HRVATSKA AKADEMSKA I ISTRAŽIVAČKA MREŽA - CARNet

CDA0027

Kako poboljšati pretraživost hrvatskog World Wide Web prostora uporabom metapodataka

Klasa: 232-000/01-1/1 **Kategorija: PREPORUKA** **Trajanje: do opoziva**
Ur.broj: 650-01-01-1
Datum nastanka: 15.01.2001. **Verzija: 1.0 (15.01.2001.)**
URL: ftp://ftp.carnet.hr/pub/CARNet/docs/advisories/CDA0027.pdf

Sažetak: *CARNet sustavno djeluje u području katalogizacije i pretraživanja hrvatskog Web prostora putem višegodišnje podrške projektima www.hr (katalog s mogućnošću pretraživanja) i CROSS (tražilica ".hr" vršne Internet domene), a od nedavna radi i na njihovom međusobnom povezivanju. Uz primjenu distribuiranog indeksiranja te prepoznavanje i korištenje metapodataka, očekuje se nastavak i poboljšanje kvalitete pretraživanja, te postupno uvođenje Dublin Core elemenata, kako unutar istraživačke i akademske zajednice, tako i u ukupnom hrvatskom Web prostoru. S tehničke strane, prvi koraci poboljšanja pretraživosti hrvatskog Web prostora – obuhvaćeni ovom preporukom – su relativno jednostavni i brzo primjenjivi, međutim, njeno prihvaćanje i stvarna izvedba u široj Web zajednici su ključni za opći napredak. CARNet preporučuje svim autorima sadržaja u hrvatskom Web prostoru korištenje metapodatakovnih elemenata u skladu s ovom preporukom.*

1. Uvod

Hrvatski World Wide Web prostor bogat je sadržajima koji pokrivaju raznovrsne interese i djelatnosti njegovih korisnika, no nažalost je nepotpuno indeksiran i teško pretraživ. Poznato je, također, da ovo nije specifičnost hrvatskog Weba, jer se globalno stanje može ocijeniti podjednako lošim zbog ogromne količine nediferenciranih elektroničkih podataka dostupnih putem Interneta. Svatko tko je pokušao pronaći informaciju koristeći neku od popularnih Web tražilica je vjerojatno bar jednom ostao nemoćan pred stotinama ili tisućama dobivenih "odgovora" na svoj upit, prilično beskorisnih bez mogućnosti njihova pročišćavanja ili pak postavljanja preciznijih upita.

Očito je da Webu u sadašnjem obliku nedostaju alati ili sredstva koja će omogućiti preciznije i uspješnije pronalaženje informacija. Bolja pretraživost se ne može postići samo usavršavanjem sustava za indeksiranje i pretraživanje, već u tome moraju sudjelovati i autori dokumenata. Cilj ove preporuke je pomoći autorima pri kreiranju jednostavnih opisnih, metapodatkovnih zapisa kojima će upotpuniti svoje dokumente, prije svega Web stranice, te doprinijeti njihovom kvalitetnijem indeksiranju i boljoj pretraživosti.

Metapodaci

Metapodaci (izvorno *metadata*) su "podaci o podacima". Metapodatkovni zapis se sastoji od skupa atributa, neophodnih za opis informacijskog izvora, i vrijednosti koje su pridružene tim atributima. Na primjer, metapodatkovni sustav knjižnica – katalog – sastoji se od metapodatkovnog zapisa s atributima koji opisuju knjigu ili drugu vrstu publikacije: autor, naslov, datum nastanka ili objavljivanja, tematsko područje i oznaku smještaja knjige na policama.

Web tražilice uz rezultate pretraživanja često nude određene metapodatke o dokumentima koje indeksiraju. Međutim, ukoliko autor ne definira metapodatke, tražilice strukturu dokumenata koje dohvaćaju s Interneta i metapodatke o njima samo nagađaju, manje ili više uspješno, i to na temelju više (npr., kako dobiti naslov) ili manje egzaktnih metoda (kako pronaći ključne riječi). Metapodaci omogućuju autoru dokumenta da sâm nedvosmisleno specificira podatke o dokumentu,a svrha ove preporuke je uputiti autore kako to učiniti.

Preporuka je podijeljena u četiri poglavlja. Nakon ovog uvoda, u drugom poglavlju opisano je zapisivanje metapodataka prema HTML specifikaciji, a u trećem poglavlju je opisan Dublin Core (DC) skup opisnih elemenata te su navedeni referentni izvori. Preporuka za primjenu DC-a unutar hrvatskog Web prostora je iznesena u četvrtom poglavlju. Na kraju dokumenta nalaze se kontakt podaci, popis kratica i popis referenci. Dodatak A sadrži primjer praktične primjene preporuke.

2. Metapodaci i HTML

HTML (Hypertext Markup Language) [4], standardni jezik za opis Web stranica, omogućuje specifikaciju metapodataka u zaglavlju dokumenta (unutar `HEAD` elementa). Element `HEAD` sadrži podatke o dokumentu kao što su naslov, metapodaci i ostali podaci koji se ne smatraju sadržajem dokumenta. Preglednik (engl. *browser*) te podatke načelno ne prikazuje, ali omogućuje korisniku da drugačije dođe do njih.

Dva načina specificiranja metapodataka unutar HEAD oznake su:

1. unutar samog dokumenta, pomoću `META` oznake ili uključivanjem RDF (Resource Description Framework) opisa [6];
2. referiranjem na vanjski metapodatkovni zapis pomoću `LINK` oznake.

Posebno treba spomenuti oznaku `TITLE`, koja je prema HTML specifikaciji obvezni dio svake `HEAD` oznake, a čija je uloga specificirati naslov dokumenta. Preglednici u pravilu prikazuju taj naslov u “title bar”-u svog prozora.

Specifikacija metapodataka unutar samog dokumenta

Specificiranje metapodatka uključuje navođenje svojstva (property) i sadržaja (value) za to svojstvo. Element `META` namijenjen je upravo za to: svojstvo se definira pomoću `NAME` atributa, a sadržaj svojstva se definira pomoću `CONTENT` atributa, npr:

```
<META name="description" content=”O uporabi metapodataka na Internetu">
```

Ovaj metapodatak specificira opis dokumenta (svojstvo "`description`").

Umjesto atributa `NAME` u nekim se slučajevima koristi atribut `HTTP-EQUIV`, kojim se specificiraju svojstva dokumenta vezana za Hypertext Transfer Protocol (HTTP) [5], kao nadopuna ili zamjena za podatke koje klijentu (pregledniku) HTTP-om šalje Web poslužitelj.

U sljedećem primjeru pokazana je primjena atributa `SCHEME`. Atribut `SCHEME` nije obvezan, a može se koristiti za pravilnu interpretaciju sadržaja svojstva, npr. metapodatak

```
<META name="identifier" content="0-8230-2355-9" scheme="ISBN">
```

specificira da je svojstvo “identifier” navedenog dokumenta zapisano kao ISBN kôd.

Još jedan važan atribut je atribut `LANG`. Atribut `LANG` omogućuje specificiranje jezika u kojem je naveden sadržaj metapodatka (treba uočiti da se ne radi o jeziku dokumenta). Na primjer, sljedeća

dva metapodatka će navesti ključne riječi za dokument u hrvatskom (vrijednost "hr") i engleskom (vrijednost "en") jeziku:

```
<META name="keywords" lang="hr" content="odmor, Hrvatska, sunce">
<META name="keywords" lang="en" content="vacation, Croatia, sunshine">
```

Specifikacija metapodataka referiranjem na vanjski metapodatkovni zapis

Element `LINK` omogućava specificiranje metapodataka o danom HTML dokumentu refrenciranjem na metapodatkovni zapis izvan samog dokumenta. Za tu potrebu se koristi atribut `REL` koji opisuje odnos referenciranog dokumenta s promatranim dokumentom. Npr, element:

```
<LINK rel="meta" href="http://myhost.carnet.hr/mydoc.rdf">
```

specificira da se metapodaci nalaze izvan dokumenta, na URL-u http://myhost.carnet.hr/mydoc.rdf.

3. Dublin Core Metadata Element Set

Vodeća inicijativa za unaprijeđenje mogućnosti pronalaženja ciljanih informacija u umreženoj okolini, pokrenuta 1995. godine, danas je poznata pod imenom Dublin Core Metadata Initiative (DCMI). Prva DCMI radionica (*Workshop*), održana u Dublinu, Ohio (SAD), okupila je knjižničare, istraživače digitalnih knjižnica, stručnjake za sadržaj, računalne mreže i specijaliste za označavanje teksta, s ciljem promoviranja standarda za pronalaženje umreženih izvora. Uslijedio je niz radionica koji je izrasao u interdisciplinarnu međunarodnu inicijativu, čijim koncenzusom je stvoren skup opisnih elemenata - Dublin Core Metadata Element Set (DCMES).

U DC terminologiji, metapodaci opisuju *informacijski izvor* (engl. *resource*). Izvorom se pritom naziva “bilo što s identitetom” [1]. U smislu primjene na Webu, izvor nudi *informaciju* ili *uslugu.* Treba reći da se metapodaci mogu koristiti za raznolike namjene, od lociranja izvora koji odgovara željenoj informaciji, do procjene njegove pogodnosti za uporabu.

Dublin Core Metadata Element Set (DCMES) je definiran kao minimalni skup *opisnih elemenata izvora*. Cilj primjene DCMES-a je lakše i učinkovitije pronalaženje traženih izvora na Webu, ali njegova primjena nije ograničena samo na Web i HTML – naprotiv, DCMES se može primijeniti na sve oblike izdavaštva.

DCMES je 15-člani skup metapodatkovnih opisnih elemenata (Tablica 1). Opisni elementi se mogu podijeliti u tri skupine:

(1) opisni elementi koji se odnose na *sadržaj* izvora,
(2) opisni elementi koji se odnose na izvor promatran kao *intelektualno vlasništvo*,
(3) opisni elementi koji se odnose na *primjerak* izvora.

Sadržaj		Intelektualno vlasništvo	Primjerak
Title	Source	Creator	Date
Subject	Relation	Publisher	Language
Description	Coverage	Contributor	Format
Type		Rights	Identifier

Tablica 1. - Opisni elementi izvora prema DC 1.1

Svaki opisni element je ponovljiv unutar DC metapodatkovnog zapisa. Redoslijed elemenata nije određen standardom. Također, nije obvezno navesti sve opisne elemente. Vrijednosti podataka koje

se pridjeljuju pojedinim elementima se određuju prema shemama, koje su najčešće standardi, formalne klasifikacije i sl.

Potpuna definicija DCMES 1.1 je dana u *Dublin Core Metadata Element Set*, Version 1.1 [1], gdje su definirani svi navedeni opisni elementi. DC specifikacija sadrži samo formalnu semantičku definiciju svakog opisnog elementa. To znači da postoji mnogo načina zapisa i prijenosa DC metapodataka. Uobičajeni načini su HTML, XML, RDF i relacijske baze podataka. Za primjenu na Webu, najzanimljiviji je zapis u HTML-u. Zapis u HTML-u specifiran je u RFC 2731, *Encoding Dublin Core Metadata in HTML* [8].

Referentni izvori za Dublin Core

Glavni izvor podataka o DC su Web stranice Dublin Core Metadata Initiative, na URL-u:

`http://purl.org/dc/`

Pristup je javan i besplatan. DC dokumenti se prema stupnju razvoja i prihvaćenosti unutar DC zajednice dijele u četiri skupine:

- preporuke (*recommendations*)
- prijedlozi (*proposed recommendations*)
- radni nacrti (*working drafts*)
- bilješke (*notes*).

Preporuke su stabilne specifikacije spremne za provođenje unutar DC zajednice. Npr, važeća specifikacija DC opisnih elemenata, *Dublin Core Metadata Element Set, Version 1.1. 1999-07-02* [1] ima status preporuke. Prijedlozi se mogu opisati kao "preporuke u nastajanju", dok se radne nacrte i bilješke može smatrati radnim dokumentima bez formalne potpore, stavljenim na uvid kako bi se potaknuli komentari i rasprave.

4. Preporuka za korištenje metapodataka na Webu

Skup temeljnih opisnih elemenata

Ovom preporukom se predlaže uporaba skupa temeljnih opisnih elemenata, prikazanog Tablicom 2, koji je podskup DCMES 1.1 i sadrži osam od ukupno petnaest opisnih elemenata definiranih u DCMES 1.1. Neki od elemenata podržani su s pripadajućim kvalifikatorima [2]. Smatra se da je taj odabrani podskup DCMES 1.1 prijelazno rješenje prema punoj implementaciji DC. Shodno tome, korištenje i ostalih DC elemenata prema [1] je u skladu s ovom preporukom.

Sinonimi

Preporuka definira i dva sinonima, `Keywords` i `Description`, koji će se koristiti za istovremeno označavanje odgovarajućih DCMES elemenata:

`Keywords` - sinonim za `DC.Subject`

`Description` - sinonim za `DC.Description`

Razlog uvođenja sinonima je njihova široka prihvaćenost i uporabna vrijednost koju imaju pod navedenim imenima. Npr, u HTML 4.01 specifikaciji može se naći nekoliko preporuka kako

tražilicama olakšati indeksiranje Web sadržaja, a čija primjena uključuje uporabu spomenutih elemenata.

Element
DC.Title
DC.Subject
DC.Description
DC.Creator
DC.Contributor
DC.Publisher
DC.Date
DC.Language

Tablica 2. - Temeljni opisni elementi

Opis skupa temeljnih opisnih elemenata

U nastavku slijedi opis temeljnih opisnih elemenata. Treba napomenuti da smisao ovog opisa nije ponovno definiranje DCMES 1.1 [1], već, u nedostatku službenog hrvatskog prijevoda, davanje smjernica za korištenje podskupa DC temeljnih opisnih elemenata na Webu. Također su navedene preporučene maksimalne duljine njihovih vrijednosti koje će CARNetovi informacijski servisi www.hr i CROSS podržati.

`DC.Title`

Definicija: Naslov odn. naziv izvora.

Komentar: Načelno, opisni element bi trebao imati isti sadržaj kao i HTML oznaka TITLE. Preporučuje se da ukupna duljina bude najviše 140 znakova.

`DC.Creator`

Definicija: Ime osobe ili tvrtke koja je stvorila sadržaj izvora.

Komentar: Preporučeni format je: "ime prezime", "prezime, ime" ili "puni naziv tvrtke". Konzistentnost sadržaja ovog elementa je vrlo bitna.

`DC.Subject`

Definicija: Tema sadržaja izvora.

Komentar: Obično se sastoji od ključnih riječi, pojmova i izraza, odvojenih zarezima. Preporučuje se odabrati vrijednosti iz zadanog rječnika ili formalne klasifikacijske sheme. Preporučuje se duljina od najviše 20 riječi.

`DC.Description`

Definicija: Prikaz sadržaja izvora. Preporučuje se da sadrži sažetak ili neformalni tekstualni opis sadržaja izvora, u kojem slučaju se može koristiti kvalifikator `Abstract`. Ako se koristi za navođenje kazala preporučuje se korištenje kvalifikatora `TableOfContents`.

Komentar: Duljina elementa ne bi trebala preći 360 znakova.

`DC.Publisher`

Definicija: Ime osobe ili tvrtke koja je izvor učinila dostupnim, odnosno objavila.

Komentar: Preporučeni format je: "ime prezime", "prezime, ime" ili "puni naziv tvrtke". Konzistentnost sadržaja ovog elementa je vrlo bitna.

`DC.Contributor`

Definicija: Ime osobe ili tvrtke koja je pridonijela stvaranju sadržaja izvora.

Komentar: Preporučeni format je: "ime prezime", "prezime, ime" ili "puni naziv tvrtke". Konzistentnost sadržaja ovog elementa je vrlo bitna.

`DC.Date`

Definicija: Datum, odnosno vrijeme stvaranja, promjene ili objavljivanja izvora. Za specifikaciju o kojem se vremenu radi, koriste se odgovarajući kvalifikatori, redom: `Created`, `Modified` i `Issued`.

Komentar: Preporučuje se korištenje formata iz [9].

`DC.Language`

Definicija: Jezik u kojem je napisan sadržaj izvora, odnosno njegov najveći dio.

Komentar: Preporučuje se uporaba kôdova iz RFC 1766 [7].

Primjer u Dodatku A ilustrira praktičnu primjenu preporučenog skupa temeljnih opisnih elemenata na ovaj dokument. Priručnik za primjenu DC-a [3] može poslužiti kao dodatna literatura.

Preporuka autorima Web stranica u hrvatskom Web prostoru

CARNet preporučuje svim autorima sadržaja u hrvatskom Web prostoru uvođenje temeljnih opisnih DC metapodatkovnih elemenata, u skladu s ovom preporukom. CARNetovi informacijski servisi CROSS (http://cross.carnet.hr) i www.hr (http://www.hr) će navedeni podskup DCMES početi podržavati nakon objavljivanja ove preporuke, te se zato autorima Web stranica savjetuje da opisne elemente počnu koristiti odmah, kako bi njihove stranice dobile dodatnu vrijednost čim spomenuti servisi podrže preporuku.

5. Autori i kontakt

Ovaj dokument nastao je kao rezultat rada na projektima CROSS i www.hr. Autori su mu Maja Matijašević i Hrvoje Stipetić, a urednik Miroslav Milinović. Svoje komentare i pitanja vezana uz ovu preporuku možete poslati elektroničkom poštom na adresu:

`metadata@carnet.hr.`

6. Reference

[1] *Dublin Core Metadata Element Set*, Version 1.1. 1999-07-02
<http://purl.org./dc/documents/rec-dces-19990702.htm>

[2] *Dublin Core Qualifiers*, 2000-07-11
<http://purl.org/dc/documents/rec/dcmes-qualifiers-20000711.htm>

[3] Hillmann, D.*User Guide Working Draft*, 1998-07-31
<http://purl.org/dc/documents/wd-guide-current.htm>

[4] *HyperText Markup Language 4.01*, World Wide Web Consortium (W3C) Recommendation, 1999-12-24.
< http://www.w3.org/TR/1999/REC-html401-19991224>

[5] Fielding, R., J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, T. Berners-Lee. *Hypertext Transfer Protocol -- HTTP/1.1*, IETF RFC 2616, June 1999.

[6] *Resource Description Framework (RDF) Model and Syntax Specification*, 1999-02-22
< http://www.w3.org/TR/1999/REC-rdf-syntax-19990222>

[7] Alvestrand, H., *Tags for the Identification of Languages*, IETF RFC 1766, March 1995

[8] Kunze, John A. *Encoding Dublin Core Metadata in HTML*, IETF RFC 2731, December 1999.

[9] Wolf, M., C. Wicksteed *Date and Time Formats*. http://www.w3.org/TR/1998/NOTE-datetime-19980827

7. Popis kratica

DC	Dublin Core
DCMES	Dublin Core Metadata Element Set
MIME	Multipurpose Internet Mail Extensions
URI	Universal Resource Indentifier
URL	Universal Resource Locator
RDF	Resource Description Framework
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
IETF	Internet Engineering Task Force
IP	Internet Protocol
XML	eXtensible Markup Language

Dodatak A: Primjer primjene temeljnih opisnih DC elemenata

Sljedeći primjer, koji bi se mogao primijeniti za ovaj dokument, prikazuje kako se u HTML unosi preporučeni skup temeljnih opisnih elemenata.

```
<HEAD>

<TITLE>Kako poboljšati pretraživost hrvatskog World Wide Web prostora
uporabom metapodataka</TITLE>

<META http-equiv="Content-Type" content="text/html; charset=iso-8859-2">

<META name="description" content=" Preporuka objašnjava ulogu
metapodataka na Webu, prezentira Dublin Core Metadata Initiative i DCMES
1.1, te predlaže obim uporabe metapodataka. Namjera dokumenta je pomoći
autorima pri kreiranju jednostavnih metapodatkovnih zapisa kojima će
upotpuniti svoje dokumente, te doprinijeti njihovom kvalitetnijem
indeksiranju i boljoj pretraživosti.">

<META name="keywords" content="Dublin Core, DC, DCMI, DCMES, metadata,
metapodaci, RDF, CARNet preporuka, www.hr, CROSS, CARNet">

<LINK rel="scheme.DC" href="http://purl.org/dc/elements/1.1/">

<META name="DC.Title" content="Kako poboljšati pretraživost hrvatskog
World Wide Web prostora uporabom metapodataka">

<META name="DC.Description" content=" Preporuka objašnjava ulogu
metapodataka na Webu, prezentira Dublin Core Metadata Initiative i DCMES
1.1, te predlaže obim uporabe metapodataka. Namjera dokumenta je pomoći
autorima da kreiraju jednostavne metapodatkovne zapise kojima će
upotpuniti svoje dokumente, te doprinijeti njihovom kvalitetnijem
indeksiranju i boljoj pretraživosti.">

<META name="DC.Subject" content="Dublin Core, DC, DCMI, DCMES, metadata,
metapodaci, RDF, CARNet preporuka, www.hr, CROSS, CARNet">

<META name="DC.Creator" content="Hrvoje Stipetić">

<META name="DC.Creator" content="Maja Matijašević">

<META name="DC.Publisher" content="CARNet – Hrvatska akademska i
istraživačka mreža">

<META name="DC.Language" content="hr">

<META name="DC.Date.Created" content="2000-07-17">

<META name="DC.Date.Modified" content="2001-01-08">

<META name="DC.Date.Issued" content="2001-01-15">

</HEAD>
```