SOUBORNÝ KATALOG ČR V ROCE 2005

Eva Svobodová – Danuše Vyorálková, Národní knihovna ČR

Počátkem roku 2005 zaznamenal Souborný katalog ČR tři významné změny:

- přechod na formát MARC21
- využívání nových mechanismů pro import a deduplikace záznamů
- změnu přístupu k přidělování vah dodaným záznamům.

Konverze záznamů z báze SK ČR do formátu MARC 21 a přechod na nové mechanismy deduplikace a importů proběhly současně na počátku roku 2005. Během posledních dvou měsíců roku 2004 probíhaly přípravné práce, které spočívaly především v kontrolách záznamů a jejich globálních i ručních opravách. Použité kontrolní a konverzní programy byly již velmi dobře prověřeny a doladěny. K jejich prověření došlo v červnu 2004 během konverze ostatních bází Národní knihovny ČR do formátu MARC21. Existovala i celá řada zkušeností s tím, které chyby v záznamech je před konverzí nutné opravit, aby nedocházelo ke ztrátě dat nebo kolapsům při konverzi. V bázi Souborném katalogu ČR bylo opraveno několik desítek tisíc záznamů.

Vlastní konverze byla provedena s cílem zajistit co největší kvalitu záznamů. Ze Souborného katalogu ČR byly vyexportovány záznamy podle sigel (lokačních značek) dodavatelů. Z těchto záznamů byly vyřazeny záznamy patřící Národní knihovně ČR, u nichž nebyl připojen další odběratel. Důvodem byla skutečnost, že tyto záznamy byly již v létě jednou do formátu MARC 21 převedeny v rámci konverze báze NKC. Navíc v bázi NKC již prošly příslušnými opravami a byla u nich vytvořena řada nových vazeb do autorit. Jevilo se tedy jako výhodnější je znova nekonvertovat, převzít je přímo z báze NKC a naimportovat do nové verze SK ČR jako základ. Poté byly postupně do SK ČR naimportovány s novými váhami záznamy ostatních dodavatelů. Zároveň probíhala deduplikace podle nově postavených deduplikačních klíčů. Zvolený způsob v konečném důsledku vedl ke snížení celkového počtu záznamů v bázi, neboť došlo ke spojení duplicitních záznamů, které se dříve používanými deduplikačními programy spojit nepodařilo. Malá část záznamů byla ze souborného katalogu vyřazena z důvodu jejich nestandardnosti, která znemožnila jejich konverzi.

Celkový počet záznamů v Souborném katalogu ČR ke dni 31.7.2005 je **2.218.479**, z toho 121.832 tvoří záznamy českých a zahraničních seriálů a 69.853 záznamy speciálních druhů dokumentů.

S přechodem na nové mechanizmy pro deduplikaci a import záznamů do SK ČR souvisí některé drobné změny ohledně zasílání dat do SK ČR. Před importem jsou záznamy kontrolovány, zda již příslušná knihovna záznam se stejným identifikačním číslem v SK ČR nemá – tyto záznamy jsou z dalšího zpracování vyřazeny. Seznam z tohoto důvodu odmítnutých záznamů se objeví ve statistice importu a knihovna by si měla ověřit, zda záznam poslala omylem podruhé, nebo zda má v bázi duplicitu identifikačních čísel. Záznam bude odmítnut i v případě, že knihovna záznam, (který byl dříve do SK ČR již jednou přijat), ve své lokální bázi změnila (opravila) a zasílá ho znovu z důvodu, aby byla provedená změna zohledněna i v SK ČR. Importovací program, s touto variantou nepočítá. Připravuje se ale program, který by zvlášť zpracovával záznamy, které mají být v bázi opraveny, ale dosud není v provozu. Pokud má knihovna zájem o opravu chyby v bibliografickém záznamu, může tak učinit pouze ve spolupráci se správcem SK ČR.

Další změnou je vyřazení záznamů, které jsou na porovnávací klíče duplicitní uvnitř právě importované dávky (první záznam je vždy zpracován, druhý odmítnut).

Soubory zasílané do SK ČR musí být pojmenovány v souladu s názvovou konvencí.

Tj. jméno souboru musí mít podobu **aaa000kk.ddd_xx**, kde prvních 6 znaků označuje siglu instituce, znaky 7 a 8 před extenzí označují použité kódování diakritiky, první 3 znaky extenze označují formát dat, další tři znaky extenze za podtržítkem obsahují vlastní pojmenování souboru.

např. osa001lg.uis_20050618

Soubor zaslalala knihovna se siglou OSA001, soubor je ve znakové sadě PC Latin2 + GIZMO a ve formátu dat UNIMARC – ISO 2709 a byl odeslaný 18. června 2005.

Přispívající knihovny mohou zasílat záznamy do SK ČR v následujících formátech dat :

mal MARC21 - exportní soubor Aleph500
mis MARC21 - ISO 2709
dtt UNIMARC - exportní soubor Aleph500
dat UNIMARC - exportní soubor Aleph300
uis UNIMARC - ISO 2709
rum UNIMARC řádkový

(Novinkou je proti loňskému roku možnost zasílat záznamy ve formátu MARC 21. Výměnný formát již nadále není podporován a záznamy v tomto formátu již nejsou do SK ČR přijímány.)

Přispívající knihovny mohou zasílat záznamy do SK ČR v následujících kódováních češtiny :

uc Unicode UTF-8
lg PC Latin 2 (+ GIZMO)
kg kód Kamenických (+ GIZMO)
sg ISO 8859-2 (+ GIZMO)
wg CP-1250 (+GIZMO)

(Zde je novinkou kódování CP-1250.)

I nadále platí pravidlo, že počet záznamů v zasílaných souborech by neměl přesahovat 10.000 záznamů a záznamy vytvořené v rámci retrospektivních konverzí by měly být zasílány ve zvláštních souborech, obsahujících maximálně 20.000 záznamů.

Se spuštěním nových mechanizmů pro deduplikaci a import dat do SK ČR souvisí i změna vzhledu statistik o importu dat. Statistika je přístupná nadále na webu na adrese : http://sigma.nkp.cz/web/skc/sigla/sigla.htm

V adrese je třeba místo textu … sigla/sigla… doplnit skutečnou siglu knihovny. Např. statistiky o importech knihovny se siglou OSA001 je možné najít na adrese: http://sigma.nkp.cz/web/skc/osa001/osa001.htm a má následující podobu:

Přehled importovaných souborů

Soubor	Datum	Zasláno	Perio	IN	ADD	UPD	NEW	Přijato
osa001_0101	20050217	449	0	448	196	6	246	448
osa001_0102	20050217	554	0	553	118	7	396	521
osa001_1202	20050217	333	0	319	115	6	174	295
osa001_3	20050313	380	0	378	246	0	93	339
osa001_4	20050313	351	0	348	162	10	148	320
osa001_5	20050316	470	0	466	211	1	224	436
osa001_010405	20050404	452	0	451	150	11	255	416
osa001_150405	20050415	475	0	475	260	4	185	449
osa001_020505	20050503	534	0	533	227	8	273	508
osa001_ger	20050510	51	0	50	16	0	34	50
osa001_160505	20050516	411	0	410	194	7	190	391
osa001_010605	20050602	477	0	477	178	14	272	464
osa001_150605	20050616	419	0	418	136	28	244	408
osa001_010705	20050714	521	0	520	132	23	341	496
osa001_180705	20050718	361	0	360	172	13	168	353
Soubor	Datum	Zasláno	Perio	IN	ADD	UPD	NEW	Přijato

Ze statistiky lze vyčíst, že od začátku roku 2005 bylo z této knihovny naimportováno celkem 15 dávek dat, kdy byly jednotlivé dávky importovány a s jakým výsledkem (kolik záznamů bylo připsáno, kolik záznamů bylo přijato jako záznamy nové atd.). Kliknutím na jméno souboru lze přejít k detailnějším údajům o počtech vyřazených záznamů a důvodech jejich vyřazení.

Soubor	Datum	Celkem	Knihy	Spec.d.	Perio	Chyby form.1	Chyby form.2	Chyby MARC	Chyby konv.	OK-mono	Dupl. 910	Dupl. KEY1	.Dupl. KEY2	Přijato
osa001_0101	20050217	449	**449**	**0**	**0**	0	0	1	0	448	0	0	0	448
osa001_0102	20050217	554	**554**	**0**	**0**	0	0	1	0	553	32	0	0	521
osa001_1202	20050217	333	**333**	**0**	**0**	0	0	14	0	319	24	0	0	295
osa001_3	20050313	380	**380**	**0**	**0**	0	0	2	0	378	39	0	0	339
osa001_4	20050313	351	**351**	**0**	**0**	0	0	3	0	348	28	0	0	320
osa001_5	20050316	470	**470**	**0**	**0**	0	0	4	0	466	29	0	1	436
osa001_010405	20050404	452	**452**	**0**	**0**	0	0	1	0	451	35	0	0	416
osa001_150405	20050415	475	**475**	**0**	**0**	0	0	0	0	475	25	0	1	449
osa001_020505	20050503	534	**534**	**0**	**0**	0	0	1	0	533	25	0	0	508
osa001_ger	20050510	51	**51**	**0**	**0**	0	0	1	0	50	0	0	0	50
osa001_160505	20050516	411	**411**	**0**	**0**	0	0	1	0	410	19	0	0	391
osa001_010605	20050602	477	**477**	**0**	**0**	0	0	0	0	477	13	0	0	464
osa001_150605	20050616	419	**419**	**0**	**0**	0	0	1	0	418	10	0	0	408
osa001_010705	20050714	521	**521**	**0**	**0**	0	0	1	0	520	24	0	0	496
osa001_180705	20050718	361	**361**	**0**	**0**	0	0	1	0	360	7	0	0	353
Soubor	Datum	Celkem	Knihy	Spec.d.	Perio	Chyby form.1	Chyby form.2	Chyby MARC	Chyby konv.	OK-mono	Dupl. 910	Dupl. KEY1	.Dupl. KEY2	Přijato

Např. dávka z 18.7.2005 obsahovala 361 záznamů. Celkem bylo přijato do báze SK ČR 353 záznamů. U jednoho záznamu byla důvodem nepřijetí chyba v UNIMARCu (nepřítomnost povinného pole 101 – viz níže – seznam vadných záznamů) u sedmi záznamů se jednalo o záznamy, které byly již dříve do SK ČR stejnou knihovnou zaslány (proto přijaty znovu nebyly.) Kliknutím na podtržené jméno souboru je možno získat log z kontroly na správnost polí UNIMARCu :

Informace o záznamech, které nebyly přijaty do SKC
Výsledky kontroly souboru na pole MARC:

Přehled vyskytujících se chyb:
#101#: Pole '101' je povinné

Seznam vadných záznamů:

000000078_2640122632' ** #101#: Pole '101' je povinné
CONVERT 361 RECORDS
360 RECORDS OK

1 RECORDS WITH ERRORS

Duplicity identif.čísla vůči bázi (poř.č. = syst.č. v SKC)

##001023593_2640105536	**	duplicita	001	vůči	bázi
##002170915_2640122313	**	duplicita	001	vůči	bázi
##002259882_2640122332	**	duplicita	001	vůči	bázi
##001256912_2640108601	**	duplicita	001	vůči	bázi
##001256628_2640107047	**	duplicita	001	vůči	bázi
##002052156_2640115559	**	duplicita	001	vůči	bázi
##002052157_2640115560	**	duplicita	001	vůči	bázi

Zásadní změnou prošly v Souborném katalogu deduplikační programy. Pro deduplikaci jsou používány deduplikační klíče, které se vytváří před importem záznamu do báze. Klíče se vytvářejí z různých polí a podpolí záznamu (z kterých polí a podpolí záleží na druhu zpracovaného dokumentu) různými metodami a pomocí programu jsou zakódovány (pomocí hashovaní metody MD5) na řetězec s konstantní délkou 32 znaků. Klíče jsou uloženy ve speciálních polích (např KEY1, KEY2…) a je zajištěna jejich změna v případě editace záznamu. V každém záznamu existuje alespoň jeden deduplikační klíč.

Příklad : Pro monografie je možné vytvořit deduplikační klíče dva.
Zdrojem pro vytvoření prvního deduplikačního klíče (KEY1) jsou následující pole formátu
MARC 21 :

020 a Mezinárodní standardní číslo knihy (ISBN) -
245 a název
300 a fyzický popis – rozsah

Pokud nejsou v záznamu uvedena všechna potřebná pole (podpole) současně **klíč se nevytváří.**

Před použitím se údaje ze zdrojových polí „opracují“
- odstraní se z nich se veškeré znaky a mezery kromě znaků alfanumerických
- odstraní se unikódové znaky pro nelatinková písma
- znaky se speciální diakritikou (vč. české) se převedou na hodnotu prostého ASCII znaku

Jako zdroj pro vytvoření klíče u pole 020 slouží všechny uvedené výskyty tohoto pole, u polí 245 a 300 pouze první výskyt.

Z pole 020 – jsou pro porovnání použity pouze znaky uvedené do mezery
Př. 020 a 80-86518-62-0 (váz.) – použije se pouze **8086518620**

Z pole 245 jsou pro porovnání použity pouze první tři znaky pro kontrolu
Př. 245 a Šifra mistra Leonarda - fakta : - použije se pouze **Šif**

Z pole 300 jsou pro porovnání použity pouze numerické znaky z vyššího čísla v podpoli a
Př. 300 a 2, 256 s. – použije se pouze **256**

Zdrojem pro vytvoření druhého deduplikačního klíče (KEY2) jsou následující pole formátu MARC 21 :

245 a název
245 n číslo označení části/sekce díla
245 p název části/sekce díla
100 a hlavní záhlaví – osobní jméno
260 c nakladatelské údaje – datum vydání
300 a fyzický popis – rozsah

Klíč se vytváří pokud je přítomno alespoň jedno z uvedených polí (podpolí).

Stejně jako při vytváření prvního klíče se před použitím údaje ze zdrojových polí nejprve „opracují“. Jako zdroj se používají pouze první výskyty uvedených polí a podpolí .

Na rozdíl od prvního klíče se jako zdroj používá obsah celého podpole 245 a – až do znaku „dvojtečka“. U podpole n se použijí jen číselné znaky. Obsah podpole p se využije v podobě v jaké je uveden.
Z pole 100 podpole a se využívá část textu do znaku „čárka“.
Př. 100 a Cox, Simon – použije se pouze **Cox**

Obsah pole 260 podpole c – využívají se pouze numerické znaky.
Pole 300 je používáno stejným způsobem jako při tvorbě prvního klíče.

Podobně jsou vytvářeny klíče u speciálních druhů dokumentů a u seriálů, kde je ale počet vytvářených klíčů vyšší. (pro porovnávání seriálů jsou vytvářeny až 4 klíče). Při porovnávání seriálů navíc platí, že záznamy shodné na KEY3 a KEY4 jsou importovány jako nové (neduplicitní) záznamy, programem jsou označeny za potencionální duplicitu a konečné slovo při jejich deduplikaci má správce báze.

Při deduplikaci a následném importu záznamů do SK ČR je velmi důležitá otázka kvality záznamu. Žádoucí je, aby v SK ČR vždy ze dvou stejných záznamů zůstal záznam kvalitnější. Proto se záznamům před deduplikací přiděluje váha – číselné ohodnocení kvality. Dříve se přidělovala váha

celé dávce záznamů, nyní se přiděluje každému záznamu zvlášť. K základní váze, kterou má dávka, resp. knihovna, která dávku záznamů zasílá, přidělenu v tabulce (na základě předcházející analýzy dat) se přičítají „bonusové“ body za přítomnost určitých údajů v záznamu. Bodově se zvýhodňují záznamy obsahující :

- v MARC21 – podpole 7, (v UNIMARC - podpole 3) které signalizuje autoritní podobu jmenného záhlaví
- v MARC21 – pole 020/022/024 (v UNIMARC 010/011/013)
- v MARC21 některé z polí bloku 7XX s podpolem t (v UNIMARC některá pole bloku 4XX)
- v MARC21 pole 080 nebo 072 (v UNIMARC 675 nebo 615)
- jakéhokoli pole věcného popisu se slovním vyjádřením
- v MARC21 podpole 2 v poli 650 (v UNIMARC pole 606) s obsahem vyjadřujícím, že jde o některý ze schválených systémů.

Zvýhodnění je vždy o jeden bod nahoru za přítomnost některého z uvedených polí, pouze propojení na autority zvýhodňuje záznam o body dva . Maximální váha, které lze dosáhnout, má hodnotu 17.

Záznamy, které vytvořily knihovny v rámci retrokonverzí svých katalogů jsou zasílány do SK ČR obvykle s nižší základní vahou (4 – 6).

V závěru roku 2005 čeká SK ČR ještě jedna změna týkající se systému ALEPH 500 a to přechod na novou verzi systému – verzi 16. 02.

Pokud se týká novinek, byly v květnu 2005 rozšířeny služby SK ČR o možnost odeslání objednávky MVS z prostředí SK ČR. Významně se rozšířil i počet uživatelů (knihoven), využívajících možnost aktualizovat údaje o odběru seriálů v SK ČR on-line pomocí interaktivního formuláře. Podrobněji o využití tohoto formuláře i o možnostech týkajících se MVS pojednává příspěvek Z. Manouškové a D. Vyorálkové s názvem Služby Souborného katalogu ČR v roce 2005, uveřejněný v tomto sborníku.

Před dokončením je retrokonverze Retrospektivního katalogu zahraničních periodik, která probíhá od roku 1994. Díky retrokonverzi tohoto katalogu může uživatel nacházet na internetu i informace o seriálech, které v elektronické podobě zatím v jednotlivých knihovnách nejsou k dispozici. Knihovníci těchto knihoven tak získávají v SK ČR nástroj pro retrokonverzi vlastních katalogů seriálů, neboť vytvořené záznamy lze zdarma sdílet prostřednictvím protokolu Z39.50. V návaznosti na dokončení retrokonverze Retrospektivního katalogu zahraničních periodik se připravuje verifikace záznamů, které byly v průběhu retrokonverze vytvořeny. Cílem této akce je ověřit, zda odebírané seriály jsou stále ve fondech knihoven, které je před více než čtyřiceti lety nahlásily. O spolupráci na verifikaci zatím byly požádány všechny knihovny, které prošly v roce 2005 školením na využívání

interaktivního formuláře pro aktualizaci periodik a řada z nich ověřovaní odběrů periodik, jejichž záznamy vznikly v rámci retrokonverze, již zahájila. Do konce roku 2005 by o spolupráci na verifikaci měly být požádány všechny ostatní knihovny. Pevně věříme v jejich ochotu spolupracovat.

Kvalita a použitelnost souborného katalogu v mnoha ohledech závisí na systému, ve kterém je budován a na programech, které využívá. Omezení systému může správce souborného katalogu obejít, nedostatky vylepšit, chyby v programech opravit. Bez aktivní spolupráce knihoven, které se na budování SK ČR podílejí, však zůstává jeho schopnost celkového zkvalitnění a využitelnosti SK ČR velmi omezená.

Použitá literatura :

Dvořáková, Helena. Souborný katalog a Národní autority – jak obejít omezení komerčního systému. *Automatizace knihovnických procesů 2005. 10. ročník semináře. Liberec 3. a 4. května 2005.* Dostupné na World Wide Web <http://www.akvs.cz/akp-2005/14-dvorakova.pdf>