Sistem za automatsko konvertovanje pravnih akata Republike Srbije u Akoma-Ntoso format podataka

Tomislav Dobrički
Softversko inženjerstvo i informacione tehnologije
Fakultet tehničkih nauka, Univerzitet u Novom Sadu
dtoma95@uns.ac.rs

***Abstrakt*—Akoma Ntoso (*eng. Architecture for Knowledge-Oriented Management of African Normative Texts using Open Standards and Ontologies*) je standard koji omogućuje opisivanje pravnih dokumenata. Njegova glavna namena jeste da olakšava razmenu citiranje i refernciranje pravnih dokumenata između rezličitih parlamenata, odnosno država, i da definiše opštu strukturu pravnih dokumenata kao osnovu za dalji razvoj informacionih sistema. U ovom radu je pradstavljen sistem osnovan na pravilima (*eng. rule-based sytem*) koji od akata napisanih u Republici Srbiji kreira XML (eng. *eXtensible Markup Language*) datoteku po Akoma Ntoso XML šeme. Specifičnije, opisani sistem iz akata kao što su zakoni, odluke, pravilnici, propisi itd. izdvaja podatke u tri sloja: sloj metapodataka, strukturalni sloj i tekstualni sloj. Za prepoznavanje značajnih elemenata unutar dokumenta, sistem koristi REGEX (eng. *Regular expression*) paterne. Rezultat ovog rada jeste skup XML datoteka generisanih na osnovu, ukupno oko 3000, pravnih akata preuzetih sa Pravno-informacionog sistema Republike Srbije. Ovi dokumenti se dalje mogu koristiti za razvoj drugih sistema, dok sistem opisan u ovom radu može dalje da se unapređuje za primenu na raznovrsniji skup pravnih dokumenata.**

Ključne reči—akoma ntoso; sistem baziran na pravilima; REGEX; XML; pravi akt; zakon; propis;

I. Uvod

Akoma Ntoso (*Architecture for Knowledge-Oriented Management of African Normative Texts using Open Standards and Ontologies*) je okvir (eng. *framework*) razvijen uz podršku Ujedinjenih Nacija, sa ciljem unapređenja informacionih sistema parlamenata u Africi. Jedna od osnovnih uloga koju Akoma Ntoso ispunjava jeste definisanje zajedničkog standard za razmenu podataka između parlamenata, kao i olakšavanje procesa citiranja i referensiranja dokumenata različitih parlamentarnih sistema [1]. Dodatno, Akoma Ntoso može da služi kao osnovi model dokumenata za dalji razvoj informacionih sistema u kontekstu rada parlamenata. Da bi iskoristili pogodnosti Akoma Ntoso okvira potrebno je razviti bazu dokumenata zapisanih na osnovu Akoma Ntoso XML (*eXtensible Markup Language*) šeme.

Ručno konvertovanje akata je spor i nepraktičan proces, pre svega i zbog velikog broja podataka. Do sada je u Službenom glasniku Republike Srbije objavljeno više od 16000 propisa i ostalih akata [2]. Zbog toga je potrebno razviti sistem za automatsko konvertovanje pravih akata. Sistem treba da bude sposoban da izdvoji značajne informacije koje Akoma Ntoso opisuje. Ovi podaci su podeljeni u tri sloja: sloj metapodataka, strukturalni sloj i tekstualni sloj [3]. Prilikom izrade sistema, potrebno je izdvojiti elemente pravnih akata Republike Srbije koji su ekvivalenti elementima koji su specifikovani unutar Akoma Ntoso XML šeme. Dodatno, sistem mora da bude sposoban da pronalazi i prepoznaje pomenute elemente unutar dokumenta.

Pošto su pravni akti napisani formalnim jezikom, i u većini slučajeva prate predefinisanu strukturu, dobar izbor tipa sistema za konvertovanje dokumenata jesu sistemi bazirani na pravilima. U ovom radu je predstavljena implementacija ovakvog sistema, koja je sposbna da kreira Akoma Ntoso XML dokumente na osnovu podataka preuzetih sa Pravno-informacionog sistema Republike Srbije.

Ovaj rad je ogranizovan na sledeći način. U sekciji II je opisan preduzeti proces za prikupljanje i prečišćavanje podataka za dalji rad sistema. U sekciji III je opisana metodologija izvlačenja metapodataka, i njihovo preslikavanje na Akoma Ntoso strukturu. U sekciji IV je opisana hijerarhijska struktura samih akata, pored toga je predstavljeno rešenje za konvertovanje pomenute strukture u Akoma Ntoso format. U sekciji V se spominju elementi tekstualnog sloja koje sistem implementira, to su, pre svega, reference unutar teksta. I na kraju rada je u sekciji VI kratka sumarizacija rada, uz pominjanje nekih mana sistema opisanog u radu i predloženi koraci koji se mogu preuzeti za dalji rad na ovu temu.

II. Priprema podataka

U ovoj sekciji je opisan način pretprocesiranja pravnih akata. Za svrhe ovog rada, korišćen je skup javno dostupnih akata na internet stranici Pravno-informacionog sistema Republike Srbije. Ukupno je skinuto 3133 akata, što pretežno čine zakoni, odluke, pravilnici i drugi tipovi propisa. Tekst akata je predstavljen u HTML (*Hypertext Markup Language*) formatu, ali pored toga Pravno-informacioni sistem nudi i skup dodatnih informacija „o aktu“. U ove informacije spadaju stavke kao što su značajni datumi, dodatne napomene, kategorije kojima akt pripada itd.

Zbog nekonzistente strukture HTML dokumenata i zbog čestih sintaksnih grešaka, potrebno je izvršiti pretprocesiranje istih. U ovom slučaju se pretprocesiranje vrši radi pripreme dokumenata za lakšu dalju upotrebu. U Akoma Ntoso okviru, preporučuje se korišćenje nekih od

HTML elemenata za struktuiranje podataka na tekstualnom nivou. Tako da, prilikom pretprocesiranja, poželjno je sačuvati samo osnovne HTML elemente:

- Paragraf (<*p*>)
- Tabela i njeni podelementi (<*table*>, <*th*>, <*tr*> i <*td*>)
- Slika (<*img*>)

Svi ostali tipovi elementa su uklonjeni iz datoteke. Isto tako su uklonjeni svi stilistički atributi i zamenjene su tekstualne vrednosti koje su isključive HTML-u (kao * *), a ne koriste se u XML-u.

III. IZDVAJANJE METAPODATAKA

U Akoma Ntoso okviru postoji definisani skup metapodataka koji služe za opisivanje dokumenata. Okvir je namenjen za opisivanje više tipova pravnih dokumenata, ne samo akata [4]. Zbog toga je potrebno izdvojiti elemente metapodataka koji su relevantni za pravne akte u Republici Srbiji. U ovoj sekciji su opisani elementi koje sistem generiše i način na koji zaključuje njihovu vrednosti iz opisa akta na veb sajtu Pravno-informacionog sistema. U narednim podsekcijama je redom opisan proces za Akoma Ntoso elemente: *identification, publication, workflow, classification, notes i lifecycle.*

A. *Identifikacija*

Identifikacioni blok je predstavljen po FRBR (*Functional Requirements for Bibliographic Records*) modelu. FRBR je klasifikacioni standard za bibliografske zapise. FRBR model je podeljen na četiri sloja, svaki od kojih predstavlja poseban nivo abstrakcije dokumenta [5]. Sistem opisan u ovom radu implementira prva tri sloja:

- delo (eng. *work*) – abstraktni koncep akta
- izraz (eng. *expression*) – posebna verzija akta koja se razlikuje po nekoj osnovi (npr. jezik ili verzija)
- manifestacija (eng. *manifestation*) – poseban format ekspresije, u ovom slučaju je uvek Akoma Ntoso XML datoteka.

Poslednji nivo FRBR modela je stavka (eng. *item*), i odnosi se na jedinstvenu fizičku manifestaciju dokumenta. U ovom slučaju bi to bio jedna datoteka u memoriji računara. Prilikom generisanja XML datoteke, ovaj sloj je izostavljen zato što je previše nizak nivo abstrakcije i ima redundantu ulog unutar Akoma Ntoso formata. Na Slici 1 je dat primer izgenerisanog identifikacionog bloka.

```
<identification source="#somebody">
    <FRBRWork>
        <FRBRthis value="/rs/act/2002/62-42/main"/>
        <FRBRuri value="/rs/act/2002/62-42"/>
        <FRBRdate date="2002" name="Generation"/>
        <FRBRauthor as="#author" href="#somebody"/>
        <FRBRcountry value="rs"/>
    </FRBRWork>
    <FRBRExpression>
        <FRBRthis value="/rs/act/2009/29-17/srp@/main"/>
        <FRBRuri value="/rs/act/2009/29-17/srp@"/>
        <FRBRdate date="2009" name="Generation"/>
        <FRBRauthor as="#editor" href="#somebody"/>
        <FRBRlanguage language="srp"/>
    </FRBRExpression>
    <FRBRManifestation>
        <FRBRthis value="/rs/act/2009/29-17/srp@/main.xml"/>
        <FRBRuri value="/rs/act/2009/29-17/srp@.akn"/>
        <FRBRdate date="2009" name="Generation"/>
        <FRBRauthor as="#editor" href="#somebody"/>
        <FRBRformat value="xml"/>
    </FRBRManifestation>
</identification>
```

Slika 1. Primer identifikacionog bloka za jedan Pravilnik

Podaci koji su potrebni za formiranje identifikacionog bloka jesu datum usvajanja, verzija i jezik akta. Jezik se zapisanu po ISO 639-2 kodu [6], što znači da će za akte napisane na srpskom jeziku, jezik akta uvek imati vrednost "srp", nezavisno od upotrebe latiničnog ili ćiriličnog pisma [7]. Datumi i verzija se preuzimaju iz naslova akta, gde su nabrojani brojevi službenog glasnika u kojima se on pojavljuje. Za nivo dela se uzimaju vrednost najstarijeg, dok se za nivoe izraza i manifestacije uzimaju vrednosti najnovijeg broja[1].

B. *Publkacija, radni tok, klasifikacija i beleške*

Vrednosti za elemente Akoma Ntoso sloja metapodataka *publication, workflow, classification i notes* se direktno preuzima sa veb sajta Pravno-informacionog sistema. U tabeli 1. su navedena polja na veb stranici koja odgovaraju svakom od elemenata.

TABELA I. PRESLIKAVANJE VREDNOSTI PREUZETIH SA PRAVNO-INFORMACIONOG SISTEMU U AKOMA NTOSO ELEMENTE

Naziv Elementa	**Polje na veb sajtu Pravno-informacionog sistema**
publication	„Glasilo i datum objavljivanja"
workflow	„Datum stupanja na snagu osnovnog teksa", „Datum primene" i „Datum usvajanja"
classification	„Vrsta propisa", „Oblast" i „Grupa"
notes	„Napomena izdavača" i „Dodatne informacije"

C. *Životni ciklus*

U *lifecycle* element se navode datumi izmena nad tekstom akta. Zbog nedostatka informacija dostupnih vezano za dokumente, nije moguće izvući tačan datum izmena akta. Zbog toga se za životni ciklus akta navode brojevi službenog glasnika koji su navedeni u naslova dokumenta. Uz dodatno znanje o tačnom datumu objavljivanje svakog broja službenog glasnika, bilo bi moguće proširiti ovaj deo sistema konvertovanja dokumenta.

IV. HIJERARHIJSKA STRUKTURA

Strukturalni sloj u Akoma Ntoso okviru daje značenje blokovima teksta unutar pravnog dokumenta. U ovoj sekciji je opisan način na koji sistem baziran na znanju izdvaja značajne delove pravnog akta koristeći pravila zasnovana na REGEX (eng. *Regular expression*) paternima. Pravila su napisana po uzoru na Jedinstvena metodološka pravila za izradu propisa [8]. Sistem deli akt na početni deo (preambula, naslov itd.) i telo rada. U narednim potpoglavljima je opisan način na koji se generišu elementi za oba ova dela.

A. *Početni deo*

Preamblua propisa u Republici Srbiji se nalazi iznad naslova, na samom početku dokumenta [8], preambula se

[1] Prilikom navođenja broja službenog glasnika, retko se navodi pun datum već samo godina. Zbog toga se u identifikaciji navodi samo godina izdavanja.

smešta u „*preambule*“ Akoma Ntoso element. Preambulu čine obični paragrafi teksta. Nakon preambule sledi rečenica o autoreitetu koji donosi akt („*authority*“), nakon čega sledi sam naslov akta („*title*“). Posle naslova se nalazi informacija o broju službenog glasnika i datuma njegovog objavljivanja. Iako se u ovom delu teksta ne nalazi samo datum, on se smešta u „*date*“ element zbog nedostatka adekvatnije nazvanog elementa. Na slici 2 je prikazan primer izgenerisanog početnog dela pravnog dokumenta.

```
<preamble>
    <p>Полазећи од државне традиције српског народа и
       равноправности свих грађана и  етничких заједница у Србији,</p>
    <p>полазећи и од тога да је Покрајина Косово и Метохија саставни
       део територије  Србије, да има положај суштинске аутономије
       у оквиру суверене државе Србије и да  из таквог положаја
       Покрајине Косово и Метохија следе уставне обавезе свих
       државних органа да заступају и штите државне интересе Србије на
       Косову иМетохији у свим унутрашњим и спољним политичким односима,</p>
</preamble>
<preface>
    <authority>грађани Србије доносе</authority>
    <title>УСТАВ РЕПУБЛИКЕ СРБИЈЕ</title>
    <date>„Службени гласник РС“, број  98 од 10. новембра 2006.</date>
</preface>
```

Slika 2 Početni deo Ustava Republike Srbije, predstavljen u Akoma Ntoso XML formatu

B. Telo

Telo akata čini niz hijerarhijskih elemenata. Osnovni klasifikacioni element akata je član [8]. Klasifikacione jedice koje su šire od člana su deo, glava, odeljak i pododeljak. Unutrašnju strukturu člana čine stavovi, tačke, podtačke i alineje. Svaka od klasifikacionih jedinica ima jedinstven skup pravila po kojima se pišu. Zbog toga je moguće pomoću REGEX-a prepoznati koji tip elementa predstavlja jedna rečenica ili red unutar dokumenta.

Akoma Ntoso nudi širok skup strukturalnih elemenata koji mogu da se koriste prilikom kreiranja hijerarhije unutar XML datoteke. U tabeli 2 je za svaku od klasifikacionih jedinica predstavljen Akoma Ntoso element koji je odabran kao njihov ekvivalent [9]. Pored toga je dat i primer izgleda tekstualnog sadržaja elementa unutar akta, zajedno sa REGEX-om uz pomoć kojeg se on prepoznaje.

Potrebno je naglasiti da deo i glava mogu da imaju naslov ispod, a član iznad svoje oznake. Sistem baziran na pravilima mora da bude sposoban da prepoznaje postojanje naslova na osnovu redosleda elemenata. Naslovi se razlikuju od stavova po tome što, po pravilu, nemaju znak interpunkcije na kraju, dok stavovi imaju [8]. Ali ovo često nije slučaj prilikom pojava neregularnost u dokumentima.

Dodatan problem nastaje zbog čestih grešaka, ili drugačijeg korišćenja klasifikacionih jedinica, prilikom pisanja dokumenata. Pošto se tačke ponekad pišu kao odeljci, izostavljaju tačke na krjau stavova itd. Sistem baziran na znanju mora da bude robustan i da prepoznaje što više ovakvih razlika, zbog toga su uvedena dodatna pravila i dodatni paterni za često uočene razlike u pisanju. Isto tako, poseban tip propisa čine odluke, kod kojih je osnovna klasifikaciona jedinica tačka, a ne član [8]. Zbog toga je uveden poseban set pravila za akte koji koriste ovakvu strukturu.

TABELA II. PRIKAZ SVIH KLASIFIKACIONIH JEDINICA PRAVNIH AKATA, ZAJEDNO ZA EKVIVALENTIN AKOMA NTOSO ELEMENTOM [9], PRIMEROM I ODGOVARAJUČIM REGEX-OM.[2]

Naziv Klasifikacione Jedinice	**Ekvivalentni Akoma Ntoso element**	**Primer izgleda unutar dokumenta**		**Odgovarajući REGEX**	
deo	*part*	PRVI DEO		(.*) ДЕО	
glava	*chapter*	Glava I	I. UVODNE ODREDBE	Глава [MDCLXVI]+	[MDCLXVI]+(\.) .*
odeljak	*section*	1. Prelazne odredbe		([0-9]+)(\.)(.*)	
podeodeljak	*subsection*	a) Ovlašćenja nadležnog organa		(a-z)(\))(.*)	
član	*article*	Član 21.		(Члан) ([0-9]+)(\.)	
stav	*clause*	Pravni poredak je jedinstven.			
tačka	*point*	1) Svaka tačka počinje novim redom.		([0-9]+)(\))(.*)	
podtačka	*item*	(1) Svaka podtačka počinje novim redom.		(\()([0-9]+)(\))(.*)	
alineja	*alinea*	- Svaka alineja počinje novim redom.		(– ?\s?)(.*)	

V. TEKSTUALNI SLOJ

Akoma Ntoso elementi tekstualnog sloja služe za obuhvatanje teksta radi urednijeg strukturinja ili isticanja bitnih informacija. Paragrafi, tabele i slike su preuzeti iz izvorne HTML datoteke. Tabele i slike se tretiraju kao zasebni stav. Pored HTML elemenata, sistem prepoznaje i reference unutar teksta dokumenta. Reference se mogu nalaziti na bilo kojem hijerarhijskom nivou. Reference mogu da ukazuju na određeni hijerarhijski element unutar istog akta, ili na delove nekog drugog akta.

Sistem prepoznaje reference putem REGEX-a koji su napisani po uzoru na pravila iz člana 36. Jedinstvenih metodoloških pravila za izradu propisa [8]. Reference se pronalaze na nivou celog dokumenta, nakon što su svi ostali delovi dokumenta izgenerisani. Drugi aktovi se često referenciraju na osnovu imena, zbog toga problem samog uvezivanja referenci sa odgovarajućim identifikatorima postaje komplikovaniji, potreban je poptpuni skup akata. S toga, povezivanje dokumenata putem refernci je implementirano samo u slučajevima gde

[2] Napomena: Dati REGEX-i su u pojednostavljenom obliku, pošto se pravila pisanja klasifikacionih jedinica često ne poštuju striktno. S toga, moraju da se Koriste za nijansu kompleksniji REGEX-i.

se referenciraju delovi istog rada ili je eksplicitno naveden broj službenog glasnika u kojem je akt objavljen. Na slici 3 je prikazana referenca sa navedenim brojem sližbenog lista.

```
<ref id="ref8" href="/rs/act/2000/23/srp@">
    Службени лист СРЈ", бр. 23/2000</ref>
```

Slika 3 Primer reference sa brojem I godinom izdanja službenog lista

VI. ZAKLJUČAK

U prethodnim sekcijama je opisana metodologija razvoja sistema za generisanje Akoma Ntoso XML datoteke za pravne akte u Republici Srbiji. Sistem se najviše oslanja na REGEX paterne za prepoznavanje pojedinih elementa u tekstu akata. Koristi se skup pravila koji su potrebni za rešavanje složenijih problema kao što su pronalaženje naslova i početka određenih delova dokumenta.

Glavni izazov prilikom razvijanja ovakvog sistema se nalazi u nekonzistentnosti u pisanju pravnih akata. Iako su pravila i REGEX-i pisani po uzoru na formalno propisanih pravila, oni često nisu dovoljni. Proširivanjem sistema da obrđuje granične slučajeve koji mogu da se pojave nailazi se na prekopleksan i, u budućnosti, neodrživ sistem pravila.

Rezultat ovog rada jeste skup Akoma Ntoso XML datoteka koji se dalje mogu koristiti i unapređivati za razvoj drugih informacionih sistema. Dalji rad bi mogao da podrazumeva kreiranje baze uvezanih dokumenata, što bi omogućilo razrešivanje više tipova referenci unutar teksta akata. Sistem opisan u radu se može unapređivati daljim uvođenjem pravila za neke od retkih slučajeva i struktura unutar akata. Jedan od kojih bi bilo izvlačenje informacija o potpisu na kraju rada, kao i o prilogu, elementi akta kojima se ovaj rad nije bavio.

REFERENCE

[1] Fabio Vitali, Flavio Zeni. "Towards a country-independent data format: the Akoma Ntoso experience." *Proceedings of the V legislative XML workshop*, pp. 67-86. 2007.

[2] Službeni Glasnik RS, *Statistika, Pravno informacioni Sistem* (http://www.pravno-informacioni-sistem.rs/SlGlasnikPortal/fp/statistics).

[3] Monica Palmirani, Fabio Vitali. "Akoma-Ntoso for legal documents." *Legislative XML for the semantic Web*, pp. 75-100. Springer, Dordrecht, 2011.

[4] Sartor G., Palmirani M., Francesconi E., Biasiotti M.A., "Legislative XML for the semantic web: principles, models, standards for document management" *(Vol. 4). Springer Science & Business Media*, strana 75, 2011.

[5] Hickey T.B., O'Neill E.T., Toves J., 2002. Experiments with the IFLA functional requirements for bibliographic records (FRBR). *D-Lib magazine*, 8(9), pp.1-13.

[6] Fabio Vitali, Monica Palmirani, Véronique Parisse. "Akoma Ntoso Naming Convention Version 1.0.", *OASIS Standard*, 2019. (https://docs.oasis-open.org/legaldocml/akn-nc/v1.0/akn-nc-v1.0.html)

[7] Byrum John D., "ISO 639-1 and ISO 639-2: International Standards for Language Codes. ISO 15924: International Standard for Names of Scripts.", 1999.

[8] Zakonodavni odbor Narodne skupštine Republike Srbije, "Jedinstvena metodološka pravila za izradu propisa", *Službeni glasnik RS, br. 21/2010*, 30. marta 2010.

[9] Monica Palmirani, Roger Sperberg, Grant Vergottini, Fabio Vitali, "Akoma Ntoso Version 1.0 Part 1: XML Vocabulary" *OASIS Standard*, 29 August 2018. (http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html)