Deduplikace na portálu *Knihovny.cz*

Deduplication on *Knihovny.cz* portal

Mgr. Jana Kurfürstová, Mgr. Petra Žabičková, Bc. Tomáš Cejpek /
Moravská zemská knihovna v Brně (Moravian Library), Kounicova 65a, 601 87 Brno

Resumé
Cílem tohoto článku je nastínit zpracování metadat na portálu *Knihovny.cz*, a zejména pak přiblížit proces deduplikace, tj. rozpoznávání a shlukování záznamů vztahujících se ke stejnému dokumentu. Portál *Knihovny.cz* deduplikuje vytvářením sloučených záznamů propojujících lokální záznamy z jednotlivých metadatových zdrojů. Deduplikaci tvoří posloupnost kroků porovnávajících různé kombinace deduplikačních klíčů. Článek dále popisuje způsoby, kterými portál usiluje o minimalizaci komplikací způsobených chybnými metadaty, slabiny a silné stránky zvolených postupů a výsledky jejich aplikace nad záznamy ze všech zapojených zdrojů. Na závěr jsou diskutovány předpoklady pro zlepšení současného stavu a možnosti dalšího rozvoje.

Klíčová slova: *deduplikace bibliografických záznamů, MARC 21, portál Knihovny.cz*

Summary
The aim of this article is to outline the processing of metadata on *Knihovny.cz* portal, and in particular to present the process of deduplication, i.e. the recognition and clustering of records related to the same document *Knihovny.cz* portal deduplicates by creating merged records to link local records from individual metadata sources. Deduplication consists of a sequence of steps comparing different combinations of deduplication keys. The paper further describes the ways in which the portal aims to minimize complications caused by incorrect metadata, the weaknesses and strengths of the chosen methods and the results of their application on records from all involved sources. It concludes with a discussion of assumptions for improving the current state and possibilities for further development.

Keywords: *bibliographic records de-duplication, MARC 21, Knihovny.cz portal*

Úvod

Od podzimu 2016 slouží portál *Knihovny.cz* jako discovery systém k prohledávání záznamů ze zapojených metadatových i fulltextových zdrojů a jako zprostředkování služeb knihovních katalogů. Vývoj a správu portálu zajišťuje Moravská zemská knihovna v Brně. V době psaní tohoto článku je do portálu zapojeno 85 knihoven, 27 digitálních knihoven (13 z nich poskytuje portálu i plné texty) a 31 dalších bibliografií a databází (Zapojené knihovny a zdroje).

Portál *Knihovny.cz* uchovává záznamy ze zapojených zdrojů víceméně v téže podobě, v jaké byly sklizeny, a prezentuje jejich obsah, pokud možno analogicky k tomu, co mohou uživatelé vidět v jednotlivých knihovních katalozích. Uchování a zpřístupnění lokálních záznamů, jimiž byl dokument popsán v každém zapojeném zdroji, je jednou z hlavních charakteristik portálu. Aby bylo možné zprostředkování služeb knihovních katalogů portálem, *Knihovny.cz* musí akceptovat záznamy všech dokumentů, které zapojená knihovna zpřístupňuje svým čtenářům, bez ohledu na správnost, kvalitu a rozsah metadat. Zpracování záznamů, zejména pak deduplikaci, proto bylo třeba navrhovat dle reálné podoby metadat, nikoliv jen podle podoby vycházející z předepsaných katalogizačních pravidel a předpisů pro minimální záznamy (Minimální/Doporučené záznamy RDA/MARC 21).

Pro zachování dobré použitelnosti a přehlednosti vyhledávání není možné nakládat se záznamy jako s navzájem nesouvisejícími entitami. Portál proto pracuje s konceptem sloučených záznamů, které provazují jednotlivé lokální záznamy popisující stejný dokument na úrovni provedení. V optimálním případě se uživateli zobrazí konkrétní vydání hledaného dokumentu jako jedna položka ve výsledcích vyhledávání. V zobrazení úplného záznamu je pak uživateli umožněno přepínat mezi všemi záznamy, které popisují daný dokument.

Cílem tohoto článku je popsat zpracování metadat na portálu *Knihovny.cz*, a zejména pak přiblížit proces deduplikace, tj. způsobu, jakým dochází k rozpoznávání a shlukování záznamů vztahujících se ke stejnému dokumentu. V souvislosti s nástupem konsorciálních řešení při pořizování knihovních systémů pro více institucí zároveň lze říci, že problematika deduplikace je velmi aktuální a její řešení může těžit ze zkušeností s různými přístupy ke zpracování metadat, včetně přístupu zvoleného portálem *Knihovny.cz*.

Správce zdrojů

Sklizeň a normalizaci (sjednocování) záznamů ze zapojených zdrojů, deduplikaci záznamů popisujících stejné dokumenty a indexaci údajů do portálového indexu provádí nástroj Správce zdrojů neboli Record Manager 2 (dále RM2). Jde tedy o klíčovou komponentu portálu *Knihovny.cz*.

Tento software vyvíjí tým portálu *Knihovny.cz*, který se při jeho tvorbě volně inspiroval Record Managerem vytvořeným ve Finské národní knihovně pro portál *Finna.fi* (Record Manager). Finská verze Správce zdrojů byla Moravskou zemskou knihovnou v roce 2013 použita pro zpracování dat v portálu Virtuální národní fonotéka. Systém sdílené katalogizace je ve Finsku zaveden již řadu desetiletí a klade proto nižší nároky na deduplikaci než prostředí českého knihovnictví, které je specifické vysokým podílem lokálně tvořených záznamů (Ikäheimo 2017; Olkinoura-Tauru 2017).

Ukázalo se, že pro portál *Knihovny.cz* by finský Record manager nesplňoval nároky na dostatečnou robustnost, výkonnost a rozšiřitelnost, a to včetně volby programovacího jazyka a databázového systému. Problematické bylo také nastavení jednotlivých procesů

způsobem, který neočekával výskyt různých výjimečných stavů vyplývajících především ze substandardní kvality části zpracovávaných metadat. Nevyhovující byla i deduplikace, která byla řešena spíše monoliticky nežli formou nezávislých kroků kombinovatelných pro různé situace a nepočítala s možností výskytu interních multiplicit.

Ke splnění funkčních požadavků na *Knihovny.cz* proto bylo nezbytné vyvinout od základů vlastní, robustnější a flexibilnější backendový nástroj zpracovávající data přímo na míru portálu. Pro tento účel byl zvolen jazyk Java, který je pro vývoj komplexních systému tohoto typu vhodnější než PHP, v němž byl napsán původní Record Manager. Původní databázi MongoDB nahradil PostgreSQL. Došlo také k zavedení podpory běhu ve více vláknech a optimalizaci spouštění procesů pouze v případě potřeby zpracování relevantních změn v záznamech.

RM2 je modulární software umožňující nejen samostatné spouštění všech procesů, ale i jejich relativně snadné programátorské úpravy. Základy způsobu zpracování záznamů byly položeny již v přípravné fázi vývoje portálu *Knihovny.cz* Od té doby proces zpracování záznamů prošel a prochází průběžnými úpravami, rozšířeními a optimalizacemi reflektujícími zkušenosti se zapojovanými zdroji a provozem portálu. Snadnou modifikovatelností každého úkonu RM2 podporuje možnost experimentování a přizpůsobování se novým požadavkům nebo i potenciálním novým účelům využití.

Provoz RM2 vyžaduje podporu Java 8, nástroj na automatizaci buildů Apache Maven 3, databázi PostgreSQL ve verzi 9.5 či novější a na výstupu index Solr ve verzi 8 či novější. Hardwarové nároky závisí na objemech zpracovávaných dat. RM2 portálu *Knihovny.cz*, který pracuje s desítkami milionů záznamů, z nichž více než čtvrt milionu obsahuje fulltexty, je spouštěn až v osmi vláknech s využitím 1,5 GB RAM a cca 500 GB na SSD disku. Schéma komponent umožňujících provoz portálu *Knihovny.cz* včetně role RM2 v tomto systému znázorňuje obrázek 1.

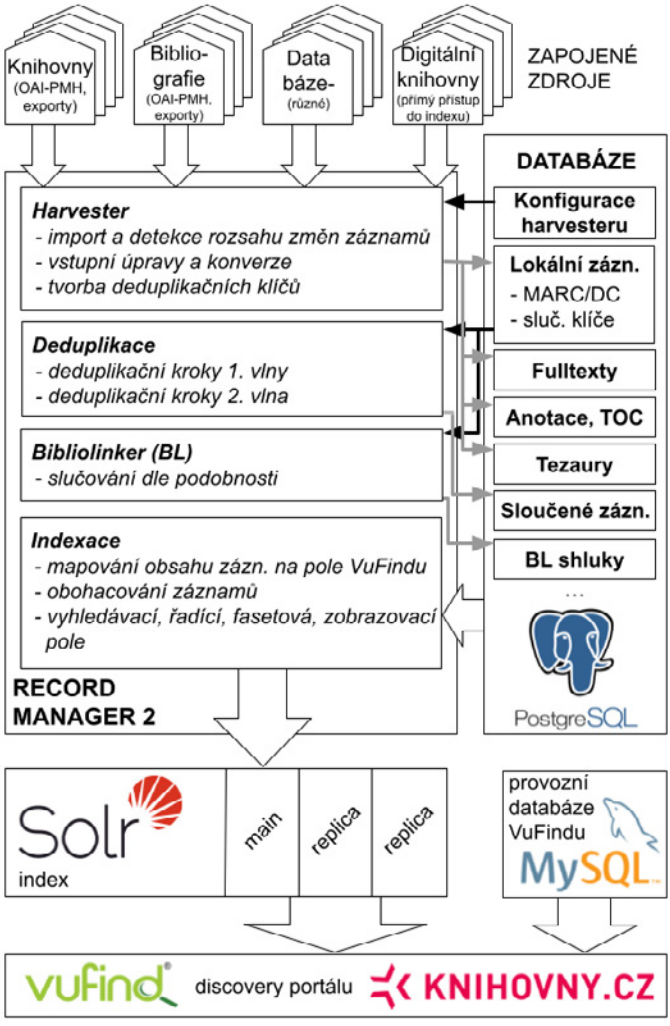
Sklízení a vstupní zpracování dat

Preferovaným způsobem získávání záznamů pro *Knihovny.cz* je sklizeň pomocí protokolu OAI-PMH. Pro případ, kdy správce zapojovaného zdroje nemá k dispozici OAI data provider, umí RM2 pracovat i s exporty nahrávanými na SFTP server. Prostřednictvím OAI-PMH lze nejsnáze zajistit nejen počáteční sklizeň nově zapojovaného zdroje, ale hlavně každodenní aktualizace přibyvších, upravených a smazaných záznamů, aby měli uživatelé portálu vždy aktuální informace o existenci a dostupnosti dokumentů.

Schopnost průběžných aktualizací je na portálu důležitá zejména kvůli exemplářovým údajům dodávaným v poli 996, k nimž jsou při prohlížení záznamu uživatelem načítány aktuální dostupnosti z API knihovních systémů. Nezřídka však dochází i k jiným změnám metadat, které je vhodné reflektovat na portálu, a to včetně nenulové pravděpodobnosti přepsání konkrétního záznamu záznamem pro jiný dokument při zachování systémového čísla, což vede k nutnosti takový záznam okamžitě znovu deduplikovat. Možnost pozměnění záznamů knihovnou v jakýkoliv okamžik je jedním z hlavních důvodů, proč tým portálu *Knihovny.cz* neprovádí žádné ruční úpravy metadat a nezasahuje ani do automatizovaných procesů deduplikace a indexace. Při aktualizaci záznamu manuálně upraveného na straně portálu by totiž nebylo možné automaticky rozhodnout, zda má být manuální úprava zachována či přepsána aktualizovanými metadaty.

Pro úsporu strojového času proto provádí RM2 na vstupu kontrolu aktualizovaných záznamů. Došlo-li ke změně v těle záznamu, kontrola rozhodne, zda stačí záznam pouze reindexovat nebo je nutné jej i znovu deduplikovat.

Záznamy jsou ukládány a dále zpracovávány v databázi PostgreSQL. Už během sklizně se v případě některých zapojených zdrojů provádí normalizace záznamů, při které



Obr. 1 Architektura portálu Knihovny.cz (Kurfürstová, Žabičková, Cejpek)

RM2 převádí specifická pole MARC na pole definovaná standardem (např. termíny z lokálních heslářů lze převést na volně tvořená klíčová slova v poli 653), generuje URL do pole 856 (např. odkazy do knihovních katalogů u Palmknih) apod.

Drtivou většinu sklízených a dále zpracovávaných metadat tvoří záznamy ve formátu MARC 21. RM2 obsahuje historicky také ovladač pro Dublin Core, před kterým je v případě zapojených digitálních knihoven preferován podrobnější MODS. Tento formát lze totiž transformovat do MARCu 21 a poté zpracovat včetně deduplikace tištěné a digitalizované formy dokumentu.

Data v proprietárních formátech a data z neknihovních zdrojů se pro potřeby sjednocení prezentace většinou konvertují do MARCu 21. Entity popisované těmito záznamy jsou ze své podstaty unikátní, a nemá tedy význam je následně deduplikovat (jde o metadata ze zdrojů jako Zákony pro lidi, Úřad průmyslového vlastnictví, Česká agentura pro

standardizaci, Osobnosti regionu, slovníková hesla, Adresář knihoven). Dlouhodobě je zvažováno také vyloučení exemplářově popsaných historických fondů z deduplikace, ale díky stávající formě jejich prezentace se zachováním lokální podoby záznamů k němu zatím nebyl důvod.

Deduplikační klíče

Z úkonů prováděných již při sklizni záznamů si podrobnější výklad zaslouží generování deduplikačních klíčů. Klíče se tvoří z údajů, jejichž kombinací lze identifikovat dokument. Datovým typem deduplikačních klíčů je řetězec (*varchar*) nebo celé číslo (*numeric*). Generování klíčů lze spouštět i samostatně nad již sklizenými záznamy, takže redefinice kteréhokoliv klíče s sebou nenese nutnost opětovné sklizně všech knihoven. V době psaní tohoto článku byl proces spuštěn nad 45 miliony záznamů a trval přibližně 3 dny.

U řetězových klíčů dochází k normalizaci, aby obsahovaly pouze alfanumerické znaky bez velkých písmen a diakritiky. Toto opatření snižuje výpočetní náročnost porovnávání klíčů a zároveň redukuje dopady rozdílných způsobů zápisu u téhož údaje.

Šanci na deduplikaci záznamů různě popsaných vícedílných dokumentů zvyšuje převod čísel 1–9 z arabských číslic na římské a vypouštění nevýznamových slov jako „svazek“, „díl“, „sešit“ apod. v názvových klíčích. Desetimístná ISBN se převádějí na třináctimístná, což usnadňuje jejich porovnávání a umožňuje použití číselného datového typu. Klíče pro ISBN, ISSN, ISMN a EAN jsou tvořeny pouze z údajů úspěšně zkontrolovaných validátorem.

Zvláštní typ klíčů tvoří identifikátory záznamů sloučených mimo portál *Knihovny.cz*. Jako spolehlivý klíč je použit identifikátor v poli 001 u záznamů vzniklých sdílenou katalogizací knihoven s celostátním povinným výtiskem Cluster. Od roku 2005 jsou vybrané záznamy Národní knihovny České republiky, Moravské zemské knihovny v Brně a Vědecké knihovny v Olomouci navzájem zrcadleny a liší se jen v lokálních údajích (Machačová 2005). Podobným způsobem RM2 zužitkovává údaje o záznamech deduplikovaných v Souborném katalogu ČR (Deduplikační procedury 2015), kdy jsou využívány původní identifikátory záznamů jednotlivých knihoven v podpoli 910 $w spolu s podobnostním porovnáváním názvových klíčů pro zamezení chyb vzniklých např. přeobsazením záznamů.

Následující tabulka na str. 53 naznačuje, jak jsou tvořeny deduplikační klíče ze záznamů v MARCu 21.

Deduplikační kroky

Deduplikace na portálu *Knihovny.cz* je proces vytváření sloučených záznamů sdružováním lokálních záznamů z jednotlivých metadatových zdrojů. Slučují se lokální záznamy popisující tentýž informační objekt na úrovni odpovídající entitě FRBR provedení, což je ve většině případů synonymum pro vydání. Protože jsou na této úrovni tvořeny i bibliografické záznamy, jde o nejschůdnější variantu jak pro zpracování, tak pro prezentaci dat na portálu. Ke slučování na vyšších úrovních (vyjádření, dílo) by bylo nutné reflektovat řadu komplikovaných otázek (vícedílné monografie a cykly, vydání s významně odlišným obsahem…), z nichž většina by za stávající kvality a rozsahu metadat ani nebyla řešitelná.

Sloučené záznamy jsou tvořeny pro všechny lokální záznamy včetně těch, ke kterým se vztahuje jen jeden lokální záznam. Vždy tedy vzniká hierarchie sloučeného záznamu („rodiče“) a jednoho či více lokálních záznamů („potomků“). Sloučené záznamy neobsahují v databázi žádná metadata kromě identifikátorů svých „potomků“ a času vytvoření

Tab. 1 Deduplikační klíče

Název klíče:	Tvořeno z:	Podmínky:
format	viz text	
title	245$abnp	
short_title	245$anp	existuje 245$b
anp_title	245$anp	jen knihy
isbn	020$a	validní hodnoty
issn	022$a	validní hodnoty
ismn	024$a	indikátor 1 == 2, validní hodnoty
ean	024$a	indikátor 1 == 3, validní hodnoty
cnb	015$a	prefix „cnb“ + číslo
oclc	035$a	
uuid	856$u, 911$u	911$u jen u SKC
cluster_id	001	jen u NKP, MZK, VKOL
publication_year	264$c > 260$c > 008|07-10	
author_string	100$a > 700$a	
author_auth_key	100$7 > 700$7	
pages	300$a (1. číslo)	
scale	255$a	jen mapy
language	041$a > 008|35-37	hodn. „cze“,„eng“ a „oth“ pro ostatní
source_info_t	773$t	
source_info_x	773$x > 773$z	
source_info_g	773$g	
publisher	264$b > 260$b	
publisher_number	028$a	
edition	250$a	jen knihy

shluku. Slučování záznamů na portálu lze označit jako virtuální deduplikaci v kontrastu s agregátory s klasickou deduplikací, kde se z většího počtu lokálních záznamů na vstupu stane jeden výsledný sloučený záznam na výstupu.

Deduplikace sestává z posloupnosti deduplikačních kroků. V každém deduplikačním kroku je porovnávána jiná kombinace deduplikačních klíčů. Výjimkou jsou kroky založené na porovnávání ID záznamů z Clusteru nebo namnožených záznamů institucí se sdílenou katalogizací (nyní Knihovna Akademie věd ČR, plánováno pro regionální systémy). Naopak použití nelokálních identifikátorů jako ISBN, ISSN nebo číslo ČNB coby jediného porovnávaného klíče se ukázalo jako vysoce nespolehlivé.

Správce zdrojů používá tři varianty porovnávání klíčů. Nejčastěji jsou klíče porovnávány na přesnou shodu. Někdy se však uplatňuje podmínka, kdy se za neshodu považují jen rozdíly mezi neprázdnými klíči, tj. výsledek porovnání prázdného klíče s neprázdným není považován za neshodu. Poslední způsob představuje porovnávání na podobnost.

Deduplikace probíhá ve dvou vlnách. První vlna zahrnuje deduplikační kroky porovnávající „silné“ kombinace klíčů. Za „silnou“ je považována taková kombinace klíčů, díky které lze sloučit záznamy s poměrně vysokou mírou spolehlivosti.

Do druhé vlny postupují pouze ty záznamy, které se nepodařilo sloučit v první vlně. Deduplikační kroky druhé vlny porovnávají „slabé“ kombinace klíčů s vyšší pravděpodobností sloučení záznamů, které k sobě nepatří. Shluky z první vlny jsou izolovány od shluků z druhé vlny. K zavedení druhé vlny kroků bylo přikročeno teprve nedávno s cílem zvýšit šanci na deduplikaci velmi stručných záznamů popisujících zejména starší publikace a zároveň nepoškodit shluky vytvořené spolehlivějším způsobem. Ve druhé vlně pochopitelně vznikají i nežádoucí shluky. Vzhledem k tomu, že kroky druhé vlny spojí jen něco přes 1,2 % lokálních záznamů, toto riziko bylo pro účely portálu vyhodnoceno jako přijatelné.

V rámci vlny se každý krok pokouší spojit záznamy, které dosud nebyly spojeny v předchozích krocích. Například: Mějme záznamy 1, 2, 3, 4, 5, 6, 7. Dále předpokládejme, že záznamy {1, 4} byly sloučeny v předchozím kroku do záznamu S1 a záznamy {2, 6, 7} byly sloučeny do S2. Provedením dalšího deduplikačního kroku bylo zjištěno, že záznamy {1, 3, 6} jsou si dostatečně podobné. Záznam 1 patří do S1, záznam 6 do S2. Stav systému po tomto kroku bude tedy: Menší shluk S2 byl začleněn do početnějšího shluku S1, který nyní obsahuje záznamy {1, 2, 3, 4, 6, 7}, zatímco záznam 5 zůstává sám v rodičovském záznamu S3. Porovnávány jsou vždy klíče lokálních záznamů, nikoliv celých shluků. Pokud právě provedený krok porovnával klíče A a B, záznam 5 zůstal nespojen, přestože měl hodnotu v klíči A shodnou se záznamem 1 a hodnotu v klíči B shodnou se záznamem 7.

Seznam deduplikačních kroků a klíčů, které se v nich porovnávají, je uveden v tabulce 2 v kapitole Vyhodnocení náročnosti deduplikace.

Z výše uvedeného popisu mechanismu deduplikace vyplývá, že sloučené záznamy nejsou statické a v případě aktualizací na straně knihovny může dojít k jejich dalšímu slučování nebo naopak rozpadu. Shluky tedy nemohou mít permanentní identifikátory a nemá smysl na ně tvořit jakékoliv odkazy.

Při deduplikaci se nebere v potaz původ záznamu, takže nic nebrání sloučení několika záznamů ze stejné knihovny, ať už popisují stejný dokument, obsahují chybu nebo se provázaly přes chybný záznam z jiné knihovny. Po technické stránce nezpůsobuje zpracování interních multiplicit v RM2 žádné potíže. Webové rozhraní portálu se s těmito případy vypořádává zobrazováním odkazů v náhledu úplného záznamu. Z uživatelského hlediska jsou interní multiplicity samozřejmě nežádoucí a knihovny jsou při zapojování do portálu vyzývány k jejich kontrole.

Indexace

Poslední fází zpracování záznamů RM2 je indexace, kdy se pole portálového indexu plní vybranými údaji ze záznamů uložených v databázi. Hierarchie lokálních a sloučených záznamů zůstává zachována. Většina údajů potřebných pro vyhledávání, filtrování či řazení se indexuje do sloučených záznamů. V „potomcích“ jsou obsažena především pole potřebná při zobrazování lokálních záznamů. Index je provozován na open source vyhledávací platformě Solr založené na knihovně Apache Lucene. Aktuální schéma indexu původně vycházelo z nástroje SolrMarc 2.x (solrmarc) a postupně bylo výrazně modifikováno pro potřeby portálu.

Silnou stránkou portálového indexu je vzájemné obohacování záznamů vedoucí k lepší dohledatelnosti záznamů uživatelem. Jak již bylo řečeno, většina údajů z lokálních záznamů se v indexu slévá do sloučeného záznamu. Sloučené záznamy se obohacují také o alternativní termíny jmenných i věcných autorit a plné texty obsahů poskytovaných Databází obálek knih *https://www.obalkyknih.cz*. Portál indexuje také plné texty z 13 digitálních knihoven a metadata z dalších 14. Uživatel tak může na portálu najít dokument

ze své knihovny pomocí vyhledávacího výrazu, který by měl v katalogu knihovny nulový počet výsledků (např. díky dostupnosti plného textu nebo podrobnějšího věcného popisu z propojených záznamů).

Úskalí deduplikace

Úspěšnost jakkoliv nastaveného deduplikačního algoritmu závisí na kvalitě zpracovávaných metadat, která je určena jejich syntaktickou a sémantickou správností (Understanding Metadata 2004). Kvalitní metadata odpovídají standardům použitého formátu a sdělují dostatek pravdivých, relevantních a strojově rozpoznatelných údajů o popisovaném dokumentu. Klíčovou vlastností kvalitních metadat je jejich konzistentnost, tj. jednotnost rozsahu zaznamenaných údajů a způsobu jejich zápisu ve všech záznamech téhož zdroje. Výstup deduplikace záznamů, které tyto podmínky nesplňují, bude vždy obsahovat chyby. „Ladění“ deduplikace je pak vždy otázkou kompromisu mezi vysokým podílem špatně sloučených záznamů různých dokumentů a velkým množstvím nededuplikovaných záznamů stejných dokumentů.

Většinu problémů, na které portál *Knihovny.cz* naráží při deduplikaci, se kryje s chybami způsobujícími problémy i v Souborném katalogu ČR, tj. nesprávně vyplněné názvové údaje (zejména u vícedílných dokumentů a knižních cyklů), ISBN a další klíčová pole (Svobodová 2015 a 2019). Potíže způsobují i chyby v ostatních klíčových údajích.

Z rozpisu deduplikačních kroků je patrné vzájemné oddělení různých typů dokumentů. Správné určení typu dokumentu je tedy zásadním momentem pro deduplikaci každého záznamu. Klíč „format“ nesoucí informaci o typu dokumentu (tj. kombinace údajů o typu obsahu a nosiče) je nejsložitěji tvořeným deduplikačním klíčem.

Za rozpoznáváním typů dokumentů stojí algoritmus, který kromě katalogizačních pravidel bere v potaz i širokou řadu často se vyskytujících odchylek. Kromě údajů z návěští, kontrolních polí a polí fyzického popisu nebo podpole 245 $h je dle okolností vyhodnocován i Konspekt, vydavatelské údaje (pro vydavatelství dokumentů pro zrakově postižené), žánr/forma nebo některá poznámková a propojovací pole. Výstup není ani zdaleka dokonalý. Je třeba se smířit s tím, že se nikdy nepodaří podchytit všechny výjimky a do procesu budou vždy vstupovat i záznamy, kde je deklarovaný typ dokumentu neúplný (např. blíže nespecifikovaná „zvuková deska“) nebo zcela chybný.

Jednu z nejproblémovějších oblastí katalogizace napříč knihovnami představují periodika, a to navzdory dostupnosti školení a srozumitelné dokumentaci. Nededuplikovatelnost periodik způsobuje nejčastěji nedostatek údajů pro vytvoření klíčů. Záznamy obsahující pouze hlavní název jsou zcela běžným jevem. Když už některé záznamy obsahují časový interval, často nejde o léta, kdy periodikum vycházelo, ale o ročníky dostupné v dané knihovně. Náprava stažením metadat ze Souborného katalogu patří vzhledem k relativně nízkému počtu periodik odebíraných v základních knihovnách k těm snáze proveditelným.

Dalším neduhem záznamů periodik bylo zejména v minulosti popisování svázaných ročníků monografickým způsobem. Monograficky se v mnoha knihovnách dodnes záměrně popisují komiksy. U nepravých a příležitostných periodik je dvojí způsob zápisu umožněn dokonce katalogizačními pravidly.

Deduplikace záznamů článků vykazuje obvykle nízkou spolehlivost, protože analytický popis nenabízí mnoho údajů vhodných pro tvorbu deduplikačních klíčů. Největším úskalím je nutnost pracovat s údaji o zdrojovém dokumentu s velmi nízkou granularitou. Pro informace o čísle, ročníku, datu vydání a rozsahu stran ve zdrojovém dokumentu se používá jediné podpole 773 $g. Variabilita způsobů zápisu do podpole 773 $g znemožňuje jakékoliv parsování (tj. rozpoznání syntaktické struktury) a normalizaci obsahu za

účelem tvorby snáze porovnatelných deduplikačních klíčů. Bez údajů z podpole 773 $g se nelze obejít, jinak by docházelo k chybnému slučování různých úvodníků a pravidelných rubrik vyskytujících se v každém čísle téhož periodika (podpole 773 $q nebo 773 $9 nejsou zastoupena v tolika záznamech, aby mělo význam jejich porovnávání). Deduplikaci analytických záznamů znesnadňuje také často chybějící ISSN nebo ISBN zdrojového dokumentu (u titulů, které je mají přiděleno) a roku vydání v poli 008 (i když dle obsahu 773 $g katalogizátor tento údaj znal).

Výše jmenované problémy jsou obvykle výsledkem nedůsledné katalogizace. Ne vždy jde ale o důsledek laxního přístupu k pravidlům. Obsahově chudé záznamy mohou pocházet i z retrokonverze, která nebyla provedena s knihou v ruce. Některé chyby vznikají automatizovanými procesy. Časté výskyty obsahu patřícího do podpole 245 $c v klíčových podpolích 245 $ab jsou pravděpodobně pozůstatkem konverze z UNIMARCu do MARCu 21, stejně jako vložení obsahu patřícího do podpole 245 $n na podpole 245 $h u složitěji hierarchizovaných vícedílných publikací. Některé opravy lze řešit pomocí úpravou šablon a konfigurací exportů/OAI setu. I zde jsou možnosti omezeny schopnostmi daného knihovního systému.

Samostatné téma představuje problematika vzniku nežádoucích shluků. Výskyt záznamů různých dokumentů sloučených kvůli chybě v klíčových údajích má na použitelnost portálu závažnější dopad než interní multiplicity způsobené vícenásobným popisem téhož dokumentu v jedné knihovně. Nežádoucí shluky vznikají především na klíčích tvořených z ISBN a čísel České národní bibliografie, a to ať jde o totožné hodnoty u „zdola“ popsaných vícesvazkových děl, údaje náležející jinému vydání téhož díla, identifikátory tištěné knihy uvedené u audioknihy, nebo o zpětně neopravená ISBN, o nichž se v době katalogizace nevědělo, že jsou chybná. Příčinou vzniku nesprávných shluků ale mohou být i nedostatky v ostatních klíčových údajích. Podrobnějším rozborem této rozsáhlé problematiky se může zabývat další článek. Při opravách musí mít maximální prioritu odstranění takových chyb, kvůli kterým dochází k propojení nejen vadných záznamů, ale i na ně navázaných správných záznamů.

Vyhodnocení náročnosti deduplikace

Abychom zjistili aktuální časovou náročnost deduplikace a množství záznamů sloučených v jednotlivých krocích proběhlo pro účely tohoto článku spuštění celého procesu nad všemi 45 664 320 záznamy. Vygenerování deduplikačních klíčů trvalo 3 dny. Doba trvání samotného slučování záznamů je uvedena v následující tabulce, která znázorňuje aktuální posloupnost deduplikačních kroků a porovnávaných klíčů. Klíče uvedené tučně jsou porovnávány na shodu, normální řez písma byl použit pro zákaz rozdílu v neprázdných klíčích a kurzíva označuje porovnávání na podobnost:

Tab. 2 Deduplikační kroky

Název kroku:	Porovnávané klíče:	Podmínky:	Sloučeno:	Čas:
		Vstup:	**45 664 320**	
1. vlna				
ClusterIds	**cluster_id**	jen Cluster	834 988	0:23:47
SimpleKeysCaslinRest	sloučení v SKC, title		14 034 897	7:44:50
SimpleKeysIsbn	**isbn, title, publication_ year, format**	mimo periodika	4 489 782	4:38:58
SimpleKeysCnb	**cnb, title, publication_ year, format**	mimo periodika	1 552 202	1:25:24
TitleAuth	**title/short_ title, author_ auth_key,** isbn, cnb, pages	názvové klíče > 16 znaků	2 738 176	13:09:07
SimpleKeysEan	**ean, title, publication_ year, format**	jen AV dokumenty	24 651	0:20:42
SimpleKeysBlindAudio	**title, author_ auth_key, publication_ year**	jen zvukové knihy pro nevidomé	44 244	0:15:19
SimpleKeysPub-lisherNumber	**title, publisher_ number, publication_ year, format**		91 817	0:44:52
CnbClusters	**cnb,** title, publication_year	mimo periodika	41 125	1:17:57
OclcClustersTable	**oclc**, title, publication_year	mimo periodika	106 924	0:28:09
IsmnClustersTable	**ismn**, title, publication_year	mimo periodika	2 080	0:08:25
PeriodicalsIssn	**issn, title**	jen periodika	78 933	0:43:27
PeriodicalsCnb	**cnb, title**	jen periodika	37 429	0:30:00
UuidClusters	**uuid,** title		63 687	0:59:52
PeriodicalsCnbClusters	**issn,** title	jen periodika	1 857	0:26:20
PeriodicalsIssnClusters	**cnb,** title	jen periodika	7 440	0:34:09
PeriodicalsOclcClusters	**oclc,** title	jen periodika	0	0:13:50
SimilaritesResults	**publication_ year, language,** author_string, isbn, cnb, pages, title		0	1:56:01

Název kroku:	Porovnávané klíče:	Podmínky:	Sloučeno:	Čas:
PeriodicalsSimilaritesResults	title, **publication_year**	periodika od r. 1850	41 956	0:45:19
ArticlesXG	**title, author_string, publication_year, source_info_x, source_info_g**		1 187 318	1:06:42
ArticlesTG	**title, author_string, publication_year, source_info_t, source_info_g,** source_info_x		95 742	0:38:01
2. vlna				
SimpleKeysDisadvantagedPublisher	**anp_title, author_string, publication_year, publisher**		498 049	0:28:49
SimpleKeysDisadvantagedEdition	**anp_title, author_string, publication_year, edition**		99 354	0:16:05
SimpleKeysDisadvantagedPages	**anp_title, author_string, publication_year, pages**		47 541	0:13:38
SimpleKeysDisadvantagedIsbn	**anp_title, isbn, publication_year, pages**		9 126	0:14:02
SimpleKeysDisadvantagedCnbPages	**cnb, pages**	mimo hudebniny	23 220	0:13:14
SimpleKeysDisadvantagedCnbTitle	**cnb, title**		385	0:12:57
SimpleKeysDisadvantagedIsmn	**ismn, publication_year, pages**		1 425	0:06:22
		Celkem:	**26 154 348**	**40:25:07**

Ne všechna podobnostní porovnávání jsou stejná. U názvových klíčů se vyhodnocuje Levenshteinova vzdálenost (Levenshtein 1966), a to pouze tehdy, neobsahují-li vyloučené výrazy, které by při slučování na základě podobnosti způsobovaly vznik nežádoucích shluků. Některé kroky počítají i s podobností založenou na tom, zda je jeden klíč prefixem druhého. Porovnávání klíče pro počet stran akceptuje buď procentuální či absolutní rozdíly (dle rozsahu dokumentu) čísel nebo příslušnost do určitého intervalu.

První pohled na tabulku může budit dojem, že některé kroky nevyhodnocují nic jiného, než je vyhodnocováno v krocích, které se provádějí později. Je však třeba vzít v úvahu, že záznamy sloučené v jednom kroku již nejsou porovnávány v následujících krocích. Posloupnost kroků vychází ze snahy, aby byl celý proces co nejrychlejší a výpočetně náročné kroky probíhaly nad co nejmenším množství záznamů. Nejnáročnější kroky, které pracují s Levenshteinovou vzdáleností a zároveň velkým množstvím klíčů, jsou proto zařazeny do závěrečné fáze první vlny. Ze sledování vzešlo zajímavé zjištění, že dříve užitečné kroky, na nichž se před lety spojovalo významné množství záznamů, nyní nevytváří po několika letech postupných úprav klíčů a předcházejících kroků žádné shluky, a lze tedy uvažovat o možnosti jejich vypuštění.

Při sledování výkonnosti deduplikace došlo i jejímu spuštění bez kroku využívajícího informace o sloučení v Souborném katalogu ČR. Na výstupu bylo deduplikováno 56,43 % záznamů oproti 57,34 % záznamům sloučených běžným způsobem (rozdíl činí 413 320 záznamů). S využitím Souborného katalogu vzniklo 6 458 457 shluků, zatímco bez něj 6 548 612. Podíl shluků, které obsahují interní multiplicitu (tj. několik lokálních záznamů z téže knihovny), se vždy pohyboval kolem 2,8 %. Rozdíl v časové náročnosti obou způsobů deduplikace byl zanedbatelný. Na základě tohoto experimentu lze odhadnout, že deduplikace na portálu *Knihovny.cz* je použitelná bez využití externích zdrojů. Podobnost výsledků nenapovídá, že by docházelo k významnému navýšení množství nežádoucích shluků nebo naopak ke drolení správně sloučených záznamů. Možnost využití spolehlivé deduplikace Souborného katalogu je však pro portál *Knihovny.cz* nespornou výhodou.

Shrnutí

Předcházející kapitola shrnuje měření provedená při jednorázové deduplikaci záznamů ze všech zapojených bibliografických zdrojů, tj. při úkonu, který by byl v reálném provozu nutný pouze v případě ztráty dat při nějaké mimořádné události.

Deduplikační algoritmus RM2 byl navržen způsobem, který vycházel z metod uplatněných v Souborném katalogu ČR a zkušeností z předcházejících projektů řešených původním jádrem týmu *Knihovny.cz* a dále byl upravován na základě analýz kvality zpracovávaných metadat. Průběžně prováděné rozbory se zaměřují převážně na kontrolu výstupů RM2 nad záznamy z nově zapojovaných zdrojů. Cílem těchto zkoumání je doporučit knihovnám opravu nejkritičtějších chyb spíše než vyhodnocování efektivity deduplikačního procesu, nkromě toho jde také o cenný zdroj poznatků vedoucích k postupnému vylepšování deduplikace i indexace.

Jakékoliv úpravě deduplikačních klíčů či kroků předchází testování a vyhodnocení dopadů na použitelnost portálu. Navrhované změny jsou aplikovány pouze když přinášejí zlepšení a v důsledku jejich využití se nevyprodukuje nadměrné množství nežádoucích shluků nebo naopak bezdůvodně osamocených záznamů. Kromě toho dochází pravidelně k analýzám atypicky početných (tj. pravděpodobně nežádoucích) shluků za účelem odhalování a omezování příčin jejich vzniku.

Jak již bylo řečeno, portál *Knihovny.cz* zpracovává všechny záznamy ze zapojovaných zdrojů. Řada knihoven přistupujících do portálu nepřispívá do Souborného katalogu ČR ani z něj nepřebírá mnoho záznamů. Podíl nepříliš kvalitních lokálně tvořených záznamů se sice v posledních letech výrazně snížil, ale podstatná část fondů byla katalogizována v dobách, kdy knihovny nepočítaly s možností využití svých záznamů mimo svůj vlastní katalog a kvalita metadat pro ně nebyla prioritou. Zpracování záznamů komplikují také různé zvyklosti záměrného porušování katalogizačních pravidel ospravedlňované snahou vyjít vstříc čtenáři. Kdyby byly tyto zvyklosti jednotné a napříč knihovnami by se vyskytovaly záznamy, které jsou chybné stejným způsobem, nebyl by to z hlediska deduplikace takový problém. Porušování pravidel se ale bohužel samo neřídí žádnými pravidly.

Stav popisu neknižních dokumentů v knihovnách je také tristní. Lze předpokládat, že zčásti jde o pozůstatek z dob, kdy se knihovny snažily vypořádat s knihovními systémy, kterým chyběly samostatně prodávané moduly pro jednotlivé typy speciálních dokumentů, a zčásti o důsledek méně častých školení katalogizace těchto typů dokumentů.

Zajímavým předmětem bádání by mohla být otázka, kolikrát je v českých knihovnách průměrně popisována jedna publikace a nakolik by byla situace odlišná, kdyby centrální katalogizace v tuzemském knihovnictví nefungovala jen na bázi dobrovolného přebírání záznamů ze Souborného katalogu ČR. V případě existence jediného kanonického zdroje záznamů pro všechny knihovny, který by byl budován specialisty na katalogizaci, by došlo nejen ke zvýšení kvality metadat, ale bezesporu také ke značné úspoře ruční práce a prostředků vynakládaných na plnohodnotné katalogizační moduly knihovních systémů (postačila by podpora doplňování lokálních polí dle individuálních potřeb knihovny).

Portál *Knihovny.cz* vznikl s cílem zprostředkovat jednotný přístup ke službám maxima českých knihoven, takže politika nabírání nových členů je velmi vstřícná. Vyloženě odmítnuto bylo zatím jen několik málo institucí, kde by opravy znamenaly de facto znovuvybudování celého katalogu. Nově zapojovaným knihovnám jsou posílány výstupy analýzy jejich záznamů spolu s upozorněním na obzvláště problematické nedostatky a výzvou, aby zlepšováním kvality svých metadat pomáhaly udržovat použitelnost portálového indexu.

Očekávat od knihoven s omezenými prostředky opravy velkých objemů dat vzniklých v uplynulých dekádách je nereálné. Bez vyhlídky na zásadní zlepšení tohoto stavu lze diskutovat o kvalitě vytvořeného indexu, a tedy i služeb portálu. Portál *Knihovny.cz* však i přes své nedostatky nabízí uživatelům užitečné funkce, které by v případě odstranění pravidlům nevyhovujících záznamů pozbyly smyslu. Právě úplnost indexu a vysoká míra obohacení dat dovede uživatele ke službám, které by mu jeho domovská knihovna nabídnout nemohla.

Závěr

Deduplikace na portálu *Knihovny.cz* prošla od roku 2015 velkým vývojem, který reflektoval především různorodost zpracovávaných zdrojů. Budoucí rozvoj se kromě dalšího vylepšování může ubírat například směrem k experimentům ve slučování záznamů na jiných úrovních, než je provedení. RM2 se pro podobné účely osvědčil při implementaci modulu Bibliolinker, který na portálu doporučuje záznamy na základě podobnosti díky prohloubení hierarchie shlukování.

Nástroj *Správce zdrojů pro sklízení, deduplikaci a indexaci dat* (dále Správce zdrojů) byl vyvinut na míru potřebám a specifikům portálu *Knihovny.cz*. Podstatu portálové deduplikace lze rozporovat, protože místo skutečné redukce většího počtu záznamů do jednoho sloučeného pracuje s virtuálními shluky záznamů. Výhodou tohoto mechanismu,

založeného na bezeztrátovém a opakovatelném slučování záznamů, je omezení dopadů chybné deduplikace.

Veškerý software umožňující fungování portálu *Knihovny.cz* je open source. Správce zdrojů by tedy bylo možné použít i v jiných projektech, ačkoliv by to znamenalo uzpůsobení procesů pro aplikaci v jiném kontextu. Výhodou Správce zdrojů je poměrně velká robustnost, rychlost a flexibilita. Nevýhodou pak to, že jde o poměrně komplexní nástroj, jehož využití dává smysl pouze v projektech většího rozsahu.

Z výše popisovaných skutečností je zřejmé, že portál bude vždy narážet na limity dané stavem vstupních dat, jejichž kvalita silně kolísá napříč zapojenými zdroji. Jako velmi cenné je proto vnímáno působení Souborného katalogu ČR, který dlouhodobě zvyšuje kulturu katalogizace v ČR, ať už jde o poskytování kvalitních a deduplikovaných záznamů nebo vzdělávací činnost v knihovnách.

Studie vznikla v rámci Dlouhodobé koncepce pro rozvoj výzkumné organizace – Moravská zemská knihovna v Brně.

Seznam použitých zkratek

API	Application Programming Interface (aplikační programové rozhraní)
ČČNB	číslo České národní bibliografie
EAN	European Article Nubmer
FRBR	Functional Requirements for Bibliographic Records
ISBN	International Standard Book Number
ISMN	International Standard Music Number
ISSN	International Standard Serial Number
MARC	Machine Readable Cataloging
MODS	Metadata Object Description Schema
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
RDA	Resource Description and Access
RM2	Record Manager 2 (tj. správce zdrojů)
SFTP	Secure File Transfer Protocol

Literatura

Deduplikační procedury. *Souborný katalog České republiky* [online]. Národní knihovna ČR, 2015 [cit. 2022-12-13]. Dostupné z: *https://www.caslin.cz/caslin/spoluprace/jak-prispivat-do-sk-cr/dodavani-dat/jak-probiha-davkovy-import/deduplikacni-procedury.*

IKÄHEIMO, Ulla. Metadata creation in the national metadata repository Melinda. *Doria.fi* [online]. The National Library of Finland [cit. 2022-12-13]. Dostupné z: *https://urn.fi/URN:NBN:fi-fe201709128588.*

LEVENSHTEIN, Vladimir I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics – Doklady* [online]. 1966, **10**(8), 707–710 [cit. 2022-12-13]. Dostupné z: *https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf.*

MACHAČOVÁ, Libuše. Katalogizace – tentokrát opravdu sdílená: využití systému Aleph-Cluster. *Knihovny současnosti 2005* [online]. Sdružení knihoven ČR, 2005, 268–274 [cit. 2022-12-08]. ISBN 80-86249-30-1. Dostupné z: *https://ipk.nkp.cz/docs/knihovny-soucasnosti/knihovny-soucasnosti-2005/at_download/file.*

Minimální/Doporučené záznamy RDA/MARC 21. *Národní knihovna České republiky* [online]. Národní knihovna ČR [cit. 2022-12-12]. Dostupné z: *https://text.nkp.cz/o-knihovne/odborne-cinnosti/zpracovani-fondu/katalogizacni-politika/minimalni-doporucene-zaznamy-rda-marc-21.*

OLKINOURA-TAURU, Minna. National metadata repository Melinda. *Doria.fi* [online]. The National Library of Finland, 2017 [cit. 2022-12-13]. Dostupné z: *https://urn.fi/URN:NBN:fi-fe201709128589.*

Record Manager. In: *GitHub* [online]. The National Library of Finland [cit. 2022-12-08]. Dostupné z: *https://github.com/NatLibFi/RecordManager.*

Record Manager 2. In: *GitHub* [online]. Moravská zemská knihovna v Brně [cit. 2022-12-08]. Dostupné z: *https://github.com/moravianlibrary/RecordManager2.*

SolrMarc. In: *GitHub* [online]. [cit. 2022-12-08]. Dostupné z: *https://github.com/solrmarc/solrmarc.*

SVOBODOVÁ, Eva. Důvody vzniku duplicit v SK ČR. In: *Souborný katalog České republiky* [online]. Národní knihovna ČR, 2015 [cit. 2022-12-13]. Dostupné z: *https://www.caslin.cz/caslin/dokumenty/rok-2015/duvodyvznikuduplicitvskcr1.ppt.*

SVOBODOVÁ, Eva. Spolupráce se Souborným Katalogem ČR. In: *Souborný katalog České republiky* [online]. Národní knihovna ČR, 2019 [cit. 2022-12-13]. Dostupné z: *https://www.caslin.cz/caslin/dokumenty/rok-2019/rda-spoluprace-se-sk-cr.*

Understanding Metadata [online], 2004. Bethesda: NISO [cit. 2022-12-07]. ISBN 1-880124-62-9. Dostupné z: *http://hdl.handle.net/10150/105486.*

Zapojené knihovny a zdroje. *Knihovny.cz* [online]. Moravská zemská knihovna v Brně [cit. 2022-12-13]. Dostupné z: *https://www.knihovny.cz/Content/zapojene-knihovny-a-zdroje*.

KURFÜRSTOVÁ, Jana, Petra ŽABIČKOVÁ, Tomáš CEJPEK. Deduplikace na portálu *Knihovny.cz* *Knihovna: knihovnická revue*. 2023, **34**(1), 48–62. ISSN 1802-3250.