VEŘEJNÉ INFORMAČNÍ SLUŽBY KNIHOVEN

Koordinační centrum programu a implementace Koncepce rozvoje knihoven v České republice

Obálky knih.cz - rozvoj projektu v roce 2019

Jihočeská vědecká knihovna v Českých Budějovicích
leden 2020

Příloha č. 4

Zhodnocení projektu

Projekt Obálkyknih.cz sdružuje různé zdroje informací o dokumentech do jedné, snadno použitelné webové služby. Databáze aktuálně obsahuje přes **2,08 miliónu obálek** (nárůst za rok 2019 o cca 200 000 obálek), **510 tisíc obsahů** českých a zahraničních dokumentů (nárůst za rok 2019 o cca 60 000 obsahů) a **6 500** seznamů doporučené literatury. Dále poskytuje přes **481 tisíc anotací, 3.65 miliónu hodnocení u 186 tisíc titulů, 8,5 tisíc komentářů, 56 tisíc fotografií autorit** a cca. **0,6 miliónu vygenerovaných citací** dle normy ISO 690**.** API služby projektu využívá většina knihoven v České republice, muzea, archivy, oborové projekty, CPK, aj.

Správcem projektu Obálkyknih.cz je Jihočeská vědecká knihovna v Českých Budějovicích a projekt provozuje ve spolupráci s Moravskou zemskou knihovnou v Brně. JVK i MZK do projektu z vlastních zdrojů vkládají nemalé lidské a finanční zdroje.

Přehled vlastností projektu:

- hlavní servery jsou provozovány v **Jihočeské vědecké knihovně v Českých Budějovicích**, záložní server je v **Moravské zemské knihovně v Brně**
- v případě výpadku jednoho ze serverů mají knihovní systémy možnost přejít během několika vteřin na záložní stroj bez ztráty dostupnosti služeb pro své čtenáře
- měsíčně hlavní server odbaví průměrně **50 miliónů** požadavků, cca. **2 milióny denně**, průměrně **20 dotazů za vteřinu**
- ve špičkách (9-15:00) odbavují servery **40-80 požadavků za vteřinu**
- denně do databáze je nově nahráno nebo upraveno průměrně **500 dokumentů**
- denní přírůstek dat činí **8 GB**, z nich se následně generují náhledy obálek v různých rozlišeních, PDF dokumenty s obsahy a rozpoznává se text pomocí OCR
- **20 Mbit za vteřinu** je datový tok ven ze serveru a na server což představuje cca. 85% všech dat, které projdou internetovým připojením JVK

Detailní statistiky exportu dokumentů přes skenovacího klienta za období leden - prosinec 2019:

Počet odeslaných dokumentů	**116 413**
Počet uložených obálek (COVER)	**104 250**
Počet uložených stran obsahů (TOC)	**146 067**
Počet uložených fotografií autorit	**1 642**
Počet uložených stran seznamů lit.	**24 265**

Počty odeslaných stran a titulů přes skenovacího klienta dle jednotlivých knihoven (rok 2019):

STRAN	TITULŮ	SIGLA	NÁZEV
46460	22298	CBA001	Jihočeská vědecká knihovna v Českých Budějovicích
26776	10858	OLD012	Knihovna Univerzity Palackého v Olomouci
17907	7644	ABA004	Slovanská knihovna
17306	8298	BOA001	Moravská zemská knihovna
15451	3541	ABA013	Národní technická knihovna
14340	9550	ABA001	Národní knihovna ČR
12999	8312	OLA001	Vědecká knihovna v Olomouci
10351	4256	ABA008	Národní lékařská knihovna
10087	2254	BOD010	Masarykova univerzita - Právnická fakulta
6206	6206	ABD001	Knihovna Ústavu Dálného východu Filozofické fakulty Univerzity Karlovy
5696	1775	CBD005	Teologická fakulta JCU
4620	1764	BOE020	Knihovna Ústavního soudu
4013	967	ZLD002	Univerzita Tomáše Bati ve Zlíně

3959	1035	ABA006	Vysoká škola ekonomická v Praze
3936	1851	CBD007	Akademická knihovna Jihočeské univerzity
3778	3064	ULG001	Severočeská vědecká knihovna v Ústí nad Labem
3768	1130	BOD001	Ústřední knihovna filozofické fakulty MU
3747	2960	HKA001	Studijní a vědecká knihovna v Hradci Králové
3161	1661	PNA001	Studijní a vědecká knihovna Plzeňského kraje
3021	1490	LIA001	Krajská vědecká knihovna v Liberci
2992	963	ABA007	Knihovna Akademie věd
2894	1097	ABB019	Knihovna Sociologického ústavu AV ČR
2501	799	BOD031	Masarykova univerzita, Fakulta sociálních studií, Ústřední knihovna
2483	1349	UHG001	Knihovna Bedřicha Beneše Buchlovana
2206	1579	BOD003	Ústřední knihovna Pedagogické fakulty MU
2110	1694	PAG001	Krajská knihovna v Pardubicích
1968	404	BOD004	Ústřední knihovna Přírodovědecké fakulty MU
1687	660	BOE451	Knihovna Biskupství brněnského
1495	425	ABD103	Univerzita Karlova-Fakulta sociálních věd-Středisko vědeckých informací
1461	969	HBG001	Krajská knihovna Vysočiny
964	483	KVG001	Krajská Knihovna Karlovy Vary
916	469	KLG001	Středočeská vědecká knihovna v Kladně
812	290	ABD100	ÚK ČVUT
794	242	BOD022	Středisko vědeckých informací ESF MU
683	689	TAG001	Městská knihovna Tábor
621	386	OSA001	Moravskoslezská vědecká knihovna v Ostravě
572	404	SMG506	Městská knihovna Antonína Marka Turnov
540	220	ULE301	Muzeum města Ústí nad Labem
465	461	LID001	Knihovna Technické univerzity v Liberci
381	117	ABA011	Parlamentní knihovna
345	118	ABB022	Středisko vědeckých informací Fyziologického ústavu AV ČR
305	103	BOD018	Masarykova univerzita - Fakulta informatiky
268	136	BOE303	Knihovna Moravské galerie v Brně
204	116	ULD001	Ústřední knihovna UJEP
195	122	ABA100	Všenorská knihovna a informační centrum Berounka
194	120	PNG001	Knihovna města Plzně
180	63	BOD006	Informační centrum, ústřední knihovna Mendelovy univerzity v Brně
147	147	ABD027	Evangelická teologická knihovna UK
129	74	ZLG001	Krajská knihovna Františka Bartoše ve Zlíně
120	96	SOG504	Městská knihovna Chodov
84	26	ABG312	Knihovna Jabok
79	32	ABD009	Knihovna MFF - Matematické oddělení
63	26	JID501	Knihovna Univerzitního centra Telč Masarykovy univerzity
48	16	ABD010	Knihovna MFF UK - půjčovna skript a učebnic
25	16	UOE802	Knihovna Regionálního muzea ve Vysokém Mýtě
19	10	ABB001	Knihovna Archeologického ústavu AV ČR, Praha
15	15	NAE951	Knihovna (Římskokatolická farnost – děkanství Nové Město nad Metují)
13	3	ABD170	Univerzita Karlova - Matematicko-fyzikální fakulta

10	10	CKG001	Měk v Českém Krumlově
8	4	BOD033	Univerzitní knihovna pro studenty se specifickými nároky MU
4	2	OSD002	Ústřední knihovna VŠB - TU Ostrava
1	1	PTG001	Městská knihovna Prachatice
1	1	ABE195	Knihovna Masarykovy demokratické akademie

Úkoly řešené v rámci projektu v roce 2019:

DOPORUČOVÁNÍ LITERATURY

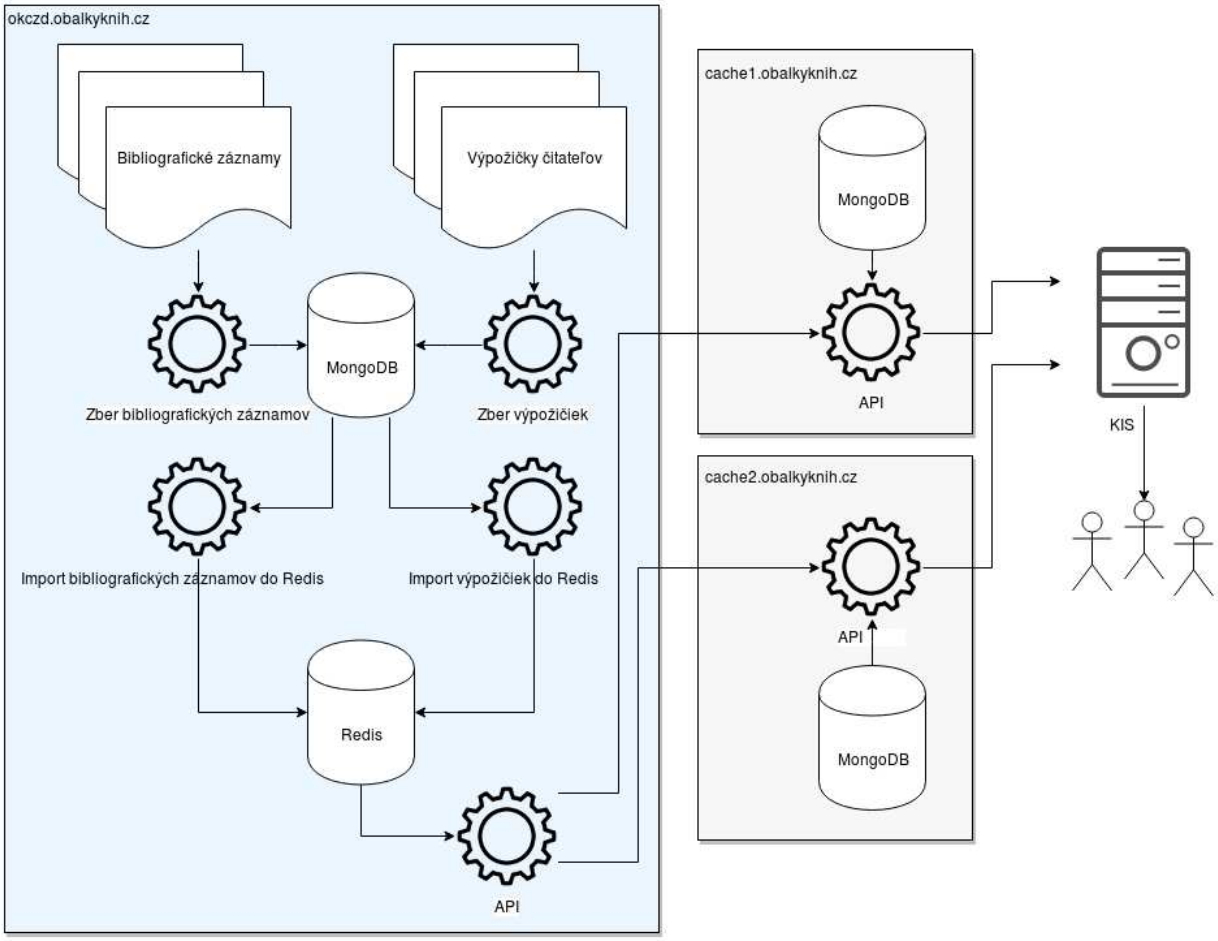
Doporučování literatury je subsystém obálek knih, který poskytuje API knižním informačním systémům s cílem obohacení o návrhy dalších podobných titulů čtenářům k četbě a to:

- doporučování na základě titulu – je nejpoužívanější způsob použití API, pro nejširší množinu čtenářů a knihoven. Ideální pro obohacení katalogu každé z knihoven, pro zatraktivnění katalogu čtenářům s minimálním dopadem na zavádění nových funkcí do stávajících katalogů.
- doporučování na základě čtenáře – kde cílem je zpřesnění doporučování pro konkrétního čtenáře na základě znalosti výpůjček. Ideální pro obohacení katalogů knihoven, které chtějí poskytovat lepší služby přihlášeným čtenářům v jejich čtenářském kontě.
- doporučování na základě preference – kde cílem je poskytnutí doporučování literatury k četbě přihlášeným čtenářům v jejich čtenářském on-line kontě i bez historie výpůjček (novým čtenářům), nebo čtenářů, kteří nechtějí poskytovat svoji historii výpůjček.

Výhodou doporučování literatury obálek knih je existence hybridního systému doporučování, které kombinuje content-based a kolaborativní doporučování. V roce 2019 byly vykonány úpravy algoritmů pro vylepšení API doporučování literatury a bylo harvestováno a pro účely kolaborativního doporučování už sklizeno cca 40 mil. anonymizovaných výpůjček spolupracujících knihoven.

Architektura subsystému doporučování literatury
V roce 2019 proběhla úprava a zdokonalení částí subsystému a to:

- Skript pro sběr bibliografických záznamů – automatické plnění plných bibliografických záznamů jako zdroje dat pro content-based doporučování
- Skript pro sběr výpůjček spolupracujících knihoven – automatické plnění anonymizovaných dat výpůjček sloužících ke kolaborativnímu doporučování. Byl doplněn sběr dat nejen harvestováním datasetů anonymizovaných výpůjček pomocí OAI, ale i importem dat jiného formátu dat z KIS způsobem týdenních inkrementů anonymizovaných výpůjček.
- Indexační skripty byly obohaceny o indexaci dalších MARC 21 tagů, zejména kvůli doporučování na základě preference.
- Byla upgradována verze databázových služeb Mongo DB a Redis DB.
- Došlo k úpravě a optimalizaci skriptu pro kolaborativní doporučování, které bylo potřebné kvůli nárůstu záznamů výpůjček více než dvou násobně. Cílem úpravy bylo použití větší množiny dat výpůjček. Dále cílem optimalizace bylo snížení doby odezvy API.



OBR. 1: Architektura subsystému doporučování literatury

Doporučování podle preference

Cílem tohoto typu doporučování je poskytnutí doporučení literatury čtenářům, kteří nemají historii výpůjček, nebo nechtějí svou historii poskytnout. Využívá se volby preference čtenáře obdobně jako u předmětového vyhledávání, tj. podle kategorie konspektu, skupiny konspektu, nebo jejich kombinací. Do API byly doplněny parametry (viz. další popis).

Dalším případem použití doporučování podle preference je v kombinaci s konkrétním čtenářem. V takovém případu dojde k filtraci titulů už přečtených daným čtenářem.
Skript pro sběr bibliografických záznamů do RedisDB byl upraven, aby došlo i k importu tagu 072.

Postup pro doporučování podle předmětového vyhledávání:

1. Vyhledání všech titulů, které obsahují kategorii, nebo skupinu zadané na vstupu.
2. Pro vyhledané tituly se zjistí počet výpůjček.
3. Tituly se seřadí podle počtu výpůjček.
4. Na výstup se dostane 50 titulů s nejvíce výpůjčkami.

Postup pro doporučování podle předmětového vyhledávání pro konkrétního čtenáře:

1. Vyhledání všech titulů, které obsahují kategorii, nebo skupinu zadané na vstupu.
2. Filtrace titulů, které už byly daným čtenářem vypůjčené.
3. Pro vyhledané tituly se zjistí počet výpůjček.
4. Tituly se seřadí podle počtu výpůjček.
5. Na výstup se dostane 50 titulů s nejvíce výpůjčkami.

Rozšíření content-based doporučování o vytěžování anotací
Cílem bylo obohacení content-based doporučování o informaci získanou z anotací dokumentů. Hlavním použitím je seřazení doporučených dokumentů podle podobnosti jednotlivých anotací dokumentů. Dosud se seřazovalo podle podobnosti tagů bibliografických záznamů, kdy každý tag má přidělenou prioritu/důležitost. Na určení podobnosti anotací se používá kombinace TF-IDF vektorů společně s kosinovou podobností.

Změny v MongoDB:
Pro ukládání anotací byla v MongoDB vytvořená nová kolekce „anotation“. Anotace se získávají pomocí API backend vrstvy obálek knih.

Popis řešení:
Na vstupu jsou dva seznamy dokumentů:

1. Seznam dokumentů, pro které se hledají podobné dokumenty.
2. Seznam dokumentů, z kterých se hledají podobné dokumenty.

Pro druhý seznam se určí pořadí podle podobnosti:

1. Pro každý dokument z obou seznamů se získá anotace z MongoDB.
2. Vypočte se TF-IDF vektor pro všechny dokumenty z obou seznamů.
3. Pro každý dokument ze seznamu 2:
    a. Vypočte se kosinová podobnost pro každý dokument ze seznamu č.1
    b. Vypočte se průměr těchto podobností.
4. Dokumenty se seřadí podle kosinové podobnosti.

Na výstupu je ve výsledku seřazený seznam podle kosinové podobnosti anotací. Pro výpočet TF-IDF je vytvořen objekt, který byl natrénován na velkém počtu anotací. TF-IDF vektor je dlouhý podle před definovaného slovníku vytvořeného ze seznamu klíčových slov dokumentů.

Rozšíření API
Vstupní parametry pro doporučování podle předmětového vyhledávání:

- kons - kategorie, nebo skupina konspektu. Kategorie musí začínat písmenem 'k'. V případě vícerých hodnot JSON pole.

Dotaz API pro doporučování podle předmětového vyhledávání:

https://cache.obalkyknih.cz/api/doporuc?multi={"kons":["527","78","k2"]}

Vstupní parametry pro doporučování podle předmětového vyhledávání pro konkrétního čtenáře:

- kons - kategorie, nebo skupiny konspektu. Kategorie musí začínat písmenem 'k'. V případě vícerých hodnot JSON pole.
- sigla - Sigla knihovny.
- user - Anonymizovaný identifikátor čtenáře.

Dotaz API pro doporučování podle předmětového vyhledávání pro konkrétního čtenáře:

https://cache.obalkyknih.cz/api/doporuc?multi={"sigla":"CBA001","user":"6mapWQXT0h1RQxB/T4iqzI7cdt3AQzUYcNW7HO2f6gnu6vQ016SUGHxyWFlNubGLVpPM80/gRzsmvWXllgo85g==","kons":["527","78","k2"]}

MONITORING SYSTÉMU

Cílem úkolu je sledování, logování a průběžné vyhodnocování stavu komponent celého systému obálek knih včetně možnosti sledování kvality poskytovaných dat. S nárůstem dostupných služeb projektu a při omezeném personálním obsazení je již velmi složité kontrolovat a monitorovat, zda všechny služby a funkce

pracují dle specifikace a prováděné upgrady funkcionality nezpůsobují výpadky a nedostupnost služeb. Kontroly je potřeba provádět automatizovaně a periodicky – min. 1x za týden (změny fontend i backend vrstvy probíhají z důvodu oprav chyb nebo nového vývoje průběžně po celý rok), nebo i častěji v případě kritických subsystémů obálek knih.

Monitoring systému sleduje komponenty systému:

1. pravidelné úkoly skriptů harvesterů, crawlerů a dalších importních skriptů - TOC OCR a jiných,
2. on demand úkoly volání importního API skenovacím klientem, webovým rozhraním
3. sledování metrik OS a stavu souborového systému, jakým je kontrola korektních nastavení ACL důležitých adresářů pro běh projektu, stavu běhu databází, apod.
4. kontrola kvality dat pomocí testovací množiny.

Monitoring systému disponuje webovým rozhraním, které umožňuje:

1. Přehled kontrolovaných částí systémů s možností plánování času a četnosti spouštění kontrol.
2. Prohlížení protokolu, kde je možné dohledat stav kontroly komponenty systému ke konkrétnímu dni v minulosti, včetně protokolu vytvořeného monitorovacím skriptem daného modulu.

Součástí monitorovacího systému je také subsystém pro aktivní varování na odhalený chybový stav. Varování jsou posílány emailem a to podle priority dané úlohy. V případě prioritní kontroly jsou odeslány okamžitě. V případě méně prioritní kontroly jsou posílány v ranních pracovních hodinách. Vždy se ale jedná o souborné hlášení nalezených problémů včetně logu vytvořeného kontrolním modulem (hlášení obsahuje stejné informace jako webové rozhraní).

KONTROLY A DALŠÍ PRÁCE

V roce 2019 byly prováděny úkoly související s kontrolou a údržbou dat. Jednalo se zejména o kontroly anotací a hodnocení, propojení na další vydání titulu, propojení vazeb e-knih a opravy a kontroly záznamů dle báze NKC NK Praha. Jedná se o práce, které nelze plně automatizovat a je nutné je provádět pomocí zaškolené obsluhy.
Výsledkem práce bylo:

1) napojení e-knih na klasické papírové tituly – aktuálně je dostupných 1766 e-knih z produkce Městské knihovny v Praze
2) propojení na další vydání titulu - propojení jednotlivých vydání shodného titulu nebo navazující série titulů – řešeno ve spolupráci s CPK
3) opravy a schvalování anotací k dokumentům – za rok 2019 přibylo v projektu 47 tisíc anotací, které byly formálně zkontrolovány a doplněny k titulům
4) kontroly dat projektu s bázemi NKP Praha - historicky projekt obálkyknih.cz získal data (obálky, obsahy, anotace,…) z různých zdrojů s různou kvalitou zpracování (popisné údaje, identifikátory). V databázi se tak vyskytují vícečetné záznamy shodného titulu, každý uložen pod jiným identifikátorem. Cílem úlohy bylo tyto záznamy spojit a sloučit dostupná data (obálka, obsah, anotace, hodnocení, komentáře). Opravou dat a doplněním identifikátorů budou obálky dostupné více knihovnám, které doposud nebyly schopny obálku a ostatní data zobrazit – nebylo je jak propojit.

Další úkoly řešené v roce 2019 mimo projekt:

- údržba a podpora skenovacího klienta pro nahrávání dat knihovnami do projektu
- kontrola skenovaných periodik a opravy nalezených problémů, metodické vedení přispěvatelů
- upgrade použitých SW na serverech za účelem větší funkcionality a vyšší stability běhu služby
- do veřejného API poskytovaného servery cache1.obalkyknih.cz a cache2.obalkyknih.cz byla přidána možnost ověření pomocí wildcard subdomény 3.úrovně. - registrace se uloží standardně na webu www.obalkyknih.cz po přihlášení (je možné vyplnit *.domena-knihovny.cz), API vyhodnotí

hvězdičku a pokud se domena-knihovna.cz shoduje s doménou v referal, bude request považován jako ověřený.

- upgrade SSL webových serverů projektů „dle požadavků a standardů Google“
- obnova a upgrade SSL certifikátů serverů - z projektu Let’s Encrypt
- aktualizace webových stránek projektu
- úprava a rozšíření servisních stánek projektu https://servis.obalkyknih.cz/
- doprogramování možnosti přidání anotace k titulu přes webové rozhraní
- kontrola a opravy položek databáze obsahů dokumentů (doplnění chybějících OCR)
- prezentace projektu mezi odbornou i laickou veřejností - 18. celostátní archivní konference Plzeň, 23.–25. dubna 2019, Setkání uživatelů ARL - květen 2019, 8th Colloquium of Library and Information Experts of the V4+ Countries Bratislava, 17th – 19th June 2019, Knihovny současnosti 2019, 12. výroční seminář SK ČR 2019, prezentace pro regionální knihovny, …
- aktualizace metodických pokynů a manuálu pro knihovny a knihovní systémy
- emailová a telefonická podpora projektu, spolupráce s tvůrci KIS, CPK

Popis řešení a veškeré kódy aplikace jsou volně dostupné jako opensource na adrese https://github.com/cbvk/obalkyknih/wiki.

V Českých Budějovicích 9. 1. 2020

Ing. Jiří Nechvátal
Jihočeská vědecká knihovna v Českých Budějovicích