

Vključitev retrospektivne baze podatkov Narodne in univerzitetne knjižnice RETRORK v vzajemno bazo podatkov COBIB

Matjaž Zalokar, Mitja Debeljak

Strokovni članek

IZVLEČEK

Opisan je postopek za vključitev retrospektivne baze podatkov Narodne in univerzitetne knjižnice (NUK) v vzajemno bazo podatkov COBIB. V njej so bili bibliografski zapisi za gradivo iz NUK-ovega najstarejšega listkovnega kataloga (do leta 1947). Navedeni so različni algoritmi za ugotavljanje potencialnih duplikatov ter iz njih izhajajoče primerjave, analize in zaključki.

KLJUČNE BESEDE

retrospektivna konverzija, Narodna in univerzitetna knjižnica, COBIB, vzajemni katalogi

Uvod

Od leta 1996 poteka v NUK-u retrospektivna konverzija listkovnih katalogov. Le-te je mogoče razdeliti na tri skupine: abecedno imenski katalog, ki je nastajal od ustanovitve knjižnice leta 1774 do 1947, katalogi za obdobje od 1948 do 1988 in katalogi zbirk. (1) V prvi fazi, ki je bila zaključena poleti 1998, je bila narejena retrospektivna konverzija najstarejšega listkovnega kataloga (do 1947), v katerem je bilo 95 000 listkov. Rezultat te

ABSTRACT

Describes the process of incorporating the National and University Library's (NUK's) retrospective database into COBIB, Slovenian union catalogue. The database contained bibliographic records for the publications from the library's oldest card catalogue until 1947. Gives a description of various algorithms for the identification of potential duplicates and of ensuing comparisons, analyses and conclusions.

KEYWORDS

retrospective conversion; National and University Library; COBIB; union catalogues;

konverzije je bilo okoli 65.000 bibliografskih zapisov, ki so bili preneseni v format COMARC in vključeni v lokalno bazo podatkov NUK. (2) V drugi fazi je bila 10. oktobra 1999 na osnovi baze podatkov RETRORK, ki so jo pripravili v NUK-u, narejena konverzija 60.044 bibliografskih zapisov in zbirnih podatkov o stanju zaloge v vzajemno bazo podatkov COBIB.

Celoten postopek retrospektivne konverzije, za katero je bil kot izvajalec



odgovoren IZUM, je - na kratko - potekal takole:

Bazo podatkov smo od NUK-a prejeli 22. julija 1998. Pri analizi smo ugotovili določene nepravilnosti oziroma pomanjkljivosti v bibliografskih zapisih (npr. napačni oziroma manjkajoči indikatorji v določenih poljih, manjkajoča obvezna polja itd.). Z NUK-om se je bilo treba uskladiti tudi glede najustrežnejšega algoritma za ugotavljanje potencialnih duplikatov, pri čemer se je pojavil problem podvojenih signatur. Medtem sta bili opravljeni dve testni konverziji. Prvo pravo konverzijo v lokalno bazo podatkov NUK smo izvedli 27. decembra 1998, vanjo pa je bilo vključenih 64.979 bibliografskih zapisov. 27. maja 1999 so nam iz NUK-a poslali dodatnih 7.585 zapisov iz retrospektivne konverzije, ki smo jih 21. julija 1999 dodali zapisom v lokalni bazi podatkov NUK. Temu so sledile natančne priprave na konverzijo vseh teh zapisov v COBIB.

V IZUM-u smo doslej izvedli 68 retrospektivnih konverzij. Prvi zapisi iz konverzij v vzajemno bibliografsko bazo podatkov COBIB so zapisi Univerzitetne knjižnice Maribor, ki so bili kreirani še pred vzpostavitvijo sistema vzajemne katalogizacije leta 1986. Do leta 1991 so bile v COBIB izvedene še konverzije monografskih publikacij za naslednje ustanove: Institut informacijskih znanosti iz Zagreba (1989, 1990), Narodno in univerzitetno knjižnico iz Ljubljane (1989), Inštitut Jožef Stefan iz Ljubljane (1990). (3)

Z omenjenimi konverzijami smo imeli dokaj negativne izkušnje, predvsem zaradi slabe kvalitete zapisov in duplikatov, zato so bili nekateri zapisi kasneje ali dopolnjeni z dodatnimi konverzijami ali pa pobrisani (npr. konverzija NUK-ovih zapisov za raziskovalne naloge, ki je bila narejena leta 1989 (4)). Po letu 1991 v COBIB nismo več izvajali konverzij. Vse so bile izvedene samo v lokalnih bazah podatkov posameznih knjižnic. (5)

Baza podatkov RETRORK se sicer z drugimi doslej opravljenimi konverzijami ne more primerjati, predvsem glede na obdobje, iz katerega so publikacije. Ta, najstarejši del NUK-ovega listkovnega kataloga namreč predstavlja pomembno slovensko kulturno dediščino. (6) Kljub temu pa je bila potrebna temeljita analiza zapisov pred sprejemom odločitve, da se konverzija opravi tudi v COBIB.

Algoritmi in sezname

V IZUM-u smo pripravili sedem različnih algoritmov, na podlagi katerih smo za podrobnejšo analizo izdelali trinajst posebnih seznamov. Zapisi v seznamih ustrezajo kriterijem, določenim s posameznimi algoritmi. Za vsak algoritem od 1 do 6 sta bila izdelana po dva seznama:

1. seznam, v katerem so zapisi z letom izida do 1830
2. seznam, v katerem so zapisi z letom izida enakim ali kasnejšim od 1831

Algoritmi so navedeni v istem vrstnem redu, kot smo jih preizkušali.

• Algoritem št. 1

BN=010a OR

TI= prvih trideset znakov iz 200a oz. do znaka "/"
ali ":" AND

LA=101a AND

PY=100c AND

AU=700a in prva črka iz 700b

Za zapise PY>=1900 pa še dodatno PU=prvih
petnajst znakov iz 210c AND

PP=(prva beseda)

Seznam:

T130 (<1831) (tj. zapisi za gradivo do leta 1831)

T130 (>=1831) (tj. zapisi za gradivo po letu 1831)

• Algoritem št. 2

BN=010a OR

TI= prvih petnajst znakov iz 200a oz. do znaka "/"
ali ":" AND

LA=101a AND

PY=100c AND

AU=700a in prva črka iz 700b

Za zapise PY>=1900 pa še dodatno PU=prvih
petnajst znakov iz 210c AND

PP=(prva beseda)

Seznam:

T115 (<1831) (tj. zapisi za gradivo do leta 1831)

T115 (>=1831) (tj. zapisi za gradivo po letu 1831)

• Algoritem št. 3

TI=prvih deset znakov iz 200a oz. do znaka "/" ali
":" AND

LA=101a AND

PY=100c

Seznam:

T110 (<1831) (tj. zapisi za gradivo do leta 1831)

T110 (>=1831) (tj. zapisi za gradivo po letu 1831)

• Algoritem št. 4

KW=prve štiri besede iz 200a oz. do
znaka "/" ali ":" , pri čemer se ne
upoštevajo besede, krajše od štirih
znakov. Med ključnimi besedami je
uporabljen logični operator AND.Tako oblikovanemu iskalnemu izrazu
dodamo še:

LA=101a AND

PY=100c

Seznam:

KW4 (<1831) (tj. zapisi za gradivo do leta 1831)

KW4 (>=1831) (tj. zapisi za gradivo po letu 1831)

• Algoritem št. 5

Upoštevajo se največ prve tri besede iz
naslova, daljše od treh znakov. Besede v
iskalnem izrazu, ki so daljše od sedem
znakov, odrežemo (pustimo prvih sedem
znakov) in jim v iskalnem izrazu dodamo
znak za krajšanje "*" (npr.: SELECT =
"dokumen*"/TI). Del iskalnega izraza za
ključne besede se oblikuje glede na
število besed v naslovu. Če je naslov
sestavljen iz treh ali več besed, daljših od
treh znakov, iščemo vse naslove, v
katerih se nahajata vsaj dve od prvih treh
besed, ki so daljše od treh znakov.



Tako oblikovanemu iskalnemu izrazu nato dodamo še del za jezik in leto izdaje: `SELECT = (ključne besede)/TI AND LA="xxx" AND PY="xxxx"`.

Seznami:

KW3 (<1831) (tj. zapisi za gradivo do leta 1831)

KW3 (>=1831) (tj. zapisi za gradivo po letu 1831)

• Algoritem št. 6

KW=prvi dve besedi iz 200a, pri čemer se ne upoštevajo besede, krajše od štirih znakov; besede v iskalnem izrazu, daljše od sedem znakov, so odrezane (pustimo prvih sedem znakov), v iskalnem izrazu pa je dodan znak za krajšanje "*" (npr.: `SELECT = "dokumen*"/TI`). Med ključnima besedama je uporabljen logični operator AND.

Tako oblikovanemu iskalnemu izrazu dodamo še:

`LA=101a AND`

`PY=100c`

Seznami:

KW2 (<1831) (tj. zapisi za gradivo do leta 1831)

KW2 (>=1831) (tj. zapisi za gradivo po letu 1831)

• Algoritem št. 7

Zapisi brez leta izdaje v podpolju 100c.

Seznami:

ZAPISI BREZ LETA IZDAJE

Statistika

Statistični podatki za algoritme št. 1 do 6:

Algoritem	Potencialni duplikati			Vsi prebrani zapisi iz RETRORK
	Do 1831	Po 1831	Skupaj	
Št. 1	67	593	660 1,05 %	63.247*
Št. 2	145	1224	1.369 2,17 %	63.247*
Št. 3	510	10.320	10.830 15,30 %	70.761**
Št. 4	86	3.371	3.457 4,89 %	70.761**
Št. 5	594	11.471	12.065 17,05 %	70.761**
Št. 6	420	9.545	9.965 14,08 %	70.761**

Statistični podatki za algoritem št. 7

Algoritem	Zapisi brez naslova	Zapisi brez leta izida	Vsi prebrani zapisi iz RETRORK
Št. 7	0	544 0,8%	70.761**

* Stanje 28. decembra 1998

** Stanje 21. julija 1999

Primerjava prvih 38 zapisov s seznama KW3 (<1831) z drugimi seznammi (do leta 1831)

Število zapisov, ki se pojavljajo...	Seznam					
	TI=30 <1831	TI=15 <1831	TI=10 <1831	KW=4 <1831	KW=3 <1831	KW=2 <1831
tudi v enem od drugih seznamov	4	9	20	6	28	26
... od tega duplikati	4	9	13	6	17	16
samo v tem seznamu	0	0	7	0	10	0
... od tega duplikati	0	0	0	0	3	0
SKUPAJ***	4	9	27	12	38	26

*** Tj. skupno število zapisov na začetku posameznih seznamov; gre za zapise od 1 do 38 v seznamu KW3 (<1831), ki se pojavljajo samo v enem od navedenih seznamov in katerih ID-ji niso večji od ID-ja 38. zapisa v seznamu KW3 (<1831)

Analiza

Seznam *TI30* (<1831), *TI30* (>=1831), *TI15* (<1831) in *TI15* (>=1831) so bili izdelani na podlagi 63.247 prebranih zapisov iz baze podatkov RETRORK. Seznam *TI10* (<1831), *TI10* (>=1831), *KW4* (<1831), *KW4* (>=1831), *KW3* (<1831), *KW3* (>=1831), *KW2* (<1831) in *KW2* (>=1831) pa so bili izdelani na podlagi baze podatkov RETRORK, ki je bila 21. julija 1999 dopolnjena s približno 7.500 novimi zapisi. V bazi podatkov je bilo vseh prebranih zapisov 70.761

Primerjave, ki sledijo, je treba zato vzeti z nekoliko previdnosti.

Analiza je v grobem potekala v dveh delih. V prvem delu smo poskusno pripravili dva algoritma, tj. algoritma št. 1 in 2. Osnovni vprašanji, na kateri naj bi odgovorili v prvi analizi, sta bili:

1. ali so v bazi podatkov RETRORK potencialni duplikati, ki jih ni v seznamu duplikatov
2. ali so v seznamu duplikatov zapisi, ki niso duplikati



Da bi lahko odgovorili na ti dve vprašanji, je bilo treba podrobneje analizirati bibliografske zapise same in jih primerjati z zapisi v vzajemni bazi podatkov.

Predhodna analiza je pokazala, da s pomočjo avtomatskih postopkov ne bo mogoče zanesljivo identificirati vseh potencialnih duplikatov. Za to so bile krive tipkarske napake v naslovih, razlike v diakritičnih znakih, drugačen prepis ločil, drugače strukturirani zapisi, različna katalogizacijska pravila, napaka pri tipkanju leta izida itd. Ravno tako tudi ni bil možen odgovor na vprašanje, koliko je v seznamu potencialnih duplikatov zapisov, ki niso duplikati. Med takšne zapise namreč spadajo zapisi za faksimila, zapisi za monografske publikacije v več delih ter zapisi, ki so prišli v seznam zaradi krajšanja naslovov in imajo zato enake začetke naslovov itd. Vse potencialne zapise bi lahko zanesljivo identificirali edino le s pomočjo skrajno zamudnega ročnega preverjanja.

Pri nadaljnji analizi in pripravi algoritmov je bilo zatorej najbolj pomembno vprašanje, kako s pomočjo avtomatskih postopkov identificirati čim večje število potencialnih duplikatov, med katerimi bi bilo tudi čim večje število pravih duplikatov. S tem namenom smo izdelali več različnih algoritmov, ki smo jih nato med seboj primerjali.

Za podrobnejšo primerjavo je bilo mogoče upoštevati predvsem seznam *KW3* (<1831), v katerem je bilo največ potencialnih duplikatov, v nekoliko manjši

meri pa tudi seznama *TI10* (<1831) in *KW2* (<1831). Narejena je bila primerjava prvih osemintridesetih zapisov v seznamu *KW3* (<1831) z zapisi v drugih petih seznamih, izdelanih po različnih algoritmih.

Primerjava je pokazala, da so se vsi zapisi, ki so bili tudi v seznamu *KW3* (<1831), nahajali na začetkih posameznih seznamov (zapisi so bili v seznamih razvrščeni naraščajoče po ID-jih). Le v seznamu *TI10* (<1831) je bilo sedem takih zapisov (oz. 25,9 %), ki se niso pojavili tudi v seznamu *KW3* (<1831) oziroma katerem koli drugem seznamu, vendar pa med njimi ni bilo duplikata. Deset zapisov (oz. 26,32 %), ki jih ni bilo v nobenem drugem seznamu, je bilo tudi v seznamu *KW3* (<1831), od tega so bili trije zapisi duplikati oziroma možni duplikati.

Najpopolnejši od šestih seznamov s potencialnimi duplikati je bil seznam *KW3* (<1831), v katerem je bilo od osemintridesetih zapisov kar dvajset duplikatov (oz. 52,46 %). Sledi mu seznam *KW2* (<1831), v katerem je bilo od šestindvajsetih zapisov šestnajst duplikatov (oz. 61,54 %), pri čemer je treba poudariti, da se je ta seznam dejanskih duplikatov v veliki meri ujemal s seznamom *KW3* (<1831). V seznamu *KW3* (<1831) so bili štirje zapisi za duplikate več kot v seznamu *KW2* (<1831) - med temi zapisi je bilo nekaj zapisov, kjer se ni bilo mogoče odločiti, ali gre za duplikate ali ne. Samo v seznamu *KW3* (<1831) so bili med potencialnimi duplikati, ki so bili samo v tem seznamu, tudi pravi duplikati (trije zapisi). Tema dvema seznamoma sledi po popolnosti seznam *TI10* (<1831), kjer je

bilo od skupno sedemindvajsetih zapisov trinajst duplikatov (oz. 48,15 %). V vseh drugih seznamih je bilo število najdenih duplikatov neprimerno manjše.

S prvim, drugim in tretjim algoritmom, ki so bili osnova za izdelavo seznamov *TI30* (<1831), *TI15* (<1831) in *TI10* (<1831), ni bilo mogoče identificirati potencialnih duplikatov, kjer prihaja do napak oziroma nedoslednosti na začetkih naslovov (podpolje 200a), tj. kjer koli med prvim in desetim, petnajstim oziroma tridesetim znakom. Da je bilo takih zapisov veliko, so pokazale primerjave z drugimi seznamami, ki so bili izdelani na osnovi četrtega, petega in šestega algoritma. Značilno zanje je bilo iskanje po ključnih besedah v naslovu namesto po začetku naslova. Algoritmi so se med sabo razlikovali predvsem po številu ključnih besed in po tem, ali so bile ključne besede rezane do določenega števila znakov ali ne. V vseh teh algoritmičnih niso bile upoštevane tudi tiste besede v naslovu, ki so bile krajše od štirih znakov, prav tako pa je bilo izpuščeno tudi iskanje po avtorju (podpolje 700a), kjer običajno prihaja do različnih napak oziroma nepravilnosti. Tako smo se npr. izognili težavam pri nedosledni uporabi členov na začetku naslovov in nepravilnostim v značnicah.

Iz medsebojne primerjave seznamov je bilo razvidno, da so bili, glede na dobljene rezultate, najboljši algoritmi z manjšim številom ključnih besed, le-te pa so bile še dodatno krajšane do določenega števila znakov. Na ta način smo se poskušali kar

nepravilnostim oziroma nedoslednostim, do katerih rado prihaja v bibliografskih zapisih v naslovi, značnicah itd. Zavedali pa smo se, da bo avtomatski postopek odpovedal v vseh tistih primerih, ko bi bila napaka kje v ključnih besedah (oziroma do mesta, kjer smo jo rezali).

Podobne primerjave kot za zapise za gradivo do leta 1831 pa nismo naredili za zapise za gradivo po letu 1831. Končni rezultat bi po naših ocenah ne mogel biti bistveno drugačen, le pri gradivu s kasnejšim letom izida smo predvideli nekoliko večji odstotek zapisov, ki niso duplikati, delno tudi na račun večjega števila ponatisov, faksimiliranih izdaj itd.

Ko smo analizirali zapise v bazi podatkov RETRORK, smo ugotovili, da v nekaterih zapisih manjka leto izdaje (podpolje 100c). Ti zapisi so bili tudi sicer precej nepopolni in jih ne bi bilo mogoče zanesljivo identificirati. Na podlagi algoritma številka 7 smo izdelali seznam *ZAPISI BREZ LETA IZDAJE*, v katerem je bilo 544 oziroma okoli 0,8 % zapisov brez leta izdaje.

Zaključki

Po primerjavi različnih algoritmov in seznamov sta v ožji izbor prišla peti in šesti algoritem, odločili pa smo se za šestega, ki se nam je zdel najustrežnejši. Čeprav je bil seznam *KW2* (<1831), narejen po šestem algoritmu, za okoli dva tisoč zapisov krajši od seznama *KW3* (<1831), ki je bil narejen po petem algoritmu, naj bi bilo v njem po naši oceni tudi manj zapisov, ki niso bili duplikati. Prav tako se je izkazalo, da je



preverjanje po dveh namesto po treh ključnih besedah bolj učinkovito, tj. obstaja manjša verjetnost, da zaradi morebitnih napak v naslovih ne bi našli nekaterih potencialnih duplikatov.

S šestim algoritmom smo tako med 70.761 zapisi, kolikor jih je bilo v bazi podatkov RETRORK, identificirali blizu 15 % (oz. okoli 9.500) takih, ki so bili potencialni duplikati. Od teh 15 % zapisov - potencialnih duplikatov pa naj bi bilo po naših ocenah kar 60 % pravih duplikatov (oz. okoli 5.700). Številke, ki smo jih dobili na ta način, so se nam zdele povsem verjetne in sprejemljive.

Na podlagi vseh teh analiz smo se odločili, da se pri vključitvi baze podatkov RETRORK v vzajemno bazo podatkov COBIB ne vpišejo potencialni duplikati (algoritem št. 6) in zapisi brez leta izdaje (algoritem št. 7). Seznami s temi zapisi pa so bili poslani v NUK za kasnejše ročno preverjanje in vnašanje teh zapisov v vzajemno bazo podatkov.

Viri

1. Žumer, M. Retrospektivna konverzija starih katalogov Narodne in univerzitetne knjižnice. V: Urbanija, J. (ur.) Zbornik razprav : 10 let Oddelka za bibliotekarstvo 1987-1997. Ljubljana : Filozofska fakulteta, Oddelek za bibliotekarstvo, 1998, 215-229.
2. Žumer, M. Positioning National and University Library for the electronic age. V: Library Systems Seminar (23 ; 1999 ; Bled). Managing multimedia collections. Ljubljana : National and University Library, 1999, 64-71. [http://www.konbib.nl/coop/elag/elag99/presentations/](http://www.konbib.nl/coop/elag/elag99/presentations/positioning_nul.html)

[positioning_nul.html](http://www.konbib.nl/coop/elag/elag99/presentations/positioning_nul.html)

3. Štok, B., Vaupotič, M. Retrospective conversions in the COBISS system. V: International Conference on Library Automation in Central and Eastern Europe (1997 ; Budapest). Libraries in the information society : conference proceedings. Luxembourg : Office for Official Publications of the European Communities, 1997, 217-224.
4. Doberšek, M. Konverzija zapisov o raziskovalnih nalogah. V: COBISS obvestila, 1997, 2(4), 16-17.
http://www.izum.si/scripts/ft?l=obvestila/1997_4/cia_97_4_06.html&lang=win
5. Op. cit. ref. 3
6. Op. cit. ref. 1 in 2