VISK 3

Informační centra veřejných knihoven - ICEKNI

Obálky knih.cz - rozvoj projektu v roce 2017

Jihočeská vědecká knihovna v Českých Budějovicích

2017

Zhodnocení projektu

Projekt ObalkyKnih.cz sdružuje různé zdroje informací o dokumentech do jedné, snadno použitelné webové služby. V současnosti pro knihovní a jiné katalogy poskytujeme:

- **Obálky knih a periodik** - ale také obálky speciálních dokumentů - map, hudebnin, CD, DVD, ..., aktuálně přes **1,57 miliónu obálek**, nárůst za rok 2017 o cca. **310 tisíc** obálek
- **Obsahy knih a periodik** - naskenované obsahy zpřístupněné v podobě PDF souborů, aktuálně přes **315 tisíc obsahů,** nárůst za rok 2017 cca. **75 tisíc** obsahů (**135 000** stran).
- **Fulltexty obsahů** - převedené obsahy na text pomocí technologie OCR a zpřístupněné pro indexaci knihovním systémem
- **Anotace** - nakladatelské, autorské i knihovnické anotace získané z různých zdrojů, aktuálně přes **430 tisíc anotací**
- **Komentáře a hodnocení** - možnost stažení dostupných komentářů a zároveň i možnost exportu komentářů z jiných systémů na obálkyknih.cz a tím zpřístupnění ostatním knihovnám, aktuálně cca. **3.12 miliónu hodnocení** u **157 tisíc titulů**, dále pak **10 tisíc komentářů**
- **Fotografie autorit -** portrétové fotografie autorů a jiných osobností z Autoritní databáze Národní knihovny ČR (http://aut.nkp.cz), aktuálně přes **50 tisíc fotografií**
- **Citace dokumentů** – citace dle normy ISO 960, aktuálně cca. **1,5 miliónu vygenerovaných citací**

Služeb projektu Obálkyknih.cz využívá většina knihoven v České republice. Dále pak muzea, archivy, veřejné projekty aj. Správcem projektu Obálkyknih.cz je Jihočeská vědecká knihovna v Českých Budějovicích (JVK) společně s Moravskou zemskou knihovnou (MZK).

Přehled vlastností projektu:

- hlavní servery jsou provozovány v Jihočeské vědecké knihovně v Českých Budějovicích, záložní server je umístěn v Moravské zemské knihovně v Brně
- denně do databáze je nově nahráno nebo je upraveno pomocí skenovacího klienta průměrně 350 dokumentů
- další dokumenty se automaticky sklízejí z externích zdrojů - nakladatelé, vydavatelé, webové portály ... průměrně denně přes 500 dokumentů
- denní přírůstek dat činí 6 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
- servery odbavují průměrně 2 milióny požadavků denně (díky optimalizaci procesů a zvyšování počtu multidotazů na více identifikátorů zároveň se počty drží na podobné úrovni jako v předchozích letech, přestože počet uživatelů roste)

Statistiky přispívání přes skenovacího klienta za období leden - prosinec 2017:

Počet odeslaných dokumentů	**98 493**
Počet uložených obálek (COVER)	**85 004**
Počet uložených stran obsahu	**145 501**
Počet uložených obrázků autorit	**1 854**

Počty odeslaných stran a titulů přes skenovacího klienta dle jednotlivých knihoven (rok 2017):

STRAN	TITULŮ	SIGLA	NÁZEV
54153	26231	CBA001	Jihočeská vědecká knihovna v Českých Budějovicích
18010	8511	ABA001	Národní knihovna ČR
16231	8802	BOA001	Moravská zemská knihovna
14451	1930	BOD001	Ústřední knihovna FF MU
13106	2786	ABA013	Národní technická knihovna

12800	7205	OLA001	Vědecká knihovna v Olomouci
12572	5441	ABA004	Slovanská knihovna
7777	2733	ABA008	Národní lékařská knihovna
6644	1738	BOD010	Masarykova univerzita - Právnická fakulta
6615	2177	OLD012	Knihovna Univerzity Palackého v Olomouci
5393	1988	BOE020	Knihovna Ústavního soudu
5375	1322	ZLD002	Univerzita Tomáše Bati ve Zlíně
4666	1683	CBD005	Teologická fakulta JCU
4555	4555	ABD001	Knihovna Ústavu Dálného východu Filozofické fakulty Univerzity Karlovy
4052	2456	KVG001	Krajská knihovna Karlovy Vary
3282	920	ABA007	Knihovna Akademie věd
3279	1074	BOD031	Masarykova univerzita, Fakulta sociálních studií, Ústřední knihovna
2861	787	BOD004	Ústřední knihovna Přírodovědecké fakulty MU
2840	726	ABA006	Vysoká škola ekonomická v Praze
2798	1529	ULG001	Severočeská vědecká knihovna v Ústí nad Labem
2551	869	ABB019	Knihovna Sociologického ústavu AV ČR, v.v.i.
2445	716	LID001	Technická univerzita v Liberci, Univerzitní knihovna
2313	1366	HBG001	Krajská knihovna Vysočiny
2222	846	BOE451	Knihovna Biskupství brněnského
2207	1364	PAG001	Krajská knihovna v Pardubicích
2175	1314	LIA001	Krajská vědecká knihovna v Liberci
2093	575	ABD100	ÚK ČVUT
1821	1145	PNA001	Studijní a vědecká knihovna Plzeňského kraje
1437	464	ABD103	Univerzita Karlova-Fakulta sociálních věd-Středisko vědeckých informací
1264	719	OSA001	MSVK v Ostravě
1019	529	KLG001	Středočeská vědecká knihovna v Kladně
924	922	TAG001	Městská knihovna Tábor
456	158	ABB001	Knihovna Archeologického ústavu AV ČR, Praha, v.v.i.
365	237	ULD001	Ústřední knihovna UJEP
315	99	BOD018	Masarykova univerzita - Fakulta informatiky
294	94	BOD006	Informační centrum, ústřední knihovna Mendelovy univerzity v Brně
286	196	SMG506	Městská knihovna Antonína Marka Turnov
277	152	ULE301	Muzeum města Ústí nad Labem
223	61	ABA011	Parlamentní knihovna
200	199	ABD027	Evangelická teologická knihovna UK
169	70	ABG312	Knihovna Jabok
162	93	BOE303	Knihovna Moravské galerie v Brně
156	111	ZLG001	Krajská knihovna Františka Bartoše ve Zlíně
133	49	ABE367	Archiv hlavního města Prahy
128	78	CHG001	Městská knihovna v Chebu, p.o.
115	115	BOD033	Univerzitní knihovna pro studenty se specifickými nároky MU
101	46	ABA100	Všenorská knihovna a informační centrum Berounka
87	36	JID501	Knihovna Univerzitního centra Telč Masarykovy univerzity
86	64	SOG504	Městská knihovna Chodov
39	29	PTG001	Městská knihovna Prachatice
4	2	ABB503	Knihovna ÚACH AVČR, v.v.i.
1	1	MEG502	Městská knihovna Neratovice

1	1	PBG001	Knihovna Jana Drdy Příbram

Úkoly řešené v rámci projektu v roce 2017:

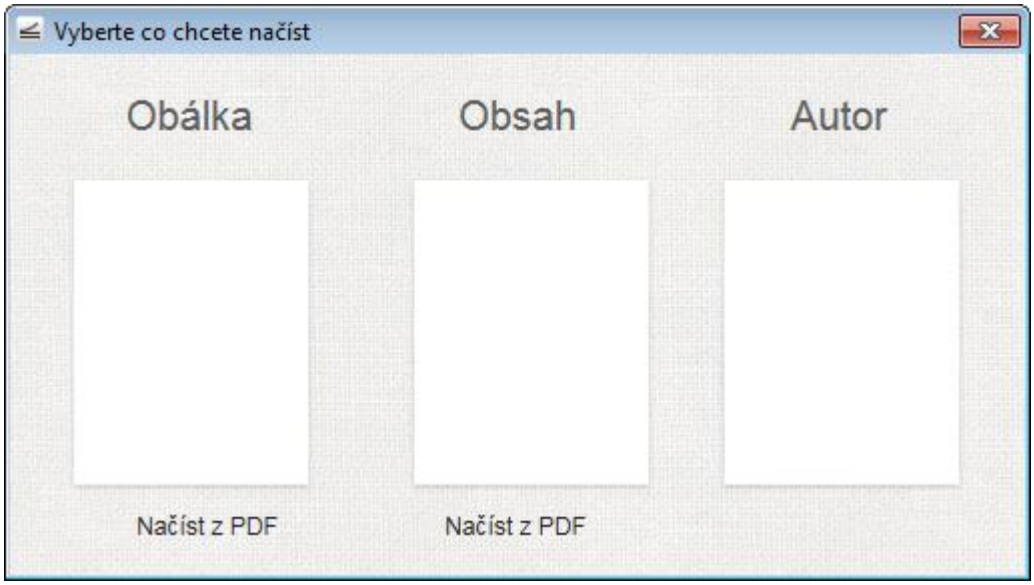
Skenovací klient

Vylepšení skenovacího klienta o nové funkce pro snadnější a rychlejší práci. Jednalo se o tři hlavní úkoly:

Vkládání obálek a obsahů z vlastních PDF souborů

- slouží pro vkládání obálek a obsahů elektronických dokumentů, které není nutné znovu skenovat
- implementováno jako rozšíření stávajícího vkládání obrázku z lokálního PC
- využívá systém GhostScript pro konverzi stránek PDF na obrázky
- nenačítá se náhled všech stránek PDF (trvalo by dlouho), ale je možné zvolit rozsah stránek k náhledu
- ze stránek náhledu je možné klasickými Windows zkratkami (pomocí kláves SHIFT a CTRL + klikání myší) zvolit jednu nebo více stránek pro vložení do klienta

Odkaz pro vložení obálek („Načíst z PDF“):



Dialogové okno pro výběr stránek PDF

Umožnění skenování všech autorit titulu

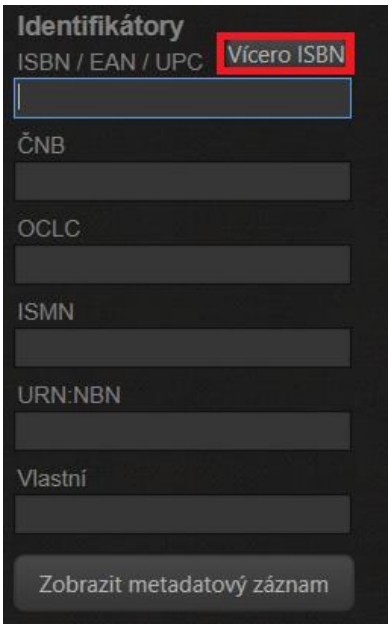
- rozšíření skenování podobizen autorit na všechny autority titulu (doposud bylo možné skenovat pouze autora s primární odpovědností)
- implementováno do stávajícího rozhraní beze změny logiky fungování skenovacího klienta
- po oskenování autority se zobrazuje vstupní drop-down pole se seznamem všech autorů k výběru právě skenovaného autora
- je možné změnit autora u každé skenované fotky, v případě skenování kvalitnější fotografie již existující osoby se obrázek automaticky nahradí

Rozhraní klienta pro výběr konkrétního autora:

Skenování více identifikátorů ISBN

- umožnění skenování více identifikátorů ISBN. Rozšíření GUI o možnost zadání více identifikátorů ISBN jednoho titulu.
- nad textovým oknem pro hlavní identifikátory (ISBN / EAN / UPC) je tlačítko "Vícero ISBN", klikem je možné přidat další ISBN identifikátory titulu

Přidání více ISBN jednoho titulu:



Sklízení dalších repositářů Kramerius

Cílem úkolu je sklízení dalších repositářů Kramerius a tím obohacení projektu Obálkyknih.cz o nové obálky a obsahy z knihoven, které nebyly do procesu sklízení zatím zapojeny. Modul Kramerius aktuálně umožňuje sběr obálek a TOC ze všech knihoven využívajících systém Kramerius. Seznam digitálních knihoven je periodicky stahován z Registru Kramériů (http://registr.digitalniknihovna.cz/) ve formátu JSON (http://registr.digitalniknihovna.cz/libraries.json), který se porovná se záznamy v lokální tabulce eshop. Zdroj neobsahuje identifikátor SIGLA, proto je nutné nejprve novou digitální knihovnu Kramerius provázat na už existující knihovnu z tabulky library. Pro každou z knihoven se pohledávají tyto typy záznamů: monograph, periodicalitem, soundrecording, archive, graphic, sheetmusic, manuscript.

Cílem crawleru je získat seznam stránek a v nich identifikovat obálku a obsah. Přitom platí:

- jako obálku vybere přednostně stránky typu FrontCover, FrontJacket, alebo TitlePage.
- pokud se nenajde ani jedna z předchozího, vybere se první stránka, která ale není je typu Spine nebo Hřbet.
- jako obsah vyber stránky typu TableOfContents.

Při výběru kvality originálu se postupuje v pořadí iiif, full, thumb a vybere se první možný. Formát thumb je přitom k dispozici vždy. V případě periodika je přitom potřebné zeptat se i na BIBLIO_MODS kořenového dokumentu tj. souborného záznamu periodika.

Příklady API Kramerius:

Vyhledání záznamů:
http://kramerius.mzk.cz/search/api/v5.0/search?q=fedora.model:monograph%20AND%20modified_date:[2017-01-10T00:00:00Z%20TO%202017-01-31T23:59:59Z]&fl=PID&wt=xml&start=0

Metadata digitalizovaného dokumentu:
http://kramerius.mzk.cz/search/api/v5.0/item/uuid:0e3f44d7-890a-4976-aeb9-783228b4a2f0

Seznam streamů digitalizovaného dokumentu:
http://kramerius.mzk.cz/search/api/v5.0/item/uuid:0e3f44d7-890a-4976-aeb9-783228b4a2f0/streams

Jeden ze streamů, konkrétně bibliografický MODS:
http://kramerius.mzk.cz/search/api/v5.0/item/uuid:0e3f44d7-890a-4976-aeb9-783228b4a2f0/streams/BIBLIO_MODS

Potomci digitalizovaného dokumentu (seznam stránek):
http://kramerius.mzk.cz/search/api/v5.0/item/uuid:0e3f44d7-890a-4976-aeb9-783228b4a2f0/children

Streamy stránky:
http://kramerius.mzk.cz/search/iiif/uuid:0e3f44d7-890a-4976-aeb9-783228b4a2f0@363/full/,510/0/default.jpg
http://kramerius.mzk.cz/search/api/v5.0/item/uuid:0e3f44d7-890a-4976-aeb9-783228b4a2f0@363/full
http://kramerius.mzk.cz/search/api/v5.0/item/uuid:0e3f44d7-890a-4976-aeb9-783228b4a2f0@363/thumb

Optimalizace běhu služeb

Jednalo se o tyto úkoly:

Zjednodušená odpověď metadatového API

Poskytnutí knihovnám jednoduchého metadatového kontejneru obsahujícího jen nejnutnější položky pro zobrazení obálek a obsahů v katalozích. Realizováno přidáním parametru "simple" do URI metadatového API.

Příklad:
http://cache1.obalkyknih.cz/api/books/?multi=[{%22isbn%22:%229788090251434%22}]&sigla=CBA001&pretty=1&simple=1

Odpověď:


```
[

 {
  "cover_preview510_url": "http://cache.obalkyknih.cz/file/cover/1686146/preview510",
  "cover_thumbnail_url": "http://cache.obalkyknih.cz/file/cover/1686146/thumbnail",
  "cover_icon_url": "http://cache.obalkyknih.cz/file/cover/1686146/icon",
  "backlink_url": "http://www.obalkyknih.cz/view?isbn=9788090251434",
  "book_id": "110273872",
  "cover_medium_url": "http://cache.obalkyknih.cz/file/cover/1686146/medium"
 }
]
```


Optimalizace obrázků

Optimalizaci bezztrátově komprimovaných obrázků uložených v DB a na diskovém úložišti bez ztráty kvality. Pro optimalizaci se používá ZopfliPNG algoritmus. Operace probíhá automatizovaně pomocí skriptu, který postupně optimalizuje všechny obrázky ve formátu png uložené v databázi. Skript si pamatuje, kde naposledy skončil. Při dalším spuštění skript pokračuje v optimalizaci. Výsledná velikost obrázku je v průměrně menší o 15% (v rozmezí 7% až 25%). Úspora na diskových úložištích dosahuje desítek GB.

Vyhledávání více identifikátorů ISBN stejného titulu.

Pomocí skenovacího klienta je aktuálně možné uložit více identifikátorů ISBN stejného titulu. Před importováním záznamu se zkontroluje existence titulu (vyhledává se podle metadat titulu + všech identifikátorů ISBN). Pokud titul v databázi už existuje, přidají se nově vložené identifikátory do tabulky product_params (tabulka obsahuje všechny "ostatní" identifikátory ISBN daného titulu, v hlavní tabulce "book" a "product" je uložen vždy jen jeden identifikátor ISBN). Při úspěšném importu záznamu se uloží ostatní identifikátory do tabulky product_params**.**

Při vyhledávaní podle ISBN se také kontroluje tabulka s ostatními ISBN identifikátory a při úspěšném nalezení zobrazí daný záznam.

Import e-knih z báze MLP.cz a jejich propojení na klasické tituly

Cílem úkolu je stažení báze e-knih z vlastní produkce Městské knihovny v Praze a e-knih z Krameria (sběr URL na fulltexty v rámci Kramerius API - ve formátu PDF) a jejich napojení na záznamy projektu obalkyknih.cz. V metadatovém kontejneru titulu je obsažena informace o dostupnosti elektronické verze titulu (bez ohledu na konkrétní vydání) s možností propojení v knihovních systémech a nabídnutí čtenáři ke stažení elektronické verze knihy.

Záznamy se ukládají do existující DB tabulky book s příznakem identifikujícím e-knihu. Došlo k vytvoření relační tabulky typu M:N pro svázání e-knihy s monografiemi. Jednu e-knihu je možné svázat s vícero monografiemi. Jedna monografie zároveň může obsahovat relaci s více e-knihami. Zároveň existuje tabulka parametrů e-knihy (URL ve formátu epub, pdb, pdf, a jiné). Seznam parametrů nebude omezený, tj. bude možné přidat do budoucna i nové formáty e-knih. K metadatovému kontejneru je připojeno pole svázaných e-knihy (i více) a u e-knihy jsou vypsány všechny dostupné formáty.
Zároveň vzniklo administrativní rozhraní, kterým je schopen správce OKCZ svazovat monografie a e-knihy. Protože neexistuje zdroj relací e-kniha - monografie, budou se automaticky vytvářet návrhy na párování (zdroj návrhů je databáze *SKC-UTF*). Skript se pokusí vyhledat podobné záznamy s totožnými identifikátory, případně titulem + autory. Administrátor OKCZ musí rozhodnout, jestli daná vazba je správná, nebo ne. Schválením návrhu se vazba vytvoří a je ihned dostupná v metadatovém kontejneru.

Příklad - titul „Cirkus Humberto“ od Bass, Eduard

E-kniha:
https://www.obalkyknih.cz/view?book_id=127833341

Příklad dostupných formátů eknihy**:** epub html pdf prc

Příklad propojení na klasické tituly:
https://www.obalkyknih.cz/view?nbn=cnb000260949
https://www.obalkyknih.cz/view?nbn=cnb000260953
https://www.obalkyknih.cz/view?isbn=8590236083226

Dotaz na metadatový kontejner konkrétního titulu – klasické knihy:
http://cache.obalkyknih.cz/api/books?multi=[{%22isbn%22:%228590236083226%22}]&sigla=CBA001&pretty=1

Vrací parametr ebook v datovém kontejneru titulu:


```
….
"ebook": [
  {
   "url": "https://web2.mlp.cz/koweb/00/04/36/09/86/cirkus_humberto.epub",
   "type": "epub"
  },
  {
   "url": "https://web2.mlp.cz/koweb/00/04/36/09/86/cirkus_humberto.html",
   "type": "html"
  },
  {
   "url": "https://web2.mlp.cz/koweb/00/04/36/09/86/cirkus_humberto.pdf",
   "type": "pdf"
  },
  {
   "url": "https://web2.mlp.cz/koweb/00/04/36/09/86/cirkus_humberto.prc",
   "type": "prc"
  }
 ],
```


Kontroly anotací

Proběhla kontrola cca. 510 tisíc nezkontrolovaných anotací, které byly získány sklízením dat od vydavatelů, obchodních portálů, exportů ze souborných katalogů, katalogů přispívajících knihoven a jiných zdrojů. Anotace obsahovali reklamní texty, chyby a nevalidní data. Kontroly anotací bylo nutno provést živými knihovníky. Bohužel nasazení automatických systémů kontrol a schvalování nefunguje spolehlivě, což bylo opakovaně testováno v reálném nasazení s negativními výsledky. Automatickým povolením anotací bez kontrol by došlo k degradaci celé služby.
Obálkyknih.cz zpřístupňují aktuálně přes 430 000 anotaci českých a zahraničních knih (nárůst o 160 000 anotací) pro zobrazení v katalozích knihoven a zároveň na indexaci pro plnotextové vyhledávání v nich. Zkontrolované anotace obohatily záznamy poskytované přes rozhraní Obálkyknih.cz a ihned po kontrole je mohly využít všechny knihovny v ČR.

Další úkoly řešené v roce 2017 mimo projekt:

- spolupráce s portálem cbdb.cz (http://www.cbdb.cz/) a import hodnocení titulů jejich čtenáři, získáno přes 3 milióny hodnocení u cca 100 tisíc titulů
- údržba a podpora skenovacího klienta pro nahrávání dat knihovnami do projektu
- úprava citací dle platné normy a připomínek uživatelů služby
- optimalizace běhu procesů backendu
- kontrola skenovaných periodik a opravy nalezených problémů, metodické vedení přispěvatelů
- upgrade serverů a použitých SW za účelem větší funkcionality a vyšší stability běhu služby
- aktualizace SSL certifikátů pro servery projektu
- aktualizace webových stránek projektu, úprava zobrazení čísel periodik (postupné načítání)
- webové rozhraní pro přispívání chybějících fotografií autorit uživateli portálu

- prezentace projektu mezi odbornou i laickou veřejností (konference Knihovny současnosti, článek v časopisu Čtenář, přednáška studentům v rámci cyklu Jinonické informační pondělky, ...)
- aktualizace metodických pokynů a manuálu pro knihovny a knihovní systémy
- emailová a telefonická podpora projektu, spolupráce s tvůrci AKS a CPK

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

V Českých Budějovicích 8.1.2018

Ing. Jiří Nechvátal
Jihočeská vědecká knihovna v Českých Budějovicích