[ONLINE] http://users.ionio.gr/ sarantos/repository/j21J-LibraryHiTech-Sitas.pdf*, 2007.

[4] DVOŘÁKOVÁ Helena. Předimportní kontroly. *[ONLINE] http://www.caslin.cz/spoluprace/importy/predimportni-kontroly/*, Srpen 2007.

[5] DVOŘÁKOVÁ Helena. Deduplikační procedury. *[ONLINE] http://www.caslin.cz/spoluprace/importy/deduplikacni-procedury*, Květen 2009.

[6] Jednotná informační brána pro hybridní knihovny. *[ONLINE] http://info.jib.cz/o-projektu/projekt/brana projekt.*.

[7] Virtual national phonoteque of the czech republic. *[ONLINE] http://projekt.narodnifonoteka.cz/homeeng*, 2014.

[8] Understanding data deduplication. *[ONLINE] http://www.druva.com/blog/understanding-data deduplication/*, 2009.

[9] Marc 21 format for holdings data: Table of contents, *[ONLINE] http://www.loc.gov/marc/holdings/*, 2000.

Metadata

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="Vyhledávání duplicit
v bibliografických metadatech" />
<meta name="DC.Creator" content="Michal Merta" />
<meta name="DC.Description" content="Tato esej popisuje
problematiku vyhledávání duplicit bibliografických metadat
v centrálních indexech. Zaměřuje se především na rozdíly
v přístupu Souborného katalogu a dvou projektů Moravské
zemské knihovny (Číst Brno a Národní fonotéka).
Text je doplněn o autorovy postřehy z praxe a příklady
reálných problémů." />
<meta name="DC.Date" content="5.12.2014" />
<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.Language" content="cs" />

<meta name="DC.Source" scheme="URL"
content="http://duha.mzk.cz/clanky/virtualni-narodni-
fonoteka-jako-projekt-moravske-zemske-knihovny-v-brne
" />
<meta name="DC.Source" scheme="URL"
content="http://duha.mzk.cz/clanky/cist-brno-virtualni
-katalog-brnenskych-knihoven
" />
<meta name="DC.Source" scheme="URL"
content="http://users.ionio.gr/~sarantos/repository/j21J
-LibraryHiTech-Sitas.pdf
" />
<meta name="DC.Source" scheme="URL"
content="http://www.caslin.cz/spoluprace/importy/predimportni
-kontroly/
" />
<meta name="DC.Source" scheme="URL"
content="http://www.caslin.cz/spoluprace/importy/deduplikacni
-procedury
" />
<meta name="DC.Source" scheme="URL"
content="http://www.druva.com/blog/understanding-data-deduplication/
" />
<meta name="DC.Source" scheme="URL"
content="http://info.jib.cz/o-projektu/projekt/brana-projekt
" />
<meta name="DC.Source" scheme="URL"
content=" http://www.loc.gov/marc/holdings/
```

```
" />\\
<meta name="DC.Source" scheme="URL"
content="http://projekt.narodnifonoteka.cz/homeeng
" />
```