DOI:10.3359/oz0704213

strokovni prispevek na konferenci COBISS 2007

# RAZLIČNOST JEZIKOV IN PISAV V SISTEMU COBISS

**Tadeja Brešar**

Institut informacijskih znanosti, Maribor

Kontaktni naslov:
tadeja.bresar@izum.si

**Izvleček**

Avtorica predstavlja projekt Poenotenje uporabe kod za jezike in pisave v COBISS.Net, ki je bil v IZUM-u zaključen v novembru 2007. Zaradi sprememb v šifrantu za jezik je bilo treba izvesti konverzijo kodiranih podatkov o jeziku. Kode za jezik besedila se uporabljajo med drugim za krmiljenje pisave bibliografskih podatkov na različnih prikazih. Ker so se spremenile ravno kode za južnoslovanske jezike in se je sprememba zato nanašala na velik dela knjižničnega gradiva s tega območja, je bilo treba pripraviti nov algoritem za krmiljenje pisave in ga vključiti v vse segmente programske opreme COBISS.

**Ključne besede**

pisave, jeziki, ISO 639-2, konverzije bibliografskih baz podatkov, programska oprema COBISS

**Abstract**

The author introduces the project called Unification of the Use of Codes for Languages and Scripts in COBISS.Net, completed at IZUM in November 2007. As a result of the changes to the list of codes for languages a conversion of encrypted data on the language had to be carried out. The codes for the language of the text are used, inter alia, for script control of bibliographic data in various display formats. With the codes for the South Slavic group of languages being the subject to change in particular, affecting a large part of the library material from this field, a new algorithm for script control had to be developed and integrated into all COBISS software segments.

**Keywords**

scripts, languages, ISO 639-2, conversion of bibliographic databases, COBISS software

## UVOD

V sistemu COBISS sodelujejo knjižnice, ki pri svojem delu uporabljajo različne jezike, pa tudi različne pisave: slovenščino in latinico uporabljamo v Sloveniji; bosanščino in hrvaščino, ki se prav tako pišeta v latinici, uporabljajo v Bosni in Hercegovini; srbščino – ta se lahko piše v cirilici ali latinici – uporabljajo v Srbiji, Črni gori ter Bosni in Hercegovini; v Makedoniji pa uporabljajo makedonščino in cirilico. Ko bo sistem COBISS vzpostavljen tudi v Bolgariji, se bo pojavil še en jezik, ki se piše v cirilici, tj. bolgarščina.

Zaradi različnih jezikov v sistemu je treba poskrbeti za prevode tako dokumentacije kot tudi programske opreme. Paziti je treba tudi pri prevzemanju zapisov, saj se določeni podatki v zapise dodajo v jeziku katalogizacijske ustanove. Če zapis prevzamemo iz drugega jezikovnega okolja, je treba te podatke prevesti. Pri prevzemanju igrajo veliko vlogo kodirani podatki. Ti so v vseh jezikih enaki in so s pomočjo prevedenih šifrantov razumljivi v vseh jezikih.

## UPORABA JEZIKOV IN PISAV V APLIKACIJAH

Ko govorimo o jezikih in pisavah v aplikacijah, mislimo na več stvari. Najprej sta tu jezik in pisava uporabniškega vmesnika. V večini aplikacij COBISS imajo uporabniki na voljo uporabniški vmesnik v domačem jeziku in dodatno še v angleščini. Kar se tiče pisave uporabniškega vmesnika in vnosa, se v programski opremi COBISS2 uporablja izključno latinica. Za srbske in makedonske knjižnice, ki uporabljajo tudi cirilico, sta v cirilici možna izpis na tiskalnik in prikaz v COBISS/OPAC-u.

V bibliografskih in normativnih bazah podatkov se jezik in pisava pojavljata kot podatek v zapisih, in to v več poljih. V bibliografskih zapisih tako vpisujemo jezik in pisavo katalogizacije, različne jezike in pisave, ki se po-

javijo na opisovani enoti, jezike povzetkov in predmetnih oznak, jezike tablic za klasifikacijo, jezike nekaterih naslovov itd. Večinoma so ti podatki kodirani, ni pa nujno. Podobno delamo v normativnih zapisih, le da je v njih teh podatkov manj. Tu zaenkrat vpisujemo jezik in pisavo katalogizacije, jezik avtorja, jezik korporacije in jezik variantne značnice.

## KODIRANI PODATKI O JEZIKU

### ISO 639-2

V zapisih je večina podatkov o jeziku kodiranih. Za normativne in bibliografske zapise se uporablja isti šifrant. Format UNIMARC, na katerem temelji tudi format COMARC, za kodiranje jezika predpisuje uporabo standarda ISO 639-2 *Codes for the representation of names of languages. Part 2, alpha-3 code.*

Standard je nastal leta 1998 na osnovi šifranta *MARC code list for languages*, ki ga vzdržuje Kongresna knjižnica. Ta je pooblaščena za zbiranje in obravnavo predlogov za nove kode in tudi za standard ISO 639-2.

Za bibliografsko kontrolo se uporabljajo tiste kode iz standarda, ki so okrajšane iz angleških izrazov.

### Kode za srbski, hrvaški in bosanski jezik

V šifrantih, ki jih uporabljamo v sistemu COBISS, je žal prišlo do sprememb ravno pri kodah za nekatere jezike, ki se uporabljajo na območju bivše Jugoslavije. Gre za srbski, hrvaški in bosanski jezik. Standard ISO 639-2 trenutno za te jezike predpisuje kode *"bos" – bosanski*, *"scr" – hrvaški* in *"scc" – srbski.*

Dokler smo imeli v bivši državi skupen sistem, smo za te jezike uporabljali kodi *"scr" – srbohrvaški (latinica)* in *"scc" – srbohrvaški (cirilica)*. Leta 1991 smo v IZUM-u od knjižnic prejeli zahtevo za ločeno kodo za hrvaški jezik in tako je bila v šifrant dodana koda *"cro" – hrvaški*. Nato je bila leta 2000 v skladu s standardom dodana še koda za bosanski jezik *"bos" – bosanski*.

Poleg tega so po ločitvi skupnega sistema v več sistemov COBISS v nekaterih državah začeli uporabljati različna pojasnila h kodam "scr" in "scc": v enih državah je pomenilo *"scr" – srbski (latinica)* in *"scc" – srbski (cirilica),* v drugih državah pa so uporabljali kodi "*scr" – srbski ali hrvaški (latinica)*, *"scc" – srbski ali hrvaški (cirilica)*.

Težav je bilo torej več. Šifrant za jezike, ki ga uporabljamo v sistemu COBISS, ni bil skladen s standardom, povrhu pa se je v različnih državah uporabljal različno. Do razlik je prihajalo ravno pri jezikih, v katerih je napisan velik delež gradiva, ki ga imajo knjižnice s tega območja. Zato smo v IZUM-u predlagali, da kode za jezike v sistemih COBISS poenotimo s standardom.

## PROJEKT *POENOTENJE UPORABE KOD ZA JEZIKE IN PISAVE V COBISS.NET*

### Časovni pregled

Projekt smo prvič uvrstili v letni program dejavnosti IZUM-a leta 2005. V resnici pa se je delo na projektu začelo že v letu prej. Takrat smo pripravili prvo specifikacijo sprememb v bibliografskih bazah in dopolnitev programske opreme. Oktobra 2004 smo predlagane spremembe opisali in dopis naslovili na nacionalne knjižnice. Le-te so se s predlogom strinjale, zato smo decembra 2004 o načrtovanih spremembah obvestili še vse knjižnice, ki sodelujejo v slovenskem sistemu COBISS.SI.

V obdobju 2005–2007 je potekalo načrtovanje nove verzije programske opreme. V začetku junija leta 2007 smo razvoj zaključili do te mere, da smo nekatere segmente lahko ponudili knjižnicam v testiranje. Testiranje je bilo relevantno predvsem za države, v katerih uporabljajo cirilico. Tem smo testiranje najavili že aprila 2007, kot rok za oddajo pripomb pa smo postavili konec avgusta. Oktobra 2007 smo v vseh sistemih, razen v slovenskem, testirali še uporabniški vmesnik COBISS/OPAC.

Ob koncu tedna, 3. in 4. novembra 2007, smo izvedli konverzijo vseh bibliografskih baz podatkov in namestili novo verzijo programske opreme.

### Naloge projekta

Pred uskladitvijo pomena kod "scr" in "scc" s standardom ISO 639-2 je bila informacija o pisavi skrita kar v kodi za jezik. S kodo "scr" smo namreč označevali le besedila v latinici, s kodo "scc" pa le besedila v cirilici. Programska oprema je bila zasnovana tako, da je bil v nekaterih primerih podatek o jeziku bistven za krmiljenje pisave prikaza.

Z načrtovano spremembo to ni bilo več res. V skladu s standardom ISO 639-2 se koda "scc" uporablja za srbski jezik, ne glede na to, ali je besedilo pisano v cirilici ali latinici. Zato smo morali poleg drugih dopolnitev programske opreme in konverzije pripraviti tudi nov algoritem za krmiljenje pisave.

Naloge, ki jih je bilo v okviru projekta *Poenotenje uporabe kod za jezike in pisave v COBISS.Net* treba opraviti, smo razdelili v pet večjih sklopov:

- konverzija vseh bibliografskih baz,
- nov algoritem za krmiljenje pisave prikaza podatkov,
- vključitev novega algoritma v vse segmente programske opreme COBISS,
- dopolnitve segmentov COBISS2/Katalogizacija, COBISS2/Izpisi in COBISS/OPAC,
- uporabniški vmesnik COBISS/OPAC-a v cirilici za vse države, ki uporabljajo cirilico.

## Algoritem za krmiljenje pisave

Algoritem za krmiljenje pisave podatkov smo skušali zasnovati tako, da bi katalogizatorjem vnos čim bolj olajšali. Upoštevali smo različne kombinacije kodiranih podatkov, iz katerih je možno razbrati, v kateri pisavi je treba prikazati posamezne podatkovne elemente. Algoritem za vsako podpolje posebej prepozna, kateri kodirani podatki so relevantni za njegov pravilen prikaz. Zato pri obdelavi gradiva večinoma ni treba paziti na to, v kateri pisavi se morajo podatki izpisati, ampak je treba zapis le pravilno kodirati.

Algoritem preverja naslednje podatke:

- *100i – Koda za transliteracijo*,
- *100h – Jezik katalogizacije*,
- *0017 – Pisava katalogizacije*,
- *101a – Jezik besedila*,
- *100l – Pisava stvarnega naslova*,
- podpolje *z – Jezik* v poljih za vsebinsko obdelavo in naslove,
- indikator v podatkih o zalogi.

V nekaterih primerih zgolj s kodiranimi podatki ni možno zagotoviti pravilnega prikaza, predvsem takrat, ko so v zapisu citirani podatki v drugem jeziku kot preostalo besedilo. Tu mora katalogizator drugo pisavo vklopiti s pomočjo kontrolnega znaka za vklop latinice ali cirilice.

Na pisavo prikaza vplivajo tudi nastavitve uporabniškega vmesnika. Če v uporabniškem vmesniku nastavimo prikaz v latinici, bodo besedila uporabniškega vmesnika in vsi podatki iz baze prikazani v latinici. Če nastavimo prikaz v cirilici, dobimo besedila uporabniškega vmesnika prikazana v cirilici, podatki iz baze pa se prikažejo v latinici ali cirilici glede na kodirane podatke, ki jih preveri algoritem za krmiljenje pisave.

Prilagamo primer bibliografskega zapisa. Kot vidimo, je celoten zapis vnesen v latinici. Poudarjeni so tisti podatki, na osnovi katerih se za posamezne dele zapisa določi ustrezna pisava za prikaz.

```
001  [a]n [b]a [c]m [d]0 [7]cc
010  [a]5-02-027215-9
100  [c]1989 [e]k [h]mac [i]b1 [l]ca
101 0 [a]rus [a]eng
102  [a]rus
105  [a]a [b]z [c]0
200 1 [a]Paleolit Kavkaza i Sever
       noî Azii [d]L_F≠The≠palaeolithic
       of Caucasus and Northern Asia
       [f][otv. redaktor Pavel Iosifo
       vič Boriskovskiî]
210  [a]Leningrad [c]Nauka [d]1989
215  [a]264 str. [c]ilustr. [d]27 sm
225 1 [a]Paleolit mira eissledovani
    ±ja po arheologii drevnego
       kamennogo veka [d]L_F≠The ≠old
       stone age of the world [e]L_F
       studies in the palaeolithic
       cultures
300  [a]Tekst na rus. i angl. jazik
320  [a]Bibliografija: str. 244-[254]
320  [a]Registri
510 1 [a]≠The ≠old stone age of the
       world [z]eng
606 1 [a]Arheološki naodi [y]Kavkaz
       [z]Paleolit
606  [a]Arheološki naodi [y]Severna
       Azija [z]Paleolit
675  [a]903(479)"632"
702 11 [a]Boriskovskiî [b]Pavel Iosi
       fovič [4]345
```

Gre za rusko knjigo, zapis pa je kataloziziran v makedonščini. Zato se morajo podatki prikazati v cirilici, kar omogočajo pravilno izbrane kode za pisavo v zapisu. Izjema so deli zapisa, ki so v angleščini. Ti morajo biti prikazani v latinici. V tem primeru mora katalogizator v podpolja z angleškim besedilom dodati kontrolni znak za vklop latinice. Prilagamo še prikaz tako urejenega zapisa:

ПАЛЕОЛИТ Кавказа и Северной Азии = The palaeolithic of Caucasus and Northern Asia / [отв. редактор Павел Иосифивич Борисковский]. - Ленинград : Наука, 1989. - 264 стр. : илустр. ; 27 см. - (Палеолит мира : исследования по археологии древнего каменного века = The old stone age of the world : studies in the palaeolithic cultures)

Текст на рус. и англ. јазик. - Библиографија: стр. 244-[254]. - Регистри

ISBN 5-02-027215-9

1. Насп. ств. насл. 2. Борисковский, Павел Иосифович

а) Археолошки наоди - Кавказ - Палеолит б) Археолошки наоди - Северна Азија - Палеолит

903(479)"632"

## Konverzija

Ob zaključku projekta smo izvedli konverzijo vseh bibliografskih baz v vseh sistemih COBISS. Poleg podatkov, ki jih je bilo treba urediti zaradi sprememb v šifrantu za jezike, smo s konverzijo uredili še nekatere druge podatke, predvsem tiste, ki smo jih kdaj v preteklosti konvertirali le v Sloveniji, ne pa tudi v preostalih sistemih. V nadaljevanju naštevamo najpomembnejše korake konverzije.

Glede na zadnjega redaktorja sta se v zapise vpisala pisava in jezik katalogizacije, ki ju uporablja posamezna knjižnica. V zapise knjižnic, ki uporabljajo cirilico, se je vpisala tudi ustrezna koda za transliteracijo. Dodana je bila pisava stvarnega naslova, in to na osnovi kode za jezik enote. V vseh podatkovnih elementih, ki so kodirani s šifrantom za jezik, je bila koda "cro", kar je neveljavna koda za hrvaški jezik, nadomeščena s kodo "scr", ki se za hrvaški jezik uporablja po standardu. Če je delo izšlo na območju Srbije, Črne gore ali Republike Srpske, je bila koda "scr", ki se je prej uporabljala za srbski jezik, pisan v latinici, in za hrvaški jezik, v polju za jezik enote nadomeščena s kodo "scc" za srbski jezik. Kodi za državi Jugoslavijo ter Srbijo in Črno goro sta bili spremenjeni v kodo za ustrezno obstoječo državo, in to na osnovi kode za regijo in na osnovi podatka o kraju izida.

Za posamezne sisteme COBISS smo pripravili seznam tistih avtorjev, pri katerih je jezik njihovih del v bibliografskih zapisih kodiran kot bosanski, hrvaški ali srbski, vendar ta koda ni v vseh zapisih v COBIB-u enaka. Sezname smo poslali v nacionalne knjižnice. Te so jih pregledale in za posamezne avtorje, kjer je to bilo možno, označile, v katerem od teh treh jezikov dejansko pišejo. Za take avtorje smo kodo za jezik besedila v bibliografskih zapisih za njihova dela programsko poenotili.

Konverzija glede na posamezne avtorje je bila opcijska. Zanjo so se odločili v Srbiji, Črni gori in Makedoniji, v Sloveniji ter Bosni in Hercegovini pa tega dela konverzije nismo izvedli.

V zapisih knjižnic, ki uporabljajo cirilične signature, je bilo treba izvesti še konverzijo indikatorja v podatkih o zalogi. Ta del konverzije je bilo treba pripraviti individualno za vsako knjižnico posebej.

## Dopolnitve programske opreme

Nov algoritem za krmiljenje pisave je bilo treba vključiti v vse glavne segmente programske opreme COBISS2 in COBISS/OPAC. Poleg tega smo v nekatere segmente dodali še druge nove funkcije.

V segment COBISS2/Katalogizacija smo dodali nekaj novih programskih kontrol, ki preverjajo pravilnost kod in prisotnost obveznih podatkov, predvsem tistih, ki so pomembni za pravilen izpis v ustrezni pisavi. Uredili smo privzete vrednosti za jezik in pisavo katalogizacije in za indikator v podatkih o zalogi. Novo je tudi to, da se pri prevzemanju zapisov v srbske lokalne baze podatkov značnica avtomatsko vpiše v pravi (fonetični ali etimološki) obliki, odvisno od tega, v kateri pisavi knjižnica vodi svoj katalog.

V vse segmente programske opreme smo dodali prenos zapisov v naboru znakov UNICODE. Inventarne knjige smo pripravili tudi v cirilici, saj so bile prej na voljo le v latinici. Dopolnili smo parametrizacijo pri pripravi bibliografij, tako da lahko knjižnice ob zlaganju bibliografij sproti izberejo, ali želijo izpis v cirilici ali latinici. Tudi spletne bibliografije so zdaj na voljo v latinici in cirilici.

V sistemih, v katerih uporabljajo cirilico, je bil precej dopolnjen tudi COBISS/OPAC. Pred projektom *Poenotenje uporabe kod za jezike in pisave v COBISS.Net* so imeli cirilični vmesnik na voljo le v Makedoniji, kar je bilo nujno, saj se makedonščina vedno piše v cirilici. Od novembra 2007 imajo tudi v Srbiji, Bosni in Hercegovini ter Črni gori možnost izbire vmesnika v cirilici. V teh sistemih je pri iskanju možno dodatno omejevanje po pisavi, v kateri je natisnjeno gradivo. V vseh sistemih COBISS pa je v novi verziji COBISS/OPAC omogočeno omejevanje po srbskem, bosanskem in hrvaškem jeziku hkrati.

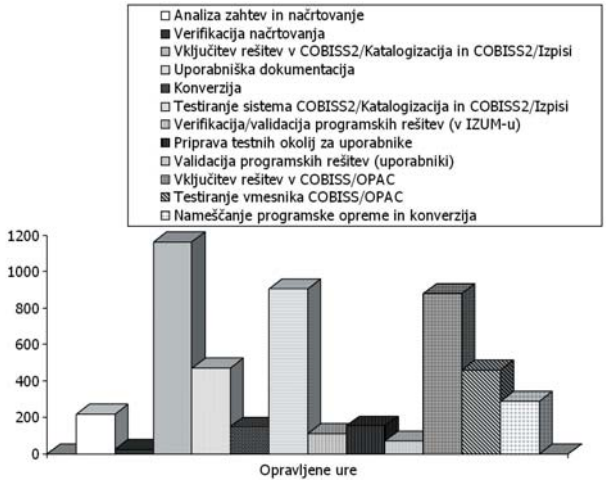
## Podatki o opravljenem delu

Projekt *Poenotenje uporabe kod za jezike in pisave v COBISS.Net* je bil zelo zahteven, tako z vsebinskega vidika kot tudi z vidika potrebnih človeških virov. Za ilustracijo navajamo podatke o opravljenem delu na tem projektu za leto 2007.

Pri projektu je sodelovalo 38 sodelavcev. Vsi skupaj so v letu 2007 opravili 4.803 delovne ure. Od tega je 19 so-

delavcev pri projektu imelo le manjše zadolžitve; vsak od teh je opravil za manj kot teden dni dela. Po drugi strani pa je šest sodelavcev opravilo za več kot 2 meseca dela vsak. Največ časa se je projektu posvečal sodelavec, ki je bil zadolžen za implementacijo nove verzije COBISS/OPAC, ta je za projekt porabil nekaj čez 550 delovnih ur.

Razporeditev opravljenih ur dela po izvedenih aktivnostih je razvidna iz slike 1.




Slika 1: Razporeditev opravljenih ur po aktivnostih

Vidimo, da je največ časa zahtevala implementacija programske opreme. Na drugem mestu je čas, porabljen za testiranje nove in spremenjene funkcionalnosti. Sledi priprava uporabniške dokumentacije. Zanimiv je zadnji, desni stolpec, ki ponazarja čas nameščanja programske opreme. Gre za 287 ur, ki so bile večinoma opravljene v zadnjih nekaj dnevih izvajanja projekta, predvsem čez vikend, ko sistem COBISS uporablja najmanj uporabnikov.

## NAČRTI

### Dopolnjevanje šifranta za jezike

Trenutno poteka revizija šifranta jezikov še za druge jezike, ne le za jezike naše bivše skupne države. Pričakujemo le manjše dopolnitve, predvsem zaradi kod, ki so bile v zadnjem času dodane v standard ISO 639-2 in jih v šifrantu COBISS še ni.

Pri dopolnjevanju šifranta se že nekaj časa držimo načela, da kod, ki jih v standardu ni, ne dodajamo! Pri jezikih, ki v šifrantu morebiti manjkajo, je treba najprej poskrbeti, da se uvrstijo v standard. Obrazec za zahtevo nove kode je na voljo na spletnih straneh Kongresne knjižnice na naslovu http://www.loc.gov/standards/iso639-2/.

### COBISS3/Katalogizacija

V pripravi je nov segment programske opreme COBISS3, COBISS3/Katalogizacija. Tehnologija COBISS3 bo omogočala vnos podatkov tudi v cirilici. Zato pri prikazu podatkov ne bo treba uporabljati nobenih algoritmov, ki bi skrbeli za pravilno pisavo.

Algoritem, ki je bil pripravljen za prikaz podatkov iz programske opreme COBISS2, pa bomo uporabili pri prenosu zapisov, kreiranih s COBISS2/Katalogizacija, v COBISS3/Katalogizacija. Z njim bo možno narediti avtomatsko pretvorbo podatkov iz latinice v cirilico. K pravilnemu prenosu podatkov bo prispevala tudi konverzija, ki smo jo izvedli v okviru projekta *Poenotenje uporabe kod za jezike in pisave v COBISS.Net*. S konverzijo smo podatke uredili in s tem izpolnili enega od osnovnih pogojev za uspešen prehod v okolje COBISS3.

### Reference

[1] UNIMARC Manual. Bibliographic Format, München: K. G. Sauer, 1994–.

[2] COMARC/B format: format za bibliografske podatke: priročnik za uporabnike. Maribor: IZUM, 1991–.

[3] COMARC/A format: za bibliografske podatke. Maribor: IZUM, 2003.

[4] MARC code list for languages. Washington: Library of Congress, Cataloging Distribution Service, 2007.

[5] International standard. ISO 639-2, Codes for the representation of names of languages. Part 2, Alpha-3 code. Geneve: International Organization for Standardization, 1998.

### Spletne povezave

1 http://www.loc.gov/standards/iso639-2/

2 http://www.loc.gov/marc/languages/langhome.html