DOPORUČOVÁNÍ PODOBNÝCH DOKUMENTŮ NA PORTÁLU KNIHOVNY.CZ POMOCÍ NÁSTROJE BIBLIOLINKER

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)
Ing. Petr Žabička; Petr.Zabicka@mzk.cz; (Moravská zemská knihovna v Brně)

Tento článek popisuje fungování nástroje Bibliolinker, určeného ke shlukování bibliografických záznamů dle podobnosti na portálu Knihovny.cz. Výstupy Bibliolinkeru se na portálu zobrazují v náhledu úplného záznamu jako seznam odkazů „Podobné". Bibliolinker je softwarovým modulem Správce zdrojů, který pro Knihovny.cz zajišťuje sklízení, deduplikaci a indexaci záznamů z institucí zapojených do portálu. Princip slučování dle podobnosti vychází z deduplikace a prohlubuje hierarchii shluků o další dvě úrovně. Článek dále zvažuje klady a zápory Bibliolinkeru v porovnání s jinými způsoby řešení téhož problému.

http://doi.org/10.52036/1335793X.2023.SC2.13-25

ÚVOD

Řada knihovních online katalogů je dnes vybavena funkcemi umožňujícími doporučit uživateli, jaké dokumenty z fondu jeho knihovny by jej mohly zajímat. Totéž pak jejich uživatel očekává i v discovery systémech agregujících záznamy z vícera knihoven a bibliografií. V České republice je takovým agregátorem portál Knihovny.cz spravovaný Moravskou zemskou knihovnou v Brně (Knihovny.cz: Základní informace). Knihovny.cz umožňují uživatelům prohledávat záznamy z více než 120 zapojených zdrojů (Knihovny.cz: Zapojené knihovny a zdroje) a přistupovat ke svým čtenářským kontům bez nutnosti navštěvovat jednotlivé databáze a katalogy knihoven. Doporučování titulů na základě podobnosti k právě prohlíženému záznamu zde bylo původně ošetřeno pomocí zabudované funkce indexu, ale její výpočetní náročnost způsobovala problémy s rychlostí odezvy indexu. Za účelem zlepšení této situace byl vyvinut nástroj Bibliolinker, který využívá a rozšiřuje již dříve zavedené procesy zpracování dat na portálu. Cílem tohoto článku je představit princip Bibliolinkeru jako jednu z alternativ pro řešení doporučovacích funkcí.

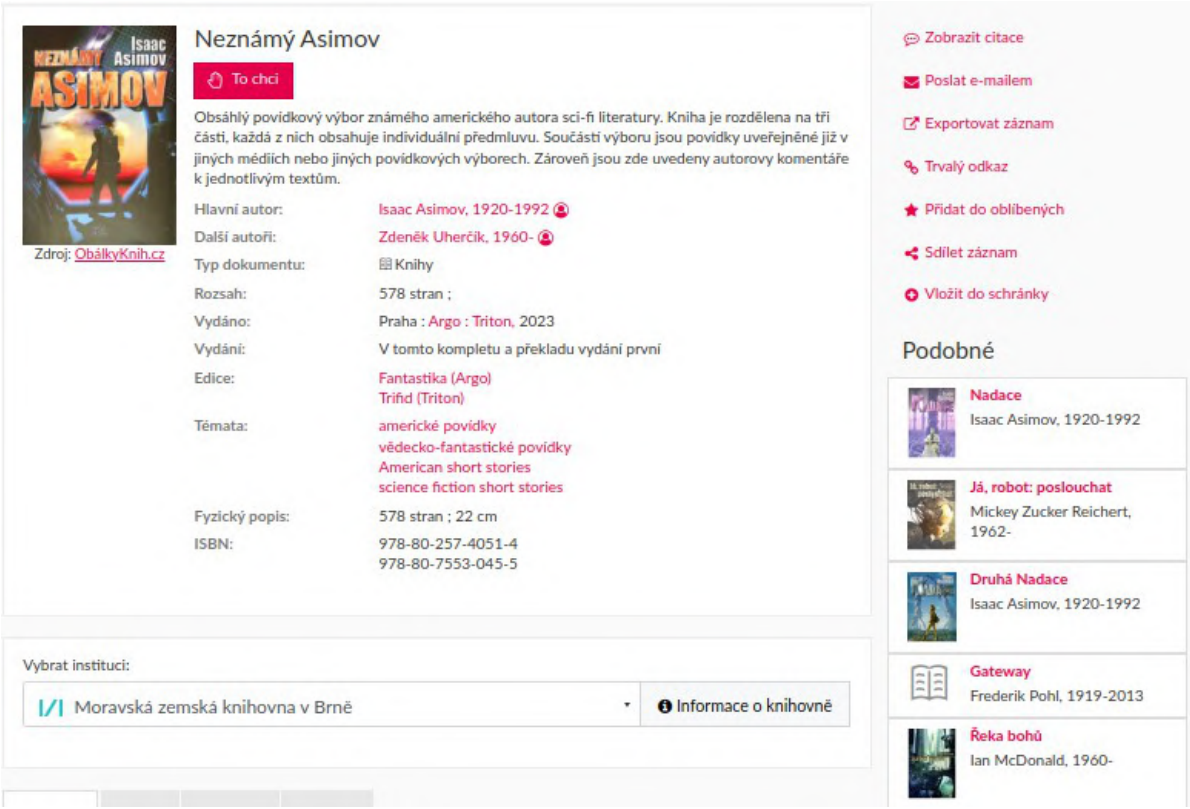
VÝCHODISKA A OMEZENÍ PRO ŘEŠENÍ DOPORUČOVACÍ FUNKCE V PROSTŘEDÍ PORTÁLU KNIHOVNY.CZ

Knihovny.cz využívají prostředí VuFind (VuFind: Search. Discover. Share), které je značně rozšířené mezi discovery systémy. Seznam doporučení podobných je zde součástí pravého panelu v náhledu úplného záznamu (viz Obrázek 1).

VuFind na portálu Knihovny.cz ke svému provozu vyžaduje index postavený na platformě Solr. Seznam s doporučeními je v původním nastavení VuFindu naplňován prostřednictvím funkce Solru MoreLikeThis (Apache Solr Reference Guide: MoreLikeThis), tj. pomocí dotazu do indexu, který v reálném čase dohledává záznamy podobné právě zobrazenému záznamu.

Toto řešení znatelně zvyšovalo vytížení indexu, a tedy i délku načítání stránek. Zároveň šlo o úkon, který pro fungování portálu není klíčový. Klíčová je pro portál vyhledávání a zjišťování informací o dokumentech, kde rychlost odezvy indexu zásadně ovlivňuje uživatelský zážitek. Proto bylo žádoucí usilovat o snížení náročnosti doporučovací funkce.

Místo automatického dohledávání seznamu doporučení v okamžiku zobrazení záznamu byla zvolena cesta doindexování seznamu podobných přímo do záznamu. Tím se otevřela možnost specifikovat vlastní pravidla pro tvorbu tohoto seznamu, která by generovala různorodější výstupy než jednotným nastavením říze-



Obrázek 1 Podobné na portálu Knihovny.cz

ná funkce MoreLikeThis. Doporučovací nástroj s pracovním názvem Bibliolinker byl plánován a realizován jako jeden z výstupů projektu NAKI II (CPK – Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven), který zastřešoval vývoj celého portálu Knihovny.cz (tehdy Centrálního portálu knihoven).

Při návrhu Bibliolinkeru bylo třeba vzít v potaz specifické prostředí portálu Knihovny.cz, které bránilo zvolit některé časté způsoby řešení doporučovacích funkcí, ale naopak umožňovalo využít zde již používané nástroje a procesy. Tento článek je tedy míněn i jako inspirace komukoliv, kdo by se potýkal s podobnými omezeními.

Knihovní katalogy mají možnost doporučovat dokumenty na základě znalosti chování čtenářů v podobě uživatelského hodnocení, údajů o historii výpůjček nebo statistik nejpůjčovanějších titulů. Knihovny.cz nedisponují použitelnými daty tohoto typu. Akce přihlášených uživatelů související s jejich čtenářskými konty probíhají na straně systémů zapojených knihoven, nikoliv na straně portálu – ten o nich proto nemá jakýkoliv přehled.

Přesto nelze říci, že by statistiky o výpůjčkách na národní úrovni nebyly pro generování seznamu doporučení použitelné. Funkční řešení využívající údaje o výpůjčkách z několika větších knihoven bylo vyvinuto v rámci projektu Obálkyknih.cz a zobrazování jeho výstupů nyní podporují katalogy knihoven se systémem ARL (O projektu Obálkyknih.cz).

Výhodou nezávislosti na údajích o výpůjčkách je schopnost tvorby doporučení u záznamů dokumentů, které jsou půjčovány jen zřídkakdy. Takové dokumenty tvoří podstatnou část fondů knihoven zapojených do portálu Knihovny.cz, zejména jde-li o knihovny s konzervační funkcí. Zobrazování odkazů na záznamy málo půjčovaných dokumentů může upozornit případného zájemce na jejich existenci.

PRINCIP FUNGOVÁNÍ BIBLIOLINKERU

Při návrhu Bibliolinkeru se předpokládalo, že nejschůdnější cestou bude stavět na již vyvinutých funkcích a nezavádět do beztak složitého systému další provozně náročný software. Sklizeň, deduplikaci a indexaci dat prohledatelných portálem Knihovny.cz zajišťuje nástroj Správce zdrojů vyvinutý v Moravské zemské knihovně v Brně (Record Manager 2). Biblio-

linker funguje jako jeho rozšíření a využívá postupy vyvinuté pro účely deduplikace.

DEDUPLIKACE NA PORTÁLU KNIHOVNY.CZ

Deduplikační algoritmus Správce zdrojů je podrobně popsán v článku Deduplikace na portálu Knihovny.cz (Kurfürstová et al., 2023). Spočívá ve vytvoření sady klíčů z bibliografických údajů, jejichž kombinace identifikují popisované dokumenty. Různé kombinace těchto klíčů jsou porovnávány sadou kroků za účelem zjistit, zda více záznamů popisuje tentýž dokument na úrovni provedení. V případě úspěchu jsou tyto lokální záznamy provázány sloučeným záznamem.

Díky každodenní sklizni musí deduplikace reflektovat průběžně přibývající, ubývající a pozměněné záznamy. Zdeduplikované shluky se mohou kdykoliv dále spojovat či rozpadat, takže sloučené záznamy nejsou permanentní – jde spíše o virtuální konstrukty, které dynamicky zachycují aktuálně platné vazby. Lokální záznamy jsou uchovávány pokud možno ve stejné podobě, v jaké byly sklizeny ze zdroje zapojeného do portálu. Ve výsledcích vyhledávání se sloučený záznam zobrazuje jako jedna položka. V náhledu úplného záznamu je pak možné přepínat mezi zobrazeními lokálních záznamů.

Velké úskalí představuje pro deduplikaci nízká kvalita, nejednotnost a neúplnost značné části zpracovávaných metadat. Způsob tvorby deduplikačních klíčů i slučovací kroky, ve kterých jsou klíče porovnávány, vychází ze snahy tuto skutečnost pokud možno kompenzovat. To především znamená, že hodnoty ukládané jako klíče jsou normalizovány za účelem sjednocení různých způsobů zápisu téhož údaje, že ke slučování záznamů nedochází výhradně na základě shody jednoho údaje s nezanedbatelnou chybovostí (ISBN, ISSN…) nebo že značná část deduplikačních kroků je určena ke zpracovávání obsahově chudých záznamů. Problematika kvality metadat je významným faktorem i v případě slučování dle podobnosti pomocí Bibliolinkeru.

Úkolem Bibliolinkeru je rozpoznat a doporučit podobné, nikoliv však stejné dokumenty. Stejnými dokumenty zde na rozdíl od deduplikace nelze myslet stejná vydání, potažmo provedení díla, ale spíše stejná vyjádření. Cílem totiž není doporučovat uživateli pět různých knižních vydání Babičky, ale raději zfilmovanou Babičku či jiná díla Boženy Němcové.

Bibliolinker proto zavádí další dvě úrovně shlukování nad rámec původní hierarchie lokálních a sloučených záznamů. Procesy Bibliolinkeru byly začleněny do každodenního zpracování nových a aktualizovaných dat, stejně jako je tomu v případě deduplikace.

SLUČOVÁNÍ PRVNÍ ÚROVNĚ

První úroveň shlukuje záznamy popisující tentýž dokument přibližně na úrovni vyjádření (Funkční požadavky na bibliografické záznamy). Stejně jako při deduplikaci vznikají sloučené záznamy porovnáváním lokálních záznamů. Vazby jsou ale identifikovány vlastními slučovacími kroky, které porovnávají jiné kombinace klíčů. Další odlišnost spočívá v propojování celých zdeduplikovaných shluků v případě, že kterékoliv jejich lokální záznamy splní podmínky pro sloučení.

Existence shluků první úrovně není patrná uživateli portálu, ale je nezbytná k zajištění relevantních výstupů druhé úrovně, které uživatel vidí coby seznam doporučení v náhledu záznamu. První úroveň má vylučovací funkci a jejím účelem je zabránit vzájemnému doporučování záznamů, které jsou si příliš podobné, takže pravděpodobně popisují stejné vyjádření dokumentu. Z toho vyplývá, že ve shlucích první úrovně spolu musí skončit i záznamy, které byly sloučeny při deduplikaci.

SLUČOVÁNÍ DRUHÉ ÚROVNĚ

Druhá úroveň slučování už slouží k nacházení souvisejících záznamů, které mohou být navzájem odkazovány v seznamu doporučení. Jak již bylo řečeno, příslušnost záznamů do stejného shluku první úrovně zabraňuje jejich provázání na druhé úrovni.

Výstupem slučování druhé úrovně je databázová tabulka se strukturou: *ID lokálního záznamu – ID lokálního záznamu přiřazeného některým krokem druhé úrovně – krok, kterým došlo k přiřazení*. Z této tabulky se při indexaci sestavuje seznam doporučení. Každý krok druhé úrovně generuje lokálnímu záznamu maximálně tolik vazeb, kolik jich zbývá k naplnění seznamu doporučení, čímž se šetří výpočetním časem.

SEZNAMY DOPORUČENÍ

Způsob sestavování výsledných seznamů a jeho délka je otázkou nastavení. Nastavení by ale mělo brát v potaz relevanci různých typů vazeb (např. vazba mezi různými díly cyklu má větší význam než společné téma) a na rozmanitost výsledného seznamu (např. v případě Babičky je žádoucí mít možnost doporučit nejen další díla od Němcové, ale i filmovou adaptaci nebo biografii autorky).

SLUČOVACÍ KLÍČE A POROVNÁVACÍ KROKY

Bibliolinker je, stejně jako celý Správce zdrojů, široce konfigurovatelný. Po vyřešení otázky principu fungování Bibliolinkeru tedy přišla na řadu problematika nastavení tvorby klíčů, jejich porovnávání ve slučovacích krocích a sestavování výsledných seznamů doporučení.

První verze pravidel vznikla na základě analýzy použitelnosti a užitečnosti potenciálně užitečných MARCových (pod)polí v metadatech knihoven zapojených do portálu Knihovny.cz. Přestože už tehdy byl z analýzy pro zjednodušení vyloučen téměř všechen věcný popis se záměrem jeho pozdějšího doplnění, výsledná nastavení nebylo možné zavést do běžného provozu z důvodu vysoké výpočetní náročnosti. Navržená pravidla byla příliš složitá a navíc generovala velké množství extrémně početných shluků, které nebyly zpracovatelné v přijatelném čase. Ze stejného důvodu bylo nutné zůstat u porovnávání klíčů na přesnou shodu a vzdát se podobnostního porovnávání s využitím Levenshteinovy vzdálenosti, které se aplikuje na názvové klíče v některých deduplikačních krocích.

Po zásadní redukci původního nastavení už je možné spouštět Bibliolinker každý den spolu s ostatními pravidelnými procesy. Relevance vygenerovaných výstupů zůstala bez uživatelem pozorovatelné újmy.

SLUČOVACÍ KLÍČE

Na rozdíl od deduplikace se při vytváření klíčů pro účely Bibliolinkeru upřednostňuje schopnost získat alespoň nějaký použitelný údaj před snahou získat pokud možno přesnou a jednoznačnou informaci. Právě proto sestávají algoritmy pro tvorbu některých klíčů z dlouhé řady navazujících podmínek, které hledají možný klíč v mnoha polích, dokud neuspějí. Tak dostávají šanci i obsahově chudší záznamy. Způsob tvorby všech klíčů používaných Bibliolinkerem je popsán v Tabulce 1.

Např. do klíče s údaji o odpovědnosti bl_author se Bibliolinker nejprve pokouší získat kód autority ze 1XX polí. Není-li k dispozici autorita, zpracovává se kombinace podpole 1XX$a a podpolí obsahující další zpřesňující údaje. U záznamů bez 1XX polí se stejným způsobem pracuje s prvními výskyty 7XX polí. U videodokumentů se berou v potaz i kódy rolí, jsou-li k dispozici (autoři literárních předloh zde mají přednost před filmařskými profesemi). Podobným způsobem jsou naplňovány i ostatní klíče.

U klíče *bl_entity* používaného pro slučování na druhé úrovni stojí za zmínku jistá odchylka od logiky vytváření ostatních klíčů. Aby bylo možné identifikovat vazby typu „dílo od autora“ a „dílo o autorovi“, je nutné plnit klíč údaji významově odlišného charakteru. Jde o klíč, který může mít více hodnot na jeden záznam a současně používá jmenné i věcné údaje. Za účelem podpory doporučování dokumentů vzniklých na půdě téhož pracoviště jsou do klíče *bl_entity* zahrnovány i afiliace autorů.

Dalším případem klíče naplňovaného různými typy údajů *bl_common_title* pro identifikaci vazeb mezi jednotlivými díly knižních cyklů, které bývají často nesprávně zapisovány jako vícedílné monografie. Příslušný klíč čerpá přednostně z pole 787 s návěštím „Z cyklu:“ a v případě neúspěchu je brán ze společných názvových údajů u záznamů obsahujících název či číslo části.

U jediného plně věcného klíče *bl_topic_key*, který pracuje s kódy termínů kontrolovaných slovníků (předmětová hesla Národních autorit, MeSH, PSH), bylo nutné zavést stop hodnoty u nejobecnějších pojmů vyskytujících se ve statisících záznamů. Jde např. o MeSH kód pro termín „lidé“, který může být přiřazen jakémukoliv dokumentu týkajícímu se humánní medicíny, nebo identifikátor věcné autority „osobnosti“ uváděný u dokumentů obsahujících biografické informace.

Bibliolinker využívá i deduplikační klíče pro názvové údaje *title*, edice *issn_series* a typ dokumentu *format*.

SLUČOVACÍ KROKY PRVNÍ ÚROVNĚ

Při klasické deduplikaci může docházet i k několika pokusům o sloučení jednoho záznamu pomocí různých kombinací klíčů. Při slučování na úrovni vyjádření se s každým záznamem pracuje většinou jen jednou. Jednotlivé kroky porovnávají záznamy na velmi obecné úrovni a na rozdíl od deduplikace brání slučování napříč různými typy dokumentů či jejich skupinami. Např. lze sloučit zvukové dokumenty s různými nosiči (např. LP a CD), ale nikoliv zvukový dokument s videodokumentem, byť jsou oba na stejném typu nosiče (např. DVD).

U periodik a analytik se provedení dokumentu obvykle kryje s vyjádřením (tzn. nejde o opakovaně publikované dokumenty), takže pro vytvoření shluků první úrovně bylo možno využít výstupů deduplikace. Periodikům, jejichž názvy a ISSN se mohou v průběhu času měnit, byl přidělen jeden krok navíc umožňující sloučení za benevolentnějších podmínek než deduplikace. Slučování založené převážně na porovnávání názvo-

bl_author jednohodnotový
100$7 > **110$7** > **111$7** > **100$ad** > **110$abcdn** > **111$acdn** > **700$7** > **710$7** > **711$7** > **700$ad** > **710$abcdn** > **711$acdn**
U videodokumentů se výše uvedená sekvence zkouší nejprve na polích s kódem role používaným pro autora předlohy či námětu (***ant*** *|* ***aut*** *|* ***ccp*** *). Z podpole* ***X00$d*** *brány jen první 4 číslice (nejčastěji rok narození). Důvodem je podchycení záznamů bez autorit, kde nedošlo k reflexi úmrtí autora.*
bl_entity vícehodnotový
100$7 > **110$7** > **111$7** > **100$ad** > **110$abcdn** > **111$acdn**
první 3 výskyty **700$7** > **710$7** > **711$7** > zbytek do prvních 3 výskytů **700$ad** > **710$abcdn** > **711$acdn**
první 3 výskyty **600$7** > **610$7** > **611$7** > zbytek do prvních 3 výskytů **600$ad** > **610$abcdn** > **611$acdn**
100$u > **700$u** > **314$a**
U ***7XX*** *polí brány v potaz jen role s kódem (* ***aut*** *|* ***cmp*** *|* ***ive*** *|* ***org*** *|* ***drt*** *|* ***ant*** *|* ***ccp*** *). U podpolí* ***X00$d*** *opět brány jen první 4 číslice.* ***X00$u*** *nese informaci o afiliaci autora (u Bibliographia Medica Čechoslovaka* ***314$a*** *).*
bl_title vícehodnotový
130$anp
730$anp
210$a
222$a
Z podpolí ***130$a*** *a* ***730$a*** *brán jen obsah vně závorek (závorky většinou obsahují údaje nepatřící do názvu). Účelem klíče je rozšíření deduplikačního klíče* ***title*** *o hodnoty užitečné pro Bibliolinker. Používáno vždy v kombinaci s klíčem* ***title*** *.*
bl_common_title vícehodnotový
787$t > **24[056]$a**, **24[56]$ab**
787$t *zpracováváno, když* ***787$i*** *obsahuje hodnotu* ***"Z cyklu: "*** ***24X$ab*** *zpracováváno jen v případě existence* ***24X$n*** *nebo* ***24X$p*** *.*
bl_series jednohodnotový
490$a > **440$a**
bl_publisher jednohodnotový
264$b > **260$b** > **260$f** > **928$a**
bl_language jednohodnotový
041$a > **041$d** > **008|35-37**
bl_topic_key vícehodnotový
zřetězené a abecedně seřazené první 3 hodnoty **650$7** > **651$7**
072$a
Indikátor 1 pole ***650*** *nesmí mít hodnotu 2 (tj. bez anglických překladů věcných autorit). U kartografických dokumentů je* ***651$7*** *vyhodnocováno před* ***650$7*** *. Klíč se užívá i ve specifickém případě adresáře knihoven s modifikovaným MARCem.*

Tabulka 1 Slučovací klíče

název kroku	typy dokumentů	porovnávané klíče	sloučeno lokál. zázn.	trvání kroku	trvání kroku (%)
blTitleAuthorStep	monografie, kartografické dokumenty	(title + bl_title) + bl_author + bl_language + format	7 467 154	3:14:39	14,41%
blTitleAuthorAudioStep	audiodokumenty	(title + bl_title) + bl_author	563 184	0:59:15	4,39%
blTitleAuthorVideoStep	videodokumenty	(title + bl_title) + bl_author	40 031	0:36:12	2,68%
blTitleAuthorMusical ScoreStep	hudebniny	(title + bl_title) + bl_author	521 980	0:56:27	4,18%
blTitleLangPeriodical Step	periodika	(title + bl_title) + bl_language	762 671	1:01:38	4,56%
blRestDedupStep	*vše dosud nesloučené*	*deduplikace*	*29 509 929*	*15:42:33*	*69,78%*

Tabulka 2 Slučovací kroky první úrovně

vých a autorských údajů s sebou nese riziko vzniku příliš početných shluků. S extrémním množstvím klíčů v těchto shlucích je náročné pracovat na druhé úrovni slučování. Problematické shluky se zatím stále vyskytují v přijatelném množství a s kompletní sadou jejich klíčů se navíc pracuje pouze za předpokladu, že se na druhé úrovni nepodaří sloučit jejich nejkvalitnější záznamy (viz dále). Přehled slučovacích kroků první úrovně je uveden v Tabulce 2.

SLUČOVACÍ KROKY DRUHÉ ÚROVNĚ

Identifikaci kandidátů do seznamu doporučení zajišťuje jedenáct slučovacích kroků rozepsaných v Tabulce 3.

Prvních osm kroků porovnává pouze klíče pravděpodobně nejkvalitnějších záznamů ze shluků první úrovně. Jedná se o záznamy, kterým byla vypočtena nejvyšší váha dle původu záznamu, deklarované úplnosti popisu (LDR|17, existence povinných polí, užití autorit, souladu indikátoru názvového pole se záhlavím apod.

První krok usiluje o rozpoznání dokumentů ze stejné série, cyklu či vícedílné monografie porovnáváním autora a hlavních názvů spolu s názvy sérií.

Cílem následujících tří kroků je nalézt vazby typu kniha-audiokniha, knižní předloha-adaptace a hudební nahrávka-hudebnina. V těchto krocích je povoleno pouze vzájemné propojování odlišných typů dokumentů a každý krok smí naplnit nejvýše jedno místo v seznamu doporučení

Pátý a šestý krok vyhodnocuje, zda jiná publikace náleží do stejné edice. Tyto kroky nepropojují shluky první úrovně, ale pouze zdeduplikované záznamy. Edice jsou totiž vázány na konkrétní vydání a není žádoucí navzájem doporučovat díla, která mají společné jen tolik, že jednou vyšla ve stejné edici (obzvláště v případech široce pojatých edic jako Světová četba).

Sedmý krok využívá klíče *bl_entity* a hledá vazby přes osoby, korporace a akce ve jmenném a věcném popisu, zatímco osmý krok pracuje s údaji věcného popisu v klíči *bl_topic_key*. V případě textových dokumentů nyní sedmý a osmý krok zpracovává pouze dokumenty v češtině.

Devátý a desátý krok porovnávají klíče *bl_entity* a *bl_topic_key* spolu s jazykem nad všemi lokálními záznamy ve shluku, a nikoliv jen nad záznamy s největší váhou. K těmto krokům se přistupuje pouze tehdy, když se v prvních osmi krocích nepodařilo najít dostatek záznamů k naplnění seznamu doporučení. Tímto způsobem je výrazně redukován počet případů, kdy je nutné zpracovávat všechny klíče mnohočetných shluků.

První až desátý krok vždy vytváří vazby mezi záznamy z téhož zdroje, protože lze předpokládat, že si uživatel prohlíží záznam ze své knihovny a zajímají jej opět záznamy z jeho knihovny. Při slučování tedy nejde o to, z jakých lokálních záznamů pocházely porovnávané klíče, ale jaké shluky první úrovně byly provázány. Ke každému lokálnímu záznamu jsou preferenčně doporučovány záznamy ze stejného zdroje. Tzn. i když byla podobnost dvou shluků první úrovně rozpoznána např. pomocí klíčů ze záznamů Národní knihovny, kutnohorskému záznamu z jednoho shluku se přiřadí kutnohorský záznam z druhého shluku (pokud se tam takový vyskytuje).

název kroku	typ vazby	typy dokumentů	porovnávané klíče	prolinkováno lokál. zázn.	trvání kroku	trvání kroku (%)
KROKY NAD LOKÁLNÍMI ZÁZNAMY S NEJVYŠŠÍ VÁHOU ZE SHLUKŮ 1. ÚROVNĚ						
blAuthorCommonTitleStep	cykly, vícedíly	vše kromě periodik	bl_author + bl_common_title + bl_language + format	1 898 611	0:51:32	1,84%
blAuthorTitleAudioMusicalScoreStep	hudebnina - interpretace	audiodokumenty, hudebniny	bl_author + (title + bl_title)	7 916	0:05:19	0,19%
blAuthorTitleAudioBookStep	kniha - audiokniha	audiodokumenty, monografie	bl_author + (title + bl_title)	319 324	0:11:23	0,41%
blAuthorTitleVideoBookStep	film - předloha	videodokumenty, monografie	bl_author + (title + bl_title)	9 506	0:12:52	0,46%
blIssnSeriesStep	edice	monografie	issn_series	205 539	0:09:38	0,34%
blSeriesPublisherStep	edice	vše kromě periodik a analytik	bl_series + bl_publisher	5 560 995	2:36:04	5,58%
blEntityStep	osoba / korporace / akce	audiodokumenty, videodokumenty, hudebniny, monografie (cze), kartografické dokumenty (cze), analytika (cze)	bl_entity	15 503 698	14:53:55	31,97%
blTopicKeyStep	téma	viz blEntityStep	bl_topic_key	1 168 142	1:22:04	2,94%
blLibrariesStep	knihovna	knihovna ADR záznam	bl_topic_key	6 508	0:00:24	0,01%
KROKY NAD VŠEMI LOKÁLNÍMI ZÁZNAMY DOSUD NESLOUČENÝCH SHLUKŮ 1. ÚROVNĚ						
blEntityLangRestStep	osoba / korporace / akce	vše dosud nesloučené	bl_entity + bl_language	7 771 711	13:11:21	28,30%
blTopicKeyLangRestStep	téma	vše dosud nesloučené	bl_topic_key + bl_language	2 688 473	3:30:12	7,52%
KROK PŘIŘAZUJÍCÍ LOKÁLNÍ ZÁZNAM JINÉ KNIHOVNY DLE PŘEDCHOZÍCH KROKŮ						
blRestStep	cokoliv z výše uvedných	vše dosud nesloučené	biblio_linker_id	3 929 695	9:31:08	20,43%

Tabulka 3 Slučovací kroky druhé úrovně

Nelze-li seznam naplnit záznamy z téže instituce, přistupuje se k jedenáctému kroku, který se pokusí o nalezení podobných záznamů z kteréhokoli jiného zdroje přes identifikátory shluků první úrovně.

ZPŮSOB SESTAVOVÁNÍ SEZNAMU DOPORUČENÍ

Počet položek seznamu doporučení je na portálu Knihovny.cz nastaven na pět, protože na stejnou délku bylo nastaveno předchozí řešení pomocí funkce MoreLikeThis a zároveň jde o délku vyhovující rozložení stránky s náhledem úplného záznamu.

Při tvorbě seznamu podobných se jako první berou v potaz vazby nalezené krokem pro identifikaci k sobě patřících části vícedílných monografií a cyklů. Vzhledem k vysoké užitečnosti této informace pro čtenáře je umožněno zaplnit doporučeními na ostatní díly většího celku celý seznam.

U druhého kroku je povoleno vygenerovat jedno doporučení od každého typu dokumentu. Cílem je zobrazit u knihy odkaz na její zfilmovanou podobu a na audioknihu, u zvukového dokumentu příslušnou hudebninu apod.

Následujícím čtyřem krokům je umožněno obsadit buď po jednom místě seznamu doporučení, nebo mohou naplnit celý zbytek seznamu, pokud už nejsou k dispozici další typy vazeb.

Záznamům, u kterých stále ještě zbývalo volné místo do nastaveného počtu položek v seznamu, jsou dopočítávány vazby na základě vyhodnocování všech lokálních záznamů spojených na první úrovni a poté ještě záznamy napříč knihovnami (viz sedmý až devátý krok druhé úrovně). Těmito záznamy je při indexaci naplněn zbytek seznamu.

VYHODNOCENÍ VÝSTUPŮ BIBLIOLINKERU

Při vyhodnocování jakýchkoliv shluků, které vytváří Správce zdrojů, je třeba brát v potaz jejich proměnlivý charakter. Pro posouzení výstupů proto stačí uvádět jen zaokrouhlené počty.

NÁROČNOST PROCESŮ BIBLIOLINKERU

Pro účely zjištění údajů o počtech sloučených záznamů pro tento článek došlo k mimořádnému spuštění Bibliolinkeru nad všemi záznamy v portálu Knihovny .cz, kterých v tu chvíli bylo necelých 39 milionů. Slučování první a druhé úrovně trvalo v součtu téměř 70 hodin (první úroveň 22 hodin 31 minut a druhá úroveň 46 hodin 36 minut). Nutno dodat, že Bibliolinker nikdy nezpracovává všechny záznamy najednou (vyjma úplně prvního spuštění). Aby byla informace o náročnosti jednotlivých kroků relevantní i pro reálný provoz, čas je v Tabulce 2 a Tabulce 3 uveden také v procentech, nejen v absolutních hodnotách.

Větší informační hodnotu mají logy týkající se každodenních (či spíše každonočních) aktualizací. Průměrný čas vyžadovaný pro celý proces zahrnující deduplikaci a shlukování Bibliolinkeru za uplynulých 100 dní (červen až září 2023) byl 5 hodin 31 minut. Deduplikace trvala v průměru 2 hodiny, slučování první úrovně 52 minut a slučování druhé úrovně 2 hodiny 35 minut. Úkony Bibliolinkeru tedy zabírají přibližně 62 % času vynaloženého na slučování záznamů. To není málo, ale pro provoz portálu je mnohem únosnější provádět tyto výpočty v rámci rutinního zpracování dat v databázi nežli zátěž přenášet na index pod vytížením uživatelů.

VYHODNOCENÍ PRVNÍ ÚROVNĚ SLUČOVÁNÍ

Na první úrovni slučování bylo zpracováváno 19 milionů zdeduplikovaných záznamů obsahujících 38,9 milionu lokálních záznamů. Většině zdeduplikovaných záznamů zde nebyl nalezen žádný další dostatečně blízký záznam. 2,2 milionu zdeduplikovaných záznamů (obsahujících 8,5 milionu lokálních záznamů) utvořilo shluky s jinými zdeduplikovanými záznamy. Ostatní zdeduplikované záznamy neutvořily na druhé úrovni žádné další větší celky, přičemž u dokumentů, kde se vyjádření kryje s provedením (periodika, analytika, jedenkrát vydané dokumenty), je to zcela záměrné.

Mezi velikostí shluků a jejich počtem platí nepřímá úměrnost. Shluků o 2 až 10 zdeduplikovaných záznamech bylo v době zkoumání 1,9 milionu, shluků o 11–50 členech bylo 177,5 tisíc a shluků o 51 a více členech bylo už jen necelých 35,5 tisíc. Výskyt největších shluků, čítajících vyšší stovky zdeduplikovaných záznamů, je přitom ojedinělý. Tyto extrémně velké shluky obsahují i záznamy, jejichž sloučení není žádoucí, ale nelze mu bohužel zabránit. Většinou jde ale o žádoucí shluky mnohokrát vydávaných děl nejznámějších autorů (Jirásek, Mácha, Kafka, Shakespeare, Brontë, Steinback apod.).

Nevhodné vazby se týkají též menších shluků a mají stejně tak negativní dopady, ale vzhledem k vysokému podílu správně utvořených vazeb není při zkoumání náhodných vzorků tolik patrný. Podíl nežádoucích shluků však nepřekračuje únosnou mez pro daný účel. Z deseti největších shluků jsou jen dva zcela bezproblémové. V náhodném výběru deseti shluků čítajících

200–51 zdeduplikovaných záznamů jich bylo korektních sedm. Obdobný byl výsledek u dalších deseti náhodně vybraných shluků o 50–11 zdeduplikovaných záznamech (osm správných) a 10–2 zdeduplikovaných záznamech (sedm správných).

Slučování na úrovni vyjádření lze většinou realizovat pouze porovnáváním autora, názvu a typu dokumentu. Kvalita jeho výstupů tedy koresponduje s kvalitou zápisu těchto údajů v bibliografických záznamech. Problémy vyskytující se ve shlucích první úrovně lze shrnout do následujících kategorií:

Problémy přenesené z deduplikace. Došlo-li při deduplikaci ke sloučení lokálních záznamů, pak je tato vazba přenesena do Bibliolinkeru i v případě, kdy to není z uživatelského hlediska žádoucí. Typicky jde o smíšení typů dokumentů, které je sice zakázáno Bibliolinkerem, ale povoleno v některých deduplikačních krocích. Sloučení napříč typy dokumentů je na první úrovni nežádoucí, protože brání jejich vzájemnému doporučení na druhé úrovni.

Souborná díla, přívazky a vícejazyčné dokumenty. Největší shluky první úrovně byly způsobeny propojením skrze dokumenty obsahující více děl (Shakespearovy hry, Goetheho díla, party Verdiho oper či romány Boženy Němcové) nebo totéž dílo ve více jazycích (typicky zjednodušená dvojjazyčná vydání známých beletristických děl pro účely výuky jazyků). Negativním důsledkem je např. nemožnost doporučit Hamleta k Othellovi či Divou Báru k Babičce.

Edice, cykly a vícedílné monografie. Některé shluky první úrovně způsobovaly záznamy, které měly zapsán název edice nebo série jako hlavní název a název knihy z edice nebo části vícedílné monografie jako podnázev. Zčásti jde o problém přenesený z deduplikace, zčásti o důsledek sloučení skrze benevolentnější pravidla Bibliolinkeru. Komplikace způsobují komplety neboli „omnibusy“, tj. společná vydaní celých sérií typu Hra o trůny v jednom boxu, nebo díla, která byla vydávaná někdy vcelku a jindy po částech. Sloučení částí cyklů, edic a vícedílných monografií znemožní tvorbu vysoce relevantních doporučení na druhé úrovni.

Různá díla se stejným názvem. Stejnojmenné nebo takřka stejnojmenné dokumenty bez uvedeného autora nebo s týmž autorem nelze nijak rozlišit. Největší shluk první úrovně v době psaní tohoto textu tvořily ročenky a výroční zprávy různých podniků a institucí v několika jazycích, kde nebyla korporace uvedena jako autor, ale jen jako vydavatel. Podobně se chovají různé sbírky rukopisů či úředních dokumentů se stejným názvem i autorem, kde se liší jen datace, kterou Bibliolinker záměrně nebere v potaz. Vzhledem k zanedbatelnému množství těchto případů se nejedná o problém.

Řešení výše uvedených nedostatků, kvůli kterým některým záznamům nemohou být přiřazena optimální doporučení, je prakticky neuskutečnitelné jakýmikoliv úpravami Bibliolinkeru. I když ale není výstupný seznam podobných ideální, zpravidla není prázdný a obsahuje doporučení, která s prohlíženým záznamem souvisejí alespoň stejným tématem.

VYHODNOCENÍ DRUHÉ ÚROVNĚ SLUČOVÁNÍ

Pro vyhodnocení výstupů druhé úrovně slučování a jejich porovnání s původním generováním podobných pomocí funkce MoreLikeThis byl zvolen následující postup:

Každému z 11 kroků byl náhodně vybrán seznam 20 doporučení za účelem podrobnějšího zkoumání relevance výstupů jednotlivých kroků.

Pro širší zhodnocení efektivity Bibliolinkeru a porovnání s výstupy MoreLikeThis byla u takto již vybraných 220 záznamů hodnocena všechna doporučení vygenerovaná oběma metodami. Tzn. 220 × až 5 doporučení podle Bibliolinkeru (celkem 1 077) a 220 × 5 doporučení podle MoreLikeThis (celkem 1 100). Rozdíl v počtech doporučení je způsoben tím, že slučovacím pravidlům Bibliolinkeru se nepodaří pokaždé naplnit všech 5 míst seznamu doporučení, zatímco MoreLikeThis najde v indexu dostatečný počet aslepoň nějakých doporučení takřka vždy.

Z podstaty výběru zkoumaných doporučení je zřejmé, že vzorek není reprezentativní z hlediska rozložení typů dokumentů ani vygenerovaných typů doporučení. Např. jen 0,02 % bibliolinkerových doporučení je založeno na vazbě audiodokument–hudebnina, zatímco ve zkoumaném vzorku mají téměř stonásobné zastoupení. Pro účely všeobecného posouzení výstupů však tato nevyrovnanost nepředstavuje zásadní problém.

Při *doporučování částí v rámci vícedílných dokumentů a sérií* je kritická první úroveň slučování. Nevznikl-li na první úrovni nežádoucí shluk různých dílů (tzn. různé díly se nedoporučí) nebo naopak několik shluků pro stejný díl (tzn. doporučí se stejný díl), pak dojde téměř vždy k jejich správnému sloučení na druhé úrovni.

Nezáměrná, avšak nikoliv vyloženě chybná doporučení vznikají nad záznamy splňujícími podmínky pro vygenerování klíče *bl_common_title*, ačkoliv nepopisují cykly či vícedíly v předpokládaném smyslu slova (ve vzorku se ocitly např. edice zapsané jako cykly nebo noty téhož hudebního díla v různých aranžmá).

Slučování napříč typy dokumentů funguje nejlépe v případě knih a audioknih. K propojování filmů a jejich předloh nedochází v takové míře, která by byla optimální, a to dílem kvůli problematickému určování typu dokumentu, dílem kvůli neuvedení autora předlohy se správným kódem role a dílem kvůli titulu filmu odlišnému od titulu knižní předlohy. V případě příznivé souhry všech předpokladů se filmy a předlohy skutečně navzájem doporučují. Problematika odlišných titulů se projevuje ještě výrazněji u hudebních děl, kde se přiřadí hudebnina k interpretaci jen velmi zřídka (hudebnina s titulem „Dobře temperovaný klavír" se nepropojí s albem „The Well Tempered Clavier" apod.) Problémem jsou tedy spíše neutvořená spojení, nežli špatně utvořená spojení.

Doporučování podle edice funguje po technické stránce velmi dobře, ale souvislost mezi záznamy nemusí být uživateli vždy zřejmá u velmi široce pojatých edicí. Ve vzorku se vyskytovaly např. dokumenty z kolekce televizních seriálů od pohádek až po detektivky, edice různorodé beletrie s obálkami bez jednotné grafické úpravy, sada učebnic z různých vyučovacích předmětů nebo řada audioknih zcela nesouvisejících žánrů. Kdyby měl být tento problém nějak ošetřován, bylo by zapotřebí explicitně vylučovat konkrétní edice nebo zakazovat slučování konkrétních žánrů (učebnic). Zjevná neefektivita tohoto postupu však zpochybňuje význam takového řešení.

Téměř 60 % podobností vygenerují kroky založené na porovnávání klíče *bl_entity*, tzn. *krok propojující záznamy se stejnými osobami, autoritami či akcemi ve jmenném a věcném popisu*. Tyto kroky také generují největší podíl uživatelsky nerelevantních doporučení. Kromě vazeb utvořených přes různé vedlejší autory sem spadají i doporučení děl jmenovců nebo děl, u kterých propojení přes autora není žádoucí (např. novinové články). Tato nevýhoda používání neautoritních údajů však nepřevažuje nad výhodou možnosti nalézt doporučení i k méně kvalitním záznamům.

Kroky slučující podle tématu naopak pracují jen s kontrolovanými slovníky, ačkoliv by zde rozšíření na neautorizované problémy mělo potenciál generovat relevantní doporučení. Volně tvořená klíčová slova však Bibliolinker nezpracovává z důvodu zachování únosné výpočetní náročnosti a také kvůli velkému množství velmi obecných a jinak nepoužitelných údajů (např. pole 653 se souslovími rozdělenými do samostatných podpolí po jednom slově). Nepříliš užitečné vazby vznikají i v současném nastavení, přičemž se jako nejvíce patrné jeví doporučení podle geografických autorit (vzpomínky rodáků propojené se sborníkem o lokálním průmyslu apod.)

POROVNÁNÍ VÝSTUPŮ BIBLIOLINKERU S FUNKCÍ MORELIKETHIS

Specifikace indexového požadavku MoreLikeThis, kterým byla doporučení generována před zavedením Bibliolinkeru, je rozepsána v Tabulce 4. Jak je patrné, zabudovaný nástroj Solru řeší tutéž funkčnost mnohem jednodušeji než Bibliolinker – jedná se o klasický vyhledávací dotaz s nastavitelnou relevancí (boostem) vybraných polí indexu. K náhradě využívání funkce MoreLikeThis bylo přikročeno kvůli snížení zátěže indexu. Tento cíl Bibliolinker splnil a kvalita jeho výstupů se jevila jako dostačující k produkčnímu nasazení. Pro účely tohoto článku bylo provedeno porovnání výstupů obou metod na výše popsaném vzorku doporučení.

pole indexu	boost	podpole MARC 21
subject_facet_mv	300	600$abcdfglnpqstyz, 610$abcdfgklnpstyz, 611$abcdefgklnpqstyz, 630$adfgklnpstyz2, 648$a, 650$avyz, 651$avxyz, 964$abcdefg, 967$ab
conspectus_facet_mv	100	072$x (když 072$2 obsahuje "Konspekt")
author	40	100$abcd
author2	10	110$ab, 111$ab, 700$abcd, 700$ab, 711$ab
title		245$abnp

Tabulka 4 Nastavení funkce MoreLikeThis

U všech 220 vybraných záznamů proběhlo zběžné zkoumání smysluplné souvislosti s jejich 1 077 bibliolinkerovými a 1 100 solrovými doporučeními (vizuální kontrola základních bibliografických údajů). Na základě výsledku získaného takovou metodou na malém nereprezentativním vzorku sice nelze vynášet přesné soudy o efektivitě jednoho či druhého postupu, ale pro jejich přibližné porovnávání je to dostačující. Bibliolinker i MoreLikeThis si zde vedly velmi podobně – míra zjevných podobností se u obou pohybovala kolem 85 %. Největší podíl (přibližně třetinu) doporučení z toho tvoří díla se společným autorem.

Hlavní výhodou Bibliolinkeru po obsahové stránce je schopnost vygenerovat záznamu doporučení pokud možno ze stejné knihovny. U vybraného vzorku bylo takových doporučení 91,2 %. MoreLikeThis funguje nad zdeduplikovanými, nikoliv lokálními záznamy. Pro porovnání tedy bylo nutné počítat doporučení takových zdeduplikovaných záznamů, které obsahovaly lokální záznam z téže knihovny jako záznam ze vzorku. Takových doporučení zde bylo 64,4 %. Lze tedy říci, že Bibliolinker dosahuje v tomto ohledu znatelně lepších výsledků.

Další silnou stránku Bibliolinkeru je mizivý výskyt odkazů na různá, či dokonce stejná vydání téhož díla, ať už jde přímo o shodu díla popisovaného prohlíženým záznamem s doporučením nebo o opakující se doporučení v jednom seznamu. Zatímco MoreLikeThis neobsahuje žádnou zábranu proti doporučování různých provedení téhož vyjádření, v Bibliolinkeru k tomuto účelu slouží první úroveň slučování. MoreLikeThis vygeneroval shodná či opakovaná doporučení u 28 záznamů z 220, zatímco u Bibliolinkeru se tak stalo dvakrát (jednou kvůli nezdeduplikovanému videodokumentu a podruhé kvůli nežádoucímu zdeduplikování knihy s audioknihou).

MoreLikeThis na druhou stranu překonává Bibliolinker v lepší kontrole nad relevancí doporučení. Pokud nedojde k výše popsanému prolinkování na stejné dílo, doporučení mají skutečně nejvyšší možnou relevanci k právě prohlíženému záznamu. V Bibliolinkeru sice kroky druhé úrovně naplňují seznam doporučení podle předpokládané relevance typu podobnosti, ale jakmile dojde k nalezení předdefinovaného počtu vazeb na krok, vyhodnocování je zastaveno, přestože na potenciálně nejzajímavější záznamy nemusela přijít řada. Na vysoké relevanci MoreLikeThis doporučení má značný podíl vyhodnocování indexového pole pro fasetu tématu, které zahrnuje téměř celý věcný popis. Výstupům Bibliolinkeru by zpracovávání širšího rozsahu věcných údajů bezpochyby prospělo, ale bohužel by způsobilo nepřijatelné zvýšení výpočetní náročnosti.

DISKUZE

Na portálu Knihovny.cz plní Bibliolinker svou roli dobře a obstojí v porovnání s obdobnými funkcemi používanými v knihovních katalozích. Doporučení většinou dávají smysl a uživatel zřídkakdy narazí na záznam bez seznamu podobných. Za zmínku stojí i využití druhé úrovně Bibliolinkeru v adresáři knihoven, kde se v seznamu doporučení zobrazují knihovny stejného typu a knihovny v okolí dané instituce. Zvažovány byly i další možnosti využití Bibliolinkeru.

V souvislosti se zpřístupňováním děl nedostupných na trhu (DNNT) byl podniknut pokus o aplikaci výstupů první úrovně slučování Bibliolinkeru při návrhu děl vhodných ke zveřejnění v digitalizované podobě. Vzhledem k citlivosti problematiky autorsky chráněných děl se však došlo k závěru, že rozhodování o zveřejnění musí být předmětem ručního kurátorství a není vhodné je svěřit automatickému procesu.

Uživatelsky zajímavou funkcí by mohlo být doporučování e-knih k tištěným knihám a Bibliolinker by mohl být prostředkem k její realizaci. Záznamy e-knih nebyly v době vzniku Bibliolinkeru dodávány na Knihovny .cz, ale nyní již s nimi portál pracuje. V případě implementace doporučování e-knih by bylo nutné zabývat se otázkou spolehlivosti vytvořených vazeb. Přísně nastavená pravidla by mohla zabránit přirazení e-knihy k jakémukoliv tištěnému dokumentu včetně stejného vydání od téhož vydavatele. Příliš benevolentní pravidla by způsobovala nejednoznačná a nesprávná propojení. Volba vhodného způsobu implementace doporučování e-knih bude předmětem dalšího zkoumání.

První úroveň slučování nabízí vhled do teoretické možnosti FRBRizace (Lichtenbergerová, 2007) na portálu Knihovny.cz. Prezentace dokumentů na úrovni vyjádření či díla byla předmětem diskuzí mezi experty stojícími u vzniku celého projektu. Pro účely doporučování nepředstavuje 20–25% podíl nežádoucích vazeb zásadní problém, protože tyto shluky první úrovně jsou skryty před zraky uživatelů. Pokud by se však takto nespolehlivě určená vyjádření měla stát základem pro vyhledávání a zobrazování záznamů na portálu, uživatel by byl zmaten a odrazen.

Častá neúspěšnost identifikace vyjádření Bibliolinkerem se do značné míry odvíjí od skutečnosti, že portál

Knihovny.cz je agregátor mnoha zdrojů s různorodou kvalitou metadat. Lze předpokládat, že nasazení téhož algoritmu nad záznamy z jedné knihovny by vedlo ke spolehlivějším výsledkům. Ty by již bylo možné použít v rámci FRBRizace katalogu, ať už by šlo o plnohodnotné shlukování na vyšších FRBR úrovních nebo jen o prezentaci souvisejících provedení v náhledu bibliografického záznamu. Menší objem zpracovávaných dat by dával prostor pro poloautomatické ošetření výjimek typu přívazků, souborných děl apod. Bibliolinker je však součástí nástroje, který je pro takto omezený účel zbytečně robustní.

V době prudkého rozmachu umělé inteligence se lze zamýšlet nad možností zapojit do procesu doporučování podobných dokumentů strojové učení. Knihovny.cz identifikují podobnost pomocí klasického programování, protože v době prvotního návrhu Bibliolinkeru před několika lety se jevilo jako nejsnadnější navázat na nástroje, které již byly vyvinuty na míru portálu. Vzhledem k obstojné funkčnosti stávajícího řešení nebyly podniknuty kroky směřující k nasazení strojového učení. V případě budoucího experimentování s umělou inteligencí by výstupy Bibliolinkeru mohly sloužit jako tréninková data.

Nový rozměr do problematiky doporučovacích nástrojů může vznést postupné zavádění principů Linked Data. Budou-li propojená data dostatečně bohatá a konzistentní, vazby mezi nimi budou již implicitně přítomny a k jejich identifikaci již nebude zapotřebí složitých algoritmů. V dohledné době však nelze očekávat zásadní proměnu charakteru bibliografických metadat tímto směrem.

ZÁVĚR

Tento článek se zabýval popisem implementace softwarového nástroje Bibliolinker určeného ke slučování bibliografických záznamů dle podobnosti za účelem doporučování podobných dokumentů uživatelům portálu Knihovny.cz. Bibliolinker nahrazuje a do značné míry překonává původní řešení téže funkce, které spočívalo v dohledávání podobných v reálném čase. Zavedením Bibliolinkeru s předindexovanými seznamy podobných došlo nejen ke snížení zátěže indexu, ale v určitých ohledech také ke zlepšení kontroly nad doporučovanými záznamy.

K vytvoření doporučení podobných, ale nikoliv stejných dokumentů používá Bibliolinker dvouúrovňovou hierarchii slučování záznamů. První, vylučovací úroveň slučuje záznamy přibližně na úrovni vyjádření a slouží jako zábrana proti doporučování různých vydání stejných děl. Druhá úroveň propojuje záznamy na základě společných hodnot ve vybraných polích jmenného i věcného popisu. Seznam doporučení je sestavován z podobnostních vazeb různého typu (série, různá vyjádření téhož díla, společná edice či tematika). Při doporučování jsou preferovány vazby na záznamy ze stejné knihovny. Naopak jsou zakázány vazby mezi záznamy, které byly propojeny na první úrovni.

Ačkoliv se jedná o software vyvinutý na míru portálu Knihovny.cz, z poznatků získaných při jeho vývoji lze odhadovat proveditelnost FRBRizace nad bibliografickými záznamy v katalozích a discovery systémech v prostředí českých knihoven. Spolehlivost vazeb generovaných první úrovní Bibliolinkeru je dostačující pro účely doporučování podobných záznamů, ale nikoliv pro přímou prezentaci vyšších FRBR úrovní v uživatelském rozhraní. Limit zde představuje především kvalita metadat a konzistentní přístup k jejich vytváření (např. u přívazků a společných vydání více děl). Dalším faktorem je pak reálně nerozhodnutelná příslušnost ke konkrétnímu vyjádření či dílu v případě některých dokumentů (např. různým způsobem vydávané série/vícedílné monografie, rozšiřovaná a doplňovaná díla atd.) O proveditelnosti FRBRizace by se dalo uvažovat nejvýše v prostředí jednotlivých knihoven, které mají buďto velmi kvalitní záznamy nebo objem dat tak omezený, že u něj lze zajistit průběžnou ruční korekci sporných případů. Prostředí agregátoru velkého množství různorodých metadatových zdrojů se pro FRBRizaci jeví jako nevhodné.

Vyvinutím Bibliolinkeru se týmu Knihovny.cz podařilo dosáhnout požadovaného cíle vytvářet index nezatěžující doporučování podobných dokumentů v náhledu katalogizačního záznamu. Zobrazování obsahově relevantního seznamu podobných zvyšuje uživatelskou přívětivost portálu Knihovny.cz a přibližuje grafické rozhraní zažitým zvyklostem z nejpoužívanějších webových aplikací typu e-shop, sociální síť apod.

DEDIKACE

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

POUŽITÁ LITERATURA

Apache Solr Reference Guide: MoreLikeThis, 2021. *Apache Solr Reference Guide* [online]. Apache Software Foundation [cit. 2023-10-12]. Dostupné z: https://solr.apache.org/guide/8_11/morelikethis.html

Funkční požadavky na bibliografické záznamy [online]. Přeložila Mgr. Ludmila Celbová. IFLA, NK ČR, 2001 [cit. 2023-10-12]. Dostupné z: https://archive.ifla.org/VII/s13/frbr/frbr-cs.pdf

CPK - Využití sémantických technologií pro zpřístupnění kulturního dědictví prostřednictvím Centrálního portálu knihoven. *Informační systém nad daty VAVAI Technologické agentury ČR* [online]. TA ČR, b.r. [cit. 2023-10-12]. Dostupné z: https://starfos.tacr.cz/cs/project/DG16P02R006

Knihovny.cz: Zapojené knihovny a zdroje. *Knihovny.cz* [online]. Brno: MZK [cit. 2023-10-12]. Dostupné z: https://www.knihovny.cz/Content/zapojene-knihovny-a-zdroje

Knihovny.cz: Základní informace. *Knihovny.cz* [online]. Brno: MZK [cit. 2023-10-12]. Dostupné z: https://www.knihovny.cz/Content/o-portalu

KURFÜRSTOVÁ, Jana, Petra ŽABIČKOVÁ a Tomáš CEJPEK. Deduplikace na portálu Knihovny.cz. In: *Knihovna: knihovnická revue* [online]. 2023, 34(1), 48-62 [cit. 2023-09-20]. ISSN 1801-3252. Dostupné z: https://knihovnarevue.nkp.cz/archiv/2023-1/recenzovane-prispevky/deduplikace-na-portalu

LICHTENBERGOVÁ, Edita. Funkční požadavky na bibliografické záznamy (FRBR) v praxi. In: *Knihovny současnosti 2007* [online]. Brno: Sdružení knihoven ČR, 2007, s. 355-367. ISBN 978-80-86249-44-5. Dostupné z: http://www.sdruk.cz/data/xinha/sdruk/2007-3-355.pdf

O projektu Obálkyknih.cz. *Obálky knih* [online]. MZK a ČBVK [cit. 2023-10-12]. Dostupné z: https://obalkyknih.cz/about

Record Manager 2. *GitHub* [online]. Brno: MZK [cit. 2022-12-08]. Dostupné z: https://github.com/moravianlibrary/RecordManager2

VuFind: Search. Discover. Share [online], Ithaca. Villanova University's Falvey Memorial Library [cit. 2023-10-12]. Dostupné z: https://vufind.org/vufind/

■ Článok bol recenzovaný