INTERNÍ MULTIPLICITY NA PORTÁLU KNIHOVNY.CZ

Mgr. Jana Kurfürstová; Jana.Kurfurstova@mzk.cz; (Moravská zemská knihovna v Brně)
Mgr. Petra Žabičková; Petra.Zabickova@mzk.cz; (Moravská zemská knihovna v Brně)
Bc. Tomáš Cejpek; Tomas.Cejpek@mzk.cz; (Moravská zemská knihovna v Brně)

Účel — Tento článek se věnuje problematice interních multiplicit bibliografických záznamů na portálu Knihovny.cz, protože se jedná o nežádoucí jev zhoršující dohledatelnost dotčených dokumentů. Za interní multiplicitu považujeme jev, kdy deduplikační algoritmus sloučí dva či více bibliografických záznamů z jednoho zdroje na základě rozhodnutí, že popisují stejný dokument (ať už tomu tak je, či nikoliv).

Design/Přístup/Metody — Poznatky shrnuté v tomto článku vycházejí ze zkušeností nabytých za léta provozu portálu Knihovny.cz.

Výsledky — Článek rozebírá různé příčiny vzniku interních multiplicit jak na straně knihoven, tak na straně portálu Knihovny.cz, popisuje jejich dopad na použitelnost portálu a zabývá se možnými způsoby jejich rozpoznávání, včetně výhledového využití umělé inteligence.

Originalita/Hodnota — Tým portálu Knihovny.cz by tímto rád přispěl svými poznatky ku prospěchu dalších projektů potýkajících se s touto problematikou.

http://doi.org/10.52036/1335793X.2024.2.70-77

ÚVOD

Od roku 2016 provozuje Moravská zemská knihovna v Brně (MZK) portál Knihovny.cz umožňující uživatelům prohledávat záznamy ze zapojených metadatových i fulltextových zdrojů (Knihovny.cz: Zapojené knihovny a zdroje) a spravovat svá čtenářská konta bez nutnosti procházení jednotlivých databází a katalogů knihoven.

Pro zachování přehlednosti výsledků vyhledávání používá portál Knihovny.cz deduplikační algoritmus rozpoznávající záznamy popisující stejný dokument. Uživatel pak vidí každé vydání hledaného dokumentu jako jednu položku výsledků vyhledávání. Podobně činí i ostatní báze agregující větší množství zdrojů bibliografických metadat, ať už jde o Souborný katalog ČR, nebo různé discovery systémy. Pravidla pro deduplikaci a způsob nakládání se zdeduplikovanými záznamy se v různých vyhledávačích liší a vzhledem rozdílným účelům i politikám těchto nástrojů je nedává smysl navzájem porovnávat.

Předmětem kritiky agregátorů bibliografických záznamů z více zdrojů bývají především viditelné multiplicity, kdy záznamy popisující stejný dokument nejsou rozpoznány a zobrazují se jako různé položky výsledků vyhledávání. Stejně, ne-li více závažným problémem jsou ale také chybné interní multiplicity, kdy se navzájem propojí záznamy týkající se různých dokumentů. Na rozdíl od viditelných multiplicit zde totiž dochází ke skrytí, nebo dokonce i ztrátě informace.

Obecně platí, že striktní deduplikační pravidla vyžadující vysokou shodu mezi záznamy omezí slučování záznamů, které k sobě nepatří. Zároveň ale nezdeduplikují větší množství neúplných, chybových či nekvalitních záznamů popisujících stejný dokument. Benevolentní pravidla pak dávají šanci na sloučení stručným či chybným záznamům, ale hrozí u nich častější vznik shluků propojujících záznamy různých dokumentů. Ojediněle pak může dojít ke sloučení i v případech, kdy ani jeden z těchto záznamů neobsahoval jakoukoliv chybu.

K plnohodnotnému zprostředkování služeb knihovních katalogů musí Knihovny.cz přijímat všechny záznamy z knihoven, nehledě na jejich různorodou kvalitu. Deduplikační algoritmus byl proto navrhován s ohledem na reálnou podobu metadat, která mnohdy nesplňují katalogizační pravidla a nemají ani rozsah předepsaný pro minimální záznamy. Podmínky pro sloučení záznamů jsou zde tedy méně přísné než v Souborném katalogu ČR a častěji zde dochází k vytváření shluků k sobě nepatřících záznamů, kterými se dále zabývá tento článek.

Dalším tématem článku jsou interní multiplicity popisující stejné dokumenty. Deduplikační algoritmy předpokládají, že každý zdroj pošle jeden záznam o každém dokumentu ve svém fondu či bázi. Prakticky žádná knihovna či bibliografie se ale nevyhnula tomu, aby během své existence nevytvořila pro některé dokumenty více záznamů. Báze indexující záznamy z mnoha zdrojů se s touto skutečností vypořádávají různými způsoby. I když pro portál Knihovny.cz tento typ interních multiplicit nepředstavuje problém, knihovny a bibliografie by měly usilovat o minimalizaci těchto případů, ať už z důvodů zapojení do dalších projektů, nebo zlepšení přehlednosti vlastního online katalogu.

DEDUPLIKACE NA PORTÁLU KNIHOVNY.CZ

Pro sklizeň, normalizaci, deduplikaci a indexaci záznamů ze zdrojů zapojených do portálu Knihovny.cz byl na půdě MZK vyvinut softwarový nástroj Správce zdrojů (*Record Manager 2*).

Portál Knihovny.cz ukládá záznamy získané z různých zdrojů v nezměněné formě a snaží se je prezentovat co nejpodobněji tomu, jak jsou zobrazeny v původních knihovních katalozích. Deduplikace na portálu Knihovny.cz proto funguje na principu hierarchie sloučeného záznamu (rodiče) a lokálních záznamů z jednotlivých knihoven či bibliografií (potomků). Sloučený záznam hraje při deduplikaci pouze roli propojovatele lokálních záznamů a sám nenese žádnou další informaci.

Před zahájením deduplikace se vytvářejí deduplikační klíče z polí lokálních záznamů obsahujících názvové údaje, jméno autora, rok vydání a další zásadní informace, podle kterých lze rozpoznat konkrétní provedení dokumentu. Správce zdrojů tvoří více než dvacet různých typů deduplikačních klíčů, i když některé z nich jsou určeny pouze pro určité typy dokumentů.

Deduplikace je souslednost deduplikačních kroků, při kterých se porovnávají různé kombinace deduplikačních klíčů. V případě shody se porovnávané lokální záznamy stanou potomky stejného sloučeného záznamu. Deduplikační kroky porovnávající pouze jediný údaj jako ISBN, ISSN nebo číslo ČNB se ukázaly jako značně nespolehlivé, a proto jsou i tyto identifikátory vždy vyhodnocovány v kombinaci s dalšími údaji.

Jak již bylo řečeno, některá z deduplikačních pravidel na portálu Knihovny.cz jsou značně benevolentní a mají tedy větší potenciál slučovat záznamy, které k sobě nepatří. Pro potlačení tohoto negativního jevu probíhá deduplikace ve dvou vlnách.

První vlna slučuje záznamy pomocí přísnějších deduplikačních kroků, kde je pravděpodobnost vzniku nežádoucích shluků velmi nízká. Aby nebyly relativně spolehlivé výstupy první vlny negativně ovlivněny následujícími benevolentnějšími deduplikačními kroky, tak je deduplikace lokálních záznamů sloučených v první vlně prohlášena na uzavřenou.

Do druhé vlny deduplikace s volnějšími pravidly postupují pouze záznamy, které se nepodařilo sloučit v první vlně, a je jim umožněno slučovat se pouze mezi sebou. Vzhledem k tomu, že ve druhé vlně vzniká jen okolo 1,2 % shluků, tak ani množství případně chybně sloučených záznamů není natolik zásadní, aby bylo důvodem pro zrušení celé druhé vlny.

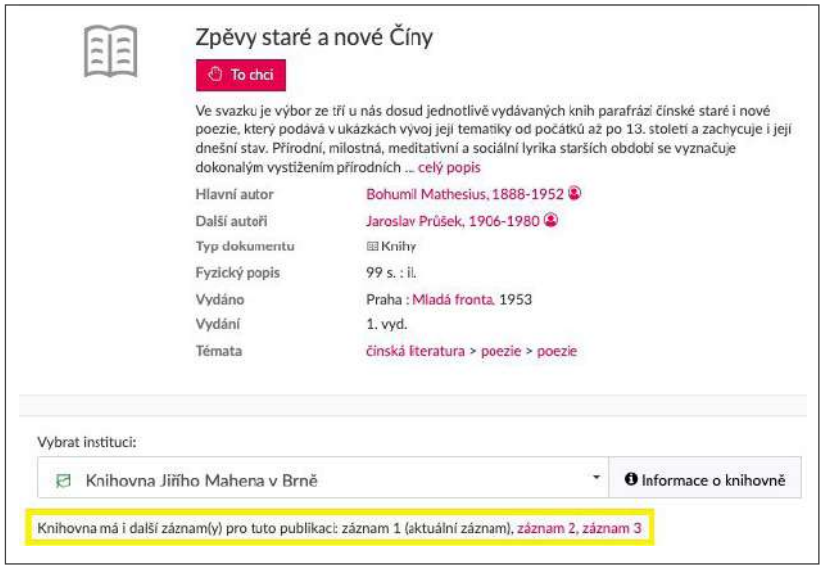
Protože Správce zdrojů sklízí nové a upravené záznamy ze zapojených zdrojů každý den, je nutné spouštět každý den i deduplikaci, aby brala v potaz nově nabytá data. Z toho plyne, že sloučené záznamy nejsou permanentní a aktualizací lokálních záznamů se mohou spojovat do větších shluků, nebo naopak rozpadat na menší shluky.

Pro nejlepší porozumění následujícího textu doporučujeme obeznámit se obsahem článku Deduplikace na portálu Knihovny.cz (Kurfürstová 2023), kde je podrobně popsán celý proces včetně způsobu tvorby deduplikačních klíčů a seznamu všech deduplikačních kroků.

ŘEŠENÍ INTERNÍCH MULTIPLICIT NA PORTÁLU KNIHOVNY.CZ

Deduplikace na portálu Knihovny.cz nebere v potaz, z jaké knihovny či bibliografie pochází každý lokální záznam. To umožňuje vznik interních multiplicit, kterými se dále zabývá tento článek. Na portálu jsou v úplném zobrazení záznamu tyto případy ošetřeny upozorněním s odkazy na ostatní propojené záznamy z téhož zdroje (viz obr. 1).

Zakázat slučování záznamů v rámci jednoho zdroje by znamenalo povolit sloučení s jakýmkoliv záznamem pouze jednomu záznamu z n-tice tvořící interní multiplicitu. Ostatních n – 1 multiplicitních záznamů by muselo zůstat vyloučených z deduplikace, a to i z deduplikace se záznamy z jiných zdrojů (protože tyto záznamy z jiných zdrojů by už byly sloučeny s jedním vybraným záznamem z n-tice). Toto opatření by navíc vyžadovalo provádět fiktivní předběžné kolo deduplikace, které by sloužilo pouze k identifikaci interních multiplicit před samotnou deduplikací obsahující zákaz slučování záznamů z jednoho zdroje. Z hlediska logiky portálu by tedy takové řešení nedávalo smysl.



Obr. 1 Odkazy na záznamy interních multiplicit na portálu Knihovny.cz

Stejně tak by nebylo vhodným řešením ponechávat na portálu jen jeden záznam z interní multiplicity a ostatní zahazovat. Jednak by došlo ke ztrátě informací o exemplářích, na které jsou navázány knihovní služby (což by bylo v rozporu s cílem portálu zastat funkce jednotlivých katalogů) a ztrácely by se také bibliografické informace v případech sloučení záznamů různých dokumentů.

Jiné souborné katalogy a discovery systémy však mohou mít dobrý důvod, proč se bránit vzniku interních multiplicit. Výsledkem deduplikace v Souborném katalogu je jeden bibliografický záznam vzniklý na základě záznamu z knihovny, která daný dokument popsala nejkvalitněji. Pro každou knihovnu, jejíž záznam se zdeduplikoval s tímto záznamem, je zde vytvořeno právě jedno pole 910 s odkazem na její původní záznam v lokálním katalogu. Pokud je cílem zobrazit seznam knihoven, ve kterých je dostupný daný dokument, přítomnost více polí 910 pro jednu knihovnu by v tomto kontextu nedávala smysl.

TYPY INTERNÍCH MULTIPLICIT

Přibližně 2,8 % sloučených záznamů na portálu Knihovny.cz obsahuje více než jeden záznam z téže knihovny. Nejhojněji jsou mezi nimi zastoupeny skutečné multiplicity, tzn. záznamy, které se sloučily správně, protože popisují stejný dokument. Případy, kdy bylo sloučení záznamů nežádoucí či sporné, jsou sice ve výrazné menšině, ale vzhledem k nepříjemnostem, které způsobují, je záhodno věnovat jim zvýšenou pozornost.

V závěru každé z následujících podkapitol je uveden podíl každého typu sloučení záznamů z jedné instituce. Tento údaj se nevztahuje k n-ticím, tzn. multiplicitám, ale ke dvojicím, čili duplicitám. Mezi více než dvěma záznamy totiž může (ale nemusí) vzniknout více typů vztahů (ve trojici může být mezi záznamy AB vztah jednoho typu a mezi záznamy AC a BC vztah jiného typu). Při klasifikaci typů duplicit je proto nutné rozložit n-tice na $\binom{n}{2}$ dvojic.

SKUTEČNÉ MULTIPLICITY

Nejjednodušším případem interních multiplicit jsou skutečné multiplicity, kdy se v knihovně vyskytuje více záznamů pro zcela stejný dokument. Příčina této situace leží často v minulosti, kdy se do jedné instituce sloučily dvě či více knihoven s průniky ve svých fondech.

Dalším historickým důvodem může být dávné používání knihovního systému nepodporujícího možnost přiřazení více knihovních jednotek jednomu bibliografickému záznamu. Lze předpokládat, že se při migraci na novější software nemuselo podařit sloučit všechny záznamy jednotlivých exemplářů do společného záznamu. Interní multiplicity, které mohly vzniknout tímto způsobem, však nelze nijak rozpoznat, takže hypotéza o jejich existenci není podložená.

V neposlední řadě může skutečná multiplicita vzniknout při časovém rozestupu mezi akvizicemi jednotlivých exemplářů, kdy knihovník v systému nezkontroluje existenci bibliografického záznamu pro daný dokument a zkatalogizuje jej znovu.

Složitější situace nastává u dotisků a dokumentů vydaných v různých vazbách, kde nejsou pravidla pro katalogizaci zcela jednoznačná a ponechávají značný prostor

úsudku knihovníka (*Dotisky a varianty tištěných monografií popisované na jednom záznamu* 2019).

Ačkoliv je z katalogizační politiky zřejmý záměr neuchylovat se k vytváření nových záznamů, pokud to není opravdu nezbytné, katalogizátoři v českých knihovnách často aplikují opačnou logiku a vytvářejí nové záznamy pro jistotu pro všechny varianty a dotisky dokumentu v rámci jednoho vydání. Jedním z důvodů může být, že lze tento přístup považovat za jednodušší z hlediska nenáročnosti na ověřování informací v databázi ČNB nebo agentuře ISBN a kontrolu obsahové shody variant dokumentu.

Situace, kdy se do společného záznamu dostane více záznamů stejného dokumentu z jedné knihovny, nepředstavují pro portál Knihovny.cz žádný jiný problém, než snížení viditelnosti záznamů, které se uživateli nezobrazí jako první. Naopak, pokud se od sebe záznamy stejných dokumentů dostatečně liší, pak se spolu ani nezdeduplikují a místo interní multiplicity vznikne klasická viditelná multiplicita, která na sebe upoutá mnohem více negativní pozornosti uživatelů.

Skutečné duplicity tvoří až 70 % interních duplicit. O zhruba polovině z nich to lze prohlásit s vysokou mírou jistoty. U druhé poloviny jde pravděpodobně o skutečné duplicity, ale bez knihy v ruce to nelze potvrdit či vyvrátit.

VÍCEDÍLNÉ MONOGRAFIE

Existují dva přístupy ke katalogizaci vícedílných monografií. Při popisu shora se vytváří jeden společný záznam pro celý knižní soubor. Tento způsob zápisu je doporučen v případě, kdy jde o *„vícesvazkové monografie, jejichž jednotlivé části nemají různé tvůrce s hlavní odpovědností, nemají názvy částí, nebo tyto názvy částí nejsou významné pro samostatné vyhledávání a všechny části přísluší k témuž vydání. Typickými představiteli těchto dokumentů jsou slovníky, encyklopedie a jiné dokumenty, které byly vydány v jednotlivých svazcích z technických, nikoli obsahových důvodů"* (*Popis vícesvazkových monografií* 2019). V ostatních případech by měl knihovník zvolit popis zdola, tzn. vytvořit samostatný záznam pro každý díl.

Podobně jako v případě různých variant téhož vydání se i zde stává, že všichni katalogizátoři nevyhodnotí tutéž situaci stejně a vícedílná monografie je pak v některých knihovnách popsána shora a v jiných zdola. Obecně lze říci, že v knihovnách dominuje praxe popisovat vícedíly zdola za všech okolností. Popis shora praktikuje pouze úzká skupina převážně velkých knihoven, z nichž většina používá knihovní systém Aleph. Nejen na portálu Knihovny.cz, ale i v Souborném katalogu ČR jsou tak k nalezení obě možnosti zápisu.

Problém nastává tehdy, když knihovník přebírá metadata vícedílné monografie ze Souborného katalogu a vybere si jinou formu popisu, než je ta, kterou hodlá použít. Případně má k dispozici pouze záznamy s popisem shora pocházející z alephovských knihoven, které jsou nejvýznamnějšími přispěvateli do Souborného katalogu. Na portálu Knihovny.cz se pak vytvoří shluk obsahující jak záznamy jednotlivých dílů, tak záznamy celého souboru. Sloučením záznamů několika dílů z jedné knihovny tak vznikne interní multiplicita.

Některé knihovny praktikují ještě třetí variantu popisu, kdy vytvářejí hierarchii souborného záznamu pro celou vícedílnou monografii a dílčích záznamů pro jednotlivé části této monografie. Úplné bibliografické údaje mohou být obsaženy jen v rodičovských záznamech, jen v potomcích, nebo ve všech těchto záznamech. Údaje o knihovních jednotkách bývají typicky obsaženy jen v potomcích. Portál Knihovny.cz není uzpůsoben na prezentaci hierarchických záznamů, takže je musí umět rozlišit, sloučit a následně s nimi zacházet jako by se jednalo o jeden záznam, který popisuje celý soubor shora. To je samozřejmě možné pouze v případě explicitně uvedené vazby mezi takovými záznamy.

Interní multiplicity nad vícedílnými monografiemi, ale i knižními sériemi, mohou vznikat také kvůli nastavení deduplikačních pravidel na portálu Knihovny.cz. Není vzácností, že různé díly téže řady mají tolik stejných či podobných bibliografických údajů, že to dostačuje k jejich sloučení (a to i přes opatření, která portál zavádí na obranu před touto situací).

Slučování záznamů různých zdola popsaných částí vícedílných monografií lze považovat za nežádoucí jev, kterému ale při stávající rozmanitosti v katalogizaci nelze zabránit bez dopadu na slučování velkého množství dalších záznamů.

Shluků navzájem propojených popisů shora a zdola nebo různých dílů popsaných zdola jsou přibližně 2 % z celkového množství interních duplicit.

UČEBNICOVÉ SOUBORY, CD PŘÍLOHY APOD.

K dalším oblastem, které trpí na možnost vágní interpretace katalogizačních pravidel, patří popis různých setů. V případě učebnicových sad je situace velmi podobná vícedílným monografiím (*Popis soubourů/jednotlivých svazků učebnic* 2019).

Řada na sebe navazujících učebnic, kde je každý díl věnován jiné látce či pokročilosti, má být popsána zdola. Pro sady, jejichž části nelze používat samostatně (učebnice a pracovní sešit, slovník, klíč nebo materiál pro učitele), má být zvolen popis shora. Zároveň je ale stanovena výjimka, že knihovna, která *„z technických nebo organizačních důvodů nemůže popisovat tyto soubory učebnic na jednom společném záznamu, může zpracovat jednotlivé svazky samostatně“* (*Popis souborů/jednotlivých svazků učebnic* 2019). Podobné pravidlo platí i pro zdroje sestávající z několika typů nosičů, např. z knižního svazku a v něm vloženého zvukového CD, které se čtenářům vždy půjčují pohromadě (VOCHOZKOVÁ 2022).

Z pohledu na záznamy dodávané na portál Knihovny.cz je zřejmé, že technické či organizační důvody pro popisování nedílných setů dokumentů zdola má většina knihoven v ČR. Vznikne-li sloučením těchto záznamů na portálu Knihovny.cz interní multiplicita, nelze hovořit o chybě. Jedná se o záznamy, které se týkají stejného dokumentu, byť každý popisuje jen jeho část.

Tento typ shluků se na celkovém počtu interních duplicit podílí více než 3 %.

PŘÍVAZKY

Zvláštní skupinu interních multiplicit na portálu Knihovny.cz tvoří záznamy přívazků. Dle pravidel je třeba každému dokumentu v přívazku tvořit samostatný záznam (Paličková 2015). Záznam se ovšem vztahuje k dokumentu coby části přívazku, nikoliv k jeho vydání. Má-li jedna knihovna dokument A ve fondu jedenkrát samostatně, podruhé svázaný s dokumentem B a potřetí svázaný s dokumentem C, pak v jejím katalogu existují tři záznamy dokumentu A, aniž by to byla chyba.

Mnoho přívazků se vyskytuje v historických fondech, u kterých je exemplářový popis běžnou praxí, protože mohou být specifické svou vazbou, vpisem či vloženým ex libris. Lze proto diskutovat nad možností, zda tyto záznamy vůbec podrobovat deduplikaci. Na portálu Knihovny.cz se záznamy historických fondů slučují, protože tím nedochází ke ztrátě informace a lokální záznamy lze prohlížet v jejich původní podobě.

I při zákazu deduplikace záznamů starších dokumentů by ale nedošlo k vyřešení slučování novějších publikací, jejichž výtisky se ocitly v přívazku. To by bylo možné zákazem deduplikace záznamů obsahujících slovo „přívazek“ v poli 787.

Interní multiplicity vzniklé slučováním záznamů přívazků lze považovat za další typ skutečných duplicit. Jejich existence není chybou na straně knihoven ani na straně portálu Knihovny.cz.

Interních duplicit, které zahrnují aspoň jeden přívazek, jsou necelá 3 %.

KRITICKÉ MULTIPLICITY

Jeden z nejzásadnějších problémů, který může nastat při deduplikaci, je sloučení záznamů různých dokumentů. Ve valné většině případů k tomu dochází tehdy, když některý ze sloučených záznamů obsahuje chybu v klíčových údajích. Typicky se jedná o záznamy s chybným ISBN či číslem ČNB, u kterých se shodují ostatní porovnávané klíče (sem patří záznamy, které nereflektovaly zjištěné omyly při přidělování identifikátorů odstraněním údaje z podpole $a, záznamy vzniklé zkopírováním a nedostatečnou editací záznamu předchozího vydání nebo třeba záznamy audioknih uvádějící ISBN tištěné předlohy).

Vzhledem k tomu, že dva z deduplikačních kroků záměrně nepracují s názvovými údaji kvůli častým záměnám různých částí názvu, edice a dalších informací uvedených ve zdroji popisu, sloučit se mohou i záznamy zcela nesouvisejících dokumentů. Přes jeden chybný záznam se pak tranzitivně propojí záznamy obou různých dokumentů z dalších knihoven, i když tyto ostatní záznamy neobsahují žádnou chybu. Pokud je problematický záznam opraven knihovnou, ze které pochází, pak během dalšího spuštění deduplikace dojde k žádoucímu rozpadu sloučeného záznamu na počet shluků odpovídající počtu popisovaných dokumentů.

Záznamy, které jsou původci problematických tranzitivních shluků, lze nejlépe identifikovat právě tehdy, upozorní-li na sebe příslušností k interní multiplicitě. Snaha o jejich odhalení je hlavním důvodem kontroly interních multiplicit při zapojování nových knihoven do portálu Knihovny.cz.

Výjimečně nastávají také případy, kdy se sloučí záznamy různých dokumentů, aniž by kterýkoliv záznam ze shluku obsahoval chybu. Stačí, aby se dva záznamy ze shluku shodovaly v takové kombinaci klíčových údajů, která je porovnávaná v některém deduplikačním kroku. Stejnojmenných či velmi podobně nazvaných publikací se stejným autorem a přibližně stejným počtem stran, které navíc vyšly v témže roce, je sice minimum, ale existují (nejčastěji jde o monograficky popisované komiksové knihy ze stejné řady). Vzniku většiny z této úzké skupiny nežádoucích shluků by bylo možno zabránit zpřísněním

deduplikačních pravidel, což by ale vedlo k rozpojení mnohem většího množství správně sloučených záznamů. Jedná se tedy o stav, jehož řešení by přineslo více škody než užitku.

Kritických interních duplicit je přibližně 7 %. I přes své relativně malé množství ale mají největší vliv na použitelnost portálu.

Lze si všimnout, že procentuální součet skutečných, vícedílných, souborových, přívazkových a kritických duplicit dohromady netvoří 100 %. Zbytek do celku připadá většinou na tranzitivní duplicity (až 13 %), které se propojily přes některou z duplicit výše jmenovaného typu (zejména pak přes kritické duplicity) a na nezařaditelné duplicity (do 1,5 %). Do nezařaditelných duplicit patří dvojice záznamů, u kterých bez knihy v ruce nelze vůbec odhadnout, zda popisují stejný dokument či nikoliv. Jde o téměř identické záznamy, kde si navzájem výrazně odporují některé z klíčových údajů (např. zcela odlišné počty stran, různá místa vydání, různí autoři apod.).

KOMUNIKACE PROBLEMATIKY KNIHOVNÁM

Každá knihovna usilující o zapojení do portálu Knihovny.cz musí ještě před svým zapojením poskytnout všechny záznamy ze svého katalogu k předběžné deduplikaci a kontrole kvality metadat. Kontrola se soustředí na části záznamů relevantní právě pro deduplikaci a indexaci na portálu. Nezabývá se zkoumáním validity a syntaktické správnosti celého záznamu, jelikož nejde o faktor ovlivňující dobrou použitelnost portálu.

Sklizené záznamy jsou deduplikovány a zaindexovány v testovací databázi a indexu. Již podíl nezdeduplikovatelných záznamů většinou naznačuje, zda má knihovna metadata v dobrém stavu, nebo lze očekávat komplikace vyžadující rozsáhlejší opravy. Další analýza pak identifikuje konkrétní chyby a na jejím základě dostane knihovna souhrn doporučení k řešení těchto problémů.

Kromě nezdeduplikovaných záznamů je věnována pozornost i interním multiplicitám. Knihovny obvykle dostávají tři seznamy interních multiplicit s různou prioritou řešení.

První seznam se týká toho typu kritických multiplicit, kde záznamy zjevně různých dokumentů sdílejí stejné ISN nebo číslo ČNB. Kromě záznamů skutečně různých dokumentů se do tohoto seznamu často dostávají také záznamy stejných dokumentů, u kterých katalogizátor špatně rozpoznal jednotlivé části názvových údajů nebo je zaměnil například s názvem edice (tzn. názvy se jeví jako zcela odlišné). Knihovny jsou důrazně nabádány, aby identifikovaly chybné záznamy v těchto shlucích a opravily údaje, kvůli kterým došlo k vytvoření interní multiplicity.

Druhý seznam se týká toho typu kritických multiplicit, kde mají záznamy zjevně stejných dokumentů odlišná ISN nebo čísla ČNB. Do seznamu se samozřejmě dostávají i publikace s více přidělenými ISBN (např. různé vazby, dotisky, více nakladatelů) nebo díla vydaná ve stejném roce nezávisle v různých nakladatelstvích. Stejně jako v případě prvního seznamu jsou knihovníci žádáni o kontrolu a případnou opravu záznamů, ale zároveň jsou srozuměni s vyšší pravděpodobností falešně pozitivních záchytů kritických interních multiplicit.

Knihovny jsou upozorňovány, že by měly řešení kritických multiplicit přikládat vysokou prioritu, aby se nestaly příčinou vzniku nežádoucích tranzitivních shluků. Provádět systematické zpětné kontroly problematických záznamů poté, co knihovna projde procesem zapojování do portálu, není v možnostech řešitelského týmu Knihovny.cz (a to nejen kvůli časové náročnosti, ale právě i kvůli nemožnosti rozpoznat skutečný důvod sloučení záznamů bez knihy v ruce). Přesto se lze domnívat, že knihovny berou apel na kontrolu kritických multiplicit vážně. Tým portálu Knihovny.cz musí průběžně aktualizovat příklady nesprávně sloučených záznamů uváděných ve svých výukových materiálech, protože dochází k jejich postupnému mizení. Posun k lepšímu byl patrný i při procházení starších seznamů kritických multiplicit při přípravě tohoto článku.

Ve třetím seznamu dostávají knihovny záznamy skutečných a pravděpodobných interních multiplicit. Pravděpodobnými multiplicitami jsou myšleny shluky záznamů, kde aspoň jednomu z nich chybí dostatek informací k tomu, aby bylo možno s jistotou prohlásit, že popisuje stejný dokument jako ostatní záznamy. Ty údaje, které jsou pro všechny záznamy ve shluku společné, ale nejsou v rozporu a nic tedy nenaznačuje, že popisují více dokumentů.

Na rozdíl od seznamů kritických multiplicit, kterých bývá v průměrné městské knihovně nanejvýš pár desítek, seznam skutečných a pravděpodobných multiplicit obvykle čítá až stovky záznamů. Vzhledem k pracnosti jejich kontroly s knihou v ruce a bezproblémovosti jejich zpracování na portálu Knihovny.cz není řešení tohoto seznamu prioritou. Knihovny se jím mohou zabývat za účelem zpřehlednění vyhledávání ve vlastním katalogu nebo v případě, že by do budoucna uvažovaly o přispívání do Souborného katalogu či projektu CARDS.

AKTUÁLNÍ ZPŮSOB ROZPOZNÁVÁNÍ INTERNÍCH MULTIPLICIT PŘI ANALÝZE METADAT KNIHOVNY ZAPOJOVANÉ DO PORTÁLU KNIHOVNY.CZ

Pro analýzu metadat knihoven zapojujících se do portálu Knihovny.cz je využíván nástroj OpenRefine. Jedná se o software umožňující čištění a pokročilou manipulaci se strukturovanými daty s funkcemi na pomezí tabulkového procesoru a databáze. Podporuje fasety, regulární výrazy, omezené propojování různých tabulek a především ukládání posloupnosti provedených kroků do skriptu ve formátu JSON k pozdějšímu použití. Tato vlastnost je klíčová pro use case časté kontroly souborů stejného typu dat, protože umožňuje předprogramovat úkony jako překlopení řádkového MARCu do tabulky se sloupcem pro každé pole, propojení s tabulkou deduplikačních klíčů a vytvoření různých kontrolních či exportních sloupců.

Časem se ukázalo jako vhodné zavést podobnou praxi i na kontrolu interních multiplicit. Skript provede nad exportem klíčových údajů interních multiplicit sérii úkonů, které navzájem porovnají všechny výskyty každého údaje na základě jejich shody či rozdílnosti (ať už úplné či částečné) a zařadí je do některé kategorie. Zpracování bere v potaz i ne-rozdíl (tzn. prázdné hodnoty), porovnává oproti sobě některé řetězce podle nejkratšího z nich, pokouší se o sjednocení údajů o vydání v poli 250, toleruje malé rozdíly mezi počty stran a roky vydání atd.

Ač jsou možnosti nástroje OpenRefine skutečně nepřeberné, variabilitu zpracovávaných záznamů nelze podchytit žádným skriptem. Lze však alespoň zachytit a označit případy, kde je zapotřebí kontrola člověkem. Každá analýza interních multiplicit tak obnáší jak automatizovanou, tak manuální část. To, co je časově únosné při kontrole menších knihoven, se však stává téměř neproveditelným v případě kontroly větších knihoven nebo potřeby analyzovat interní multiplicity na celém portálu.

VÝHLEDOVÉ ŘEŠENÍ KLASIFIKACE INTERNÍCH MULTIPLICIT

Po analýze interních multiplicit desítek zapojovaných knihoven pomocí kombinace skriptování a ruční práce se jako vhodný způsob dalšího směřování začalo jevit využití strojového učení. Jde o typický problém vícetřídní klasifikace, který nelze zcela pokrýt zpracováním pomocí klasického algoritmu, protože vyžaduje opakovaná rozhodování o klasifikaci podobných, nikoliv však stejných případů. To činí z rozpoznávání typů interních multiplicit zjevného kandidáta pro zapojení umělé inteligence.

Nyní již byly podniknuty první pokusy s tvorbou modelů pro rozpoznávání typů interních multiplicit. Jejich výstupy zatím nebyly příliš spolehlivé, ale naznačují, že je zde prostor ke zlepšování při experimentování se širší škálou metod. S referováním o úspěšnosti využití strojového učení je ještě třeba počkat na výsledky budoucích pokusů.

Největším kamenem úrazu při klasifikaci interních multiplicit i řešení obdobných problémů, které by mohly benefitovat z nasazení umělé inteligence, se zatím jeví příprava dostatečného množství tréninkových dat. Jedná se o časově náročný proces – práce na jejich vytvoření poskytla podklady pro tento článek. Pro předběžné třídění lze použít skript. Je-li ale cílem vytvořit nástroj, který překoná úspěšnost skriptu a omezí potřebu lidského vstupu, musí následovat i manuální kontrola značné části dat. Tímto způsobem byly zatím přiděleny labely několika tisícům řádků.

Analýzy interních multiplicit knihoven zapojovaných v posledních nebolo možné jednoduše použít jako tréninková data, protože byly tvořeny odlišnými metodami, a záznamy, které v nich byly uváděny, už dnes byly zčásti opraveny nebo neexistují. I v případě zpracování těchto výstupů by tímto způsobem vzniklo nanejvýš jen pár tisíc řádků, přičemž by bylo vhodné jich mít řádově více.

Další překážku při nasazení strojového učení představuje variabilita způsobů zápisu téhož typu údaje. Modely obvykle používané pro vícetřídní klasifikaci totiž pracují nejlépe tehdy, mají-li na vstupu vyčištěná data (ALDOSERI 2023). Ukázkovým příkladem takřka nevyčistitelných dat, která mohou snižovat úspěšnost klasifikace, je například údaj o vydání v podpoli 250$a. Vyskytují se zde číslice i slovní číslovky, celá slova i jejich zkratky, a to navíc v různých jazycích. Jako možné řešení tohoto problému jeví nasazení předtrénovaného jazykového modelu, který však pro své vyladění potřebuje ručně připravit vlastní sadu tréninkových dat.

Aktuální diskurz kolem využití umělé inteligence v knihovnictví je optimistický a vychází z předpokladu, že knihovny disponují obrovským množstvím dat. Zaznívají návrhy na její aplikaci při doporučování literatury z fondů knihoven čtenářům, při zavedení podpory vyhledávání v přirozeném jazyce do knihovních katalogů nebo při opravě subminimálních záznamů.

Bližší pohled ale ukáže, že samotné množství dat pro uskutečnění těchto nápadů nestačí. Čtenáři detektivek doporučí umělá inteligence literaturu téhož žánru pouze tehdy, pokud knihy, které si půjčoval, i knihy, které mu budou doporučeny, budou mít kvalitní věcný popis. Zadávání dotazů pomocí chatbotu by nebylo nutně efektivnější

než klasické vyhledávání, obzvláště nad obsahově chudými záznamy. A obsahově chudým záznamům by zjevně neprospělo, kdyby byly doplněny GPT halucinacemi.

Úspěšné projekty uskutečněné např. nad digitalizovanými dokumenty ukazují, že umělá inteligence má v knihovnictví své uplatnění. Její možnosti však nejsou neomezené.

ZÁVĚR

Tento článek rozebírá problematiku interních multiplicit v agregátorech bibliografických metadatových zdrojů, zejména pak na portálu Knihovny.cz. Ke sloučení více záznamů ze stejného zdroje dochází nejčastěji tehdy, když byl tentýž informační objekt zkatalogizován vícekrát. V takových případech není chybou, že se tyto záznamy zdeduplikují a záleží na politice každého portálu či katalogu, jak s takovou interními multiplicitami naloží.

Složitější situace nastávají tehdy, dojde-li ke vzniku interní multiplicity nad záznamy různých dokumentů. Příčinou může být chyba v některém ze záznamů, různá interpretace katalogizačních pravidel nebo i samotné nastavení deduplikačního algoritmu. Interní multiplicity tohoto typu totiž mají potenciál stát se jádrem tranzitivních multiplicit, tzn. sloučení správných záznamů různých dokumentů.

Při komunikaci ohledně multiplicit vzniklých chybou na straně knihovny je vhodné zadání dávkovat a prioritizovat, a tím pádem i zvýšit pravděpodobnost vyřešení nejzávažnějších problémů. Hledání chyb a opravy záznamů způsobujících interní multiplicity je totiž většinou nezbytné provádět s knihou v ruce, a jde tedy o časově velmi náročnou činnost. Proveditelné zadání pro opravy multiplicit je výstupem analýzy metadat přizpůsobené na míru daného agregátoru.

Ačkoliv lze značnou část tohoto procesu automatizovat, analýza se stále neobejde bez alespoň částečné manuální kontroly, což je ve větším měřítku extrémně neefektivní způsob práce. Proto je plánováno pokusit se v budoucnosti na klasifikaci typů interních multiplicit nasadit nástroje strojového učení.

Dedikace: Publikace vznikla v rámci Institucionální podpory na dlouhodobý koncepční rozvoj výzkumné organizace – Moravská zemská knihovna v Brně.

POUŽITÁ LITERATURA

ALDOSERI, Abdulaziz, Khalifa N. AL-KHALIFA a Abdel M. HAMOUDA, (2023). *Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges.* Applied Sciences. Online. vol. 13, iss. 12. ISSN 2076-3417. Dostupné z: https://doi.org/10.3390/app13127082. [cit. 2024-09-17].

Dotisky a varianty tištěných monografií popisované na jednom záznamu, (2019). Online. Národní knihovna ČR: Katalogizační politika. Dostupné z: https://text.nkp.cz/o-knihovne/odborne-cinnosti/zpracovani-fondu/katalogizacni-politika/dotisky-prosinec-2019. [cit. 2024-09-11].

Knihovny.cz: Zapojené knihovny a zdroje, (© 2015–2024). *Knihovny.cz*. Online. Brno: MZK. Dostupné z: https://www.knihovny.cz/Content/zapojene-knihovny-a-zdroje. [cit. 2024-09-11].

KURFÜRSTOVÁ, Jana, Petra ŽABIČKOVÁ a Tomáš CEJPEK, (2023). *Deduplikace na portálu Knihovny.cz*. Knihovna: knihovnická revue. Online. roč. 34, č. 1, s. 48–62. ISSN 1802-3250. Dostupné z: https://knihovnarevue.nkp.cz/archiv/2023-1/recenzovane-prispevky/deduplikace-na-portalu. [cit. 2024-09-11].

PALIČKOVÁ, Tereza, (2015). *Stručná rukověť pro katalogizaci starých tisků podle RDA*. Online. Praha: Národní knihovna ČR, Oddělení rukopisů a starých tisků. Dostupné z: http://text.nkp.cz/o-knihovne/odborne-cinnosti/oddeleni-rukopisu-a-starych-tisku/strucna-rukovet-pro-katalogizaci-starych-tisku-podle-rda. [cit. 2024-09-12].

Popis souborů/jednotlivých svazků učebnic, (2019). Online. Národní knihovna ČR: Katalogizační politika. Dostupné z: https://www.nkp.cz/o-knihovne/odborne-cinnosti/zpracovani-fondu/katalogizacni-politika/popis-souboru-jednotlivych-svazku-ucebnic-2013-brezen-2019. [cit. 2024-09-12].

Popis vícesvazkových monografií, (2019). Online. Národní knihovna ČR: Zpracování fondů. Dostupné z: https://www.nkp.cz/o-knihovne/odborne-cinnosti/zpracovani-fondu/zapisy-z-jednani/popis-vicesvazkovych-monografii. [cit. 2024-09-12].

Record Manager 2, (© 2024). GitHub. Online. Brno: MZK. Dostupné z: https://github.com/moravianlibrary/RecordManager2. [cit. 2024-17-09].

VOCHOZKOVÁ, Hana, (2022). *Dodatky.* Kurz MARC 21/RDA – Tištěné monografie – doporučená úroveň. Online. Dostupné z: https://webserver.ics.muni.cz/hanan/SKCR2022D/SKCR2022D_Dodatky.pdf. [cit. 2024-09-12].

■ **Článok bol recenzovaný.**