Kvantitativna i kvalitativna analiza Digitalnog arhiva hrvatskih mrežnih publikacija

Holub, Karolina; Pigac Ljubi, Sonja; Rudomino, Ingeborg

Kvantitativna i kvalitativna analiza Digitalnog arhiva hrvatskih mrežnih publikacija

Quantitative and qualitative analysis of Digital Archive of Croatian Web Resources

Karolina Holub
kholub@nsk.hr
Sonja Pigac Ljubi
spigac@nsk.hr
Ingeborg Rudomino
irudomino@nsk.hr

Sažetak
Rad analizira vrstu građe koja ulazi u Digitalni arhiv hrvatskih mrežnih publikacija i sadržaj samog arhiva. Predmet analize je uzorak od 3097 publikacija na stanje arhiva do kraja rujna 2009. Rezultati analize uspoređeni su s podacima iz prosinca 2005. godine na uzorku od 1024 publikacije. Svrha rada je pokazati kvantitativne i kvalitativne promjene u sadržaju i funkcionalnosti rada arhiva.
Analiza je pokazala da je integrirajuća građa porasla za 24%, omeđene publikacije bilježe pad od 7%, a najveći pad vidljiv je kod serijskih publikacija za 21 %. Vidljive su i promjene kod ostalih parametara analize sadržaja arhiva. U radu se donosi i kratak osvrt na poslovanje i razvoj arhiva.

Ključne riječi: digitalni arhiv, vrsta građe, arhiviranje građe

Summary
This work analyses all types of resources and content of Digital Archive of Croatian Web Resources. Analysis is made on the sample of 3097 archived publications from 2005. until September 2009. The results of the analysis in 2009. are compared with the research (možda bolje analysis results) results from 2005. The aim of this work is to show the quantity and quality changes in contents and functionalities of the archive.
According to the analysis results, integrating resources indicate upgrowth of 24 %, whereas monographs and serials indicate drop of 7% and 21 % accordingly. Apart from these, the paper also rapresents changes of other parameters included in the analysis.

Keywords: digital archive, types of resources, archiving

1. Uvod

Mrežnim publikacijama svojstvena je promjenjivost, dinamičnost, veličina datoteka te kratak i nepredvidiv vijek trajanja na internetu. Za rad s takvom vrstom građe odnosno za preuzimanje i arhiviranje obveznog primjerka hrvatskih mrežnih publikacija, suradnjom Nacionalne i sveučilišne knjižnice u Zagrebu i Sveučilišnog računskog centra Sveučilišta u Zagrebu uspostavljen je 2005. godine sustav za selektivno arhiviranje Digitalni arhiv hrvatskih mrežnih publikacija (DAMP)
Do 2007. godine u Nacionalnoj i sveučilišnoj knjižnici u Zagrebu mrežne su se publikacije obrađivale u knjižničnom sustavu CROLIST koristeći format UNIMARC, nakon čega se

prešlo na novi sustav Voyager kao i novi format MARC21. Prva analiza sadržaja arhiva provedena je već 2005. godine na uzorku od samo 1024 publikacije.
Tijekom pet godina rada arhiva broj jedinica (publikacija) u arhivu se gotovo utrostručio zbog čega se ukazala potreba za novom analizom njegova sadržaja. Nova analiza provedena je na cjelokupnom sadržaju arhiva (odnosno građe koja u njega ulazi) prema zastupljenosti vrsta građe, dostupnosti publikacija na webu, zastupljenosti domena arhiviranih publikacija, vrsti podataka koje publikacije sadrže te učestalosti, dubinom i parametrima pobiranja. Osim navedenih parametara koji su bili uključeni u analizu iz 2005. godine, u novu su analizu uključeni i novi aspekti pohrane i funkcionalnosti arhiva

U ovom radu analizom su obuhvaćeni podaci koji se odnose na stanje arhiva do rujna 2009. godine na uzorku od 3097 publikacija te su uspoređeni su s rezultatima iz prosinca 2005. godine [1]. Također, ti su podaci uspoređeni i s podacima o serijskim publikacijama na mreži iz 2001. godine[2].
Novi aspekt u procesu obrade i pohrane mrežnih publikacija je njihova sadržajna obrada koja je započela 2008. godine te je zastupljenost prema sadržajnim kategorijama uzeta kao novi parametar u ovoj analizi. S obzirom da rezultati analize čine tek jednu trećinu obrađenih kategorija, ne daju potpunu sliku sadržaja arhiva.
Novi parametri analize uključuju i usporedbu sadašnjih kategorija serijskih publikacija s onima iz 2000. godine. te postojanje tiskanih inačica uz mrežno izdanje.

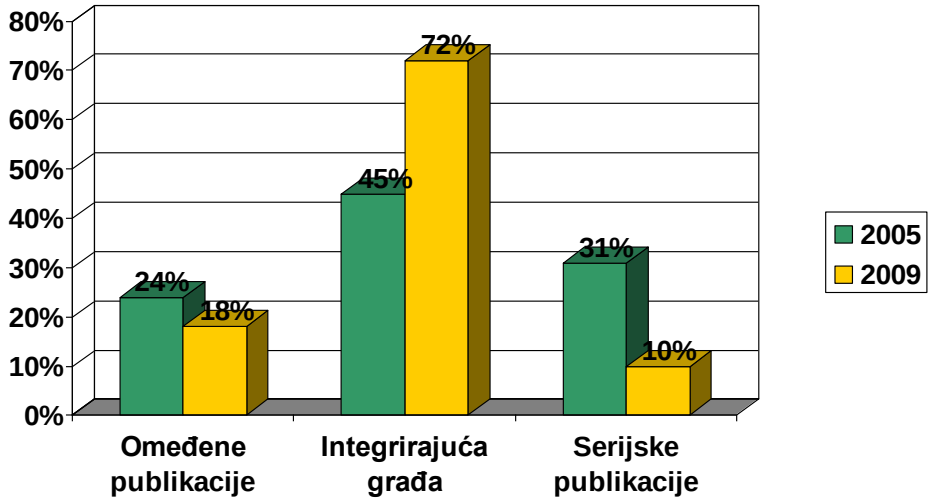
Broj arhiviranih publikacija do rujna 2009. godine iznosio je 3.097 što predstavlja porast od preko 200 % u odnosu na stanje s kraja 2005. kada je iznosio 1.024. Arhiviranih primjeraka je 28.046 što u odnosu na 2005. kada ih je bilo 3.674 čini porast oko 76%.Veličina arhiva također je porasla s 269 GB koliko je iznosila na kraju 2005., na 2,4 TB u studenom 2009. godine.
.

2. Analiza građe iz kataloga i arhiva

2.1 Vrste mrežne građe

Mrežna građa koja ulazi u Digitalni arhiv odabire se selektivno prema važećim Kriterijima odabira obveznog primjerka mrežne građe za obradu i arhiviranje[3]. Tri su vrste mrežne građe:
a) integrirajuća građa
b) omeđene publikacije
c) serijske publikacije.
Analiza je pokazala da najveći postotak sadržaja arhiva pripada integrirajućoj građi - 72%. Slijede omeđene publikacije s 18% te serijske s 10%. U usporedbi s 2005. godinom vidljiv je pad omeđenih i serijskih publikacija, dok integrirajuća građa bilježi rast (Slika 1).

[1] Pigac, Sonja; Buzina, Tanja. Selektivno arhiviranje hrvatskog weba : rezultati i otvorena pitanja // 9. seminar Arhivi, knjižnice, muzeji : mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova / uredile Mirna Willer i Ivana Zenić. Zagreb : Hrvatsko knjižničarsko društvo, 2006. Str. 28-39
[2] Klarin, Sofija; Pigac, Sonja. Hrvatske daljinski dostupne elektroničke serijske publikacije // Vjesnik bibliotekara Hrvatske 4, 43(2000.[i.e. 2001.]). Str.163
[3] Kriteriji odabira obveznog primjerka mrežne građe za obradu i arhiviranje. Dostupno na: http://www.nsk.hr/DigitalLib.aspx?id=83

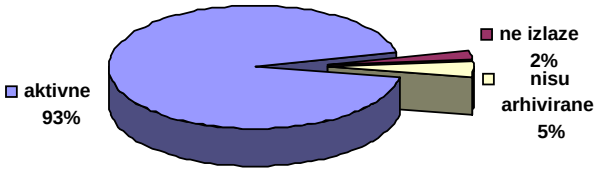


Slika 1. Usporedba vrsta arhivirane mrežne građe: 2005.-2009.

2.1.1Integrirajuća građa

Integrirajuća građa čini najzastupljeniju vrstu građe u arhivu. Nju čine portali, e-zini, mrežna mjesta udruga, gradova, raznih državnih i javnih institucija, mrežne inačice novina, najnovije osobne stranice, forumi i blogovi.
ISBD(CR) definira integrirajuću građu[4] kao „bibliografsku jedinicu koja se dopunjuje ili mijenja nadopunama koje se uključuju u cjelinu i ne ostaju izdvojene (npr., mrežna mjesta koja se osuvremenjuju).[5] Integrirajuća građa osuvremenjuje se u redovitim ili neredovitim vremenskim razmacima bez predviđenog kraja izlaženja (portali koji donose vijesti o tekućim zbivanjima, mrežna mjesta ministarstava, gradova, ustanova, dnevne novine i sl.).
Analiza iz 2005. također je pokazala da je integrirajuća građa činila najveći udio sveukupne građe u arhivu s oko 45%[6] što je u odnosu na 72% iz 2009. porast od 26%. Najveći broj integrirajuće građe uspješno je i arhiviran – čak 93% - što dokazuje uspješan razvoj tehnologije pobiranja. Od ukupnog broja integrirajuće građe u arhivu samo je 2,% jedinica nestalo s weba, a nije arhivirano njih 5% koje i dalje izlaze na webu (slika 2).

[4] U radu s mrežnim publikacijama, zbog njihove promjenjivosti i dinamičnosti te nepredviđenog roka izlaženja, pokazalo se da termin omeđena integrirajuća građa primjereniji za tiskane publikacije odnosno za publikacije koje imaju uvez sa slobodnim listovima, stoga se zbog svega navedenog u ovom radu ne razlikuju omeđena i neomeđena integrirajuća građa..
[5] ISBD(CR) : međunarodni standardni bibliografski opis serijskih publikacija i druge neomeđene građe. Zagreb : Hrvatsko knjižničarsko društvo, 2005. Str. 16.
[6] Ovaj postotak okvirni je rezultat zbroja neomeđene i omeđene integrirajuće građe iz 2005. godine



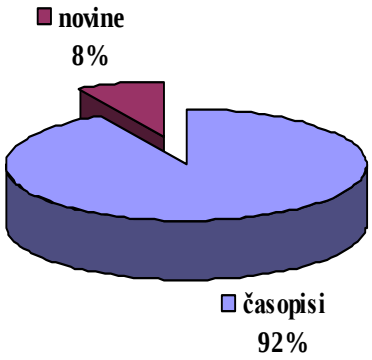
Slika 2 Uspješnost arhiviranja integrirajuće građe

2.1.2 Omeđena građa

Druge po zastupljenosti vrste građe u arhivu su omeđene publikacije s 525 jedinice (18%). U usporedbi s njihovom zastupljenošću od 24% iz 2005. očito je da bilježe pad od 6% .
Najveći je udio omeđenih publikacija koje su aktivne, još uvijek prisutne na mreži i uspješno arhivirane - njih 83%. Omeđenih mrežnih publikacija za koje postoji arhivirani primjerak, ali više ne postoje na webu ukupno je 11%. Publikacije koje su ušle u arhiv, ali nisu arhivirane je 36%.
Uspoređujući podatke iz 2005. godine broj katalogiziranih i uspješno arhiviranih omeđenih mrežnih publikacija porastao je za oko 15% (2005-68%; 2009-83%), dok je broj nearhiviranih publikacija u opadanju za oko 4,99% (2005-10,99%; 2009-6%) što potvrđuje poboljšanje uspješnosti pobiranja zahvaljujući raspoloživosti i primjeni odgovarajućih parametara te unaprijeđene verzije pobirača.

2.1.3 Serijske publikacije

Od ukupnog broja mrežne građe, serijskih publikacija je 309 (10 %) te čine najmanje zastupljenu skupinu u arhivu. U usporedbi s podacima iz 2005. kada su činile 31 % ukupne građe, evidentan je pad od 21 %.



Slika 3. Vrsta serijskih publikacija

Od ukupnog broja serijskih publikacija, velika većina otpada na časopise (92 %), a znatno manje na novine (8%). Rezultati analize pokazuju da je od ukupnog broja katalogiziranih

mrežnih serijskih publikacija, njih 5% ušlo u sustav DAMP, no još uvijek nije arhivirano. Razlozi tomu su tehnologija izrade mrežnih stranica koja otežava ili potpuno onemogućuje pobiranje te ograničenje od strane nakladnika koji za pristup publikaciji zahtijeva registraciju putem korisničkog imena i lozinke. Arhivski primjerak, ali ne više i izvornik na webu, ima njih 16%. Kao i u slučaju omeđenih mrežnih publikacija, najveći je udio serijskih publikacija koje su aktivne, još uvijek prisutne na mreži i uspješno arhivirane – njih 79%.

Iako se očekivao porast broja serijskih publikacija na mreži to se nije dogodilo. Naime, pojavom integrirajuće građe koja je na odgovarajući način mogla odražavati stalne promjene sadržaja koji se dodavao, mijenjao ili jednostavno nestajao, serijske publikacije na webu počele su gubiti svoju osnovnu značajku – izlaženje u zasebnim dijelovima.

Promjena vrste građe najbolje se vidi na nekim primjerima mrežnih novina, npr. Jutarnji.hr, Večernji.hr (u početku inačicama tiskanih izdanja) koje su napustile način objavljivanja u zasebnim dijelovima, već su svaku promjenu, dodavanje ili uklanjanje sadržaja počele izražavati putem osuvremenjenih instanci koje su integrirane u cjelinu.

Napravljena je i usporedba prema kategorijama serijskih publikacija objavljenim 2001. godine [7]. Analiza je pokazala da je znatno porastao broj časopisa državnih službi, nakon kojih slijede časopisi društava/udruga, znanstveni časopisi te časopisi fakulteta/znanstvenih instituta (Tablica 1).

Vidljiv je pad zastupljenosti novina i tjednih/dvotjednih časopisa koji se može objasniti promjenom načina objavljivanja na webu, odnosno gore spomenutom promjenom vrste građe.

KATEGORIJA ČASOPISA	2000.	2009.
Novine	24	17
Tjedni/dvotjedni časopisi	40	21
Znanstveni časopisi	7	40
Vjerski časopisi	9	8
Studentski časopisi	4	5
Časopisi društva/udruga	17	60
Časopisi fakulteta/znanstvenih instituta	14	37
Časopisi državnih službi	10	81
Časopisi poduzeća	7	27
Ostali	8	13

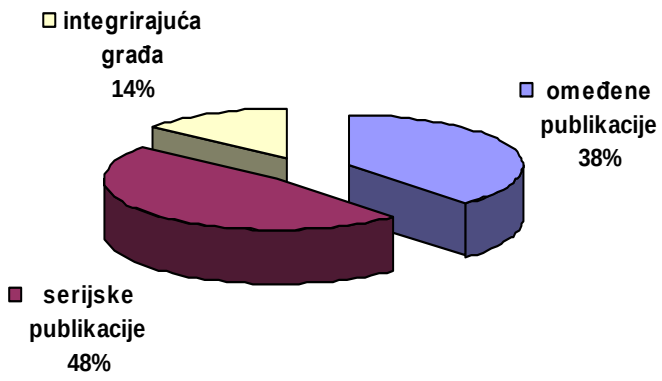
Tablica 1. Usporedba kategorizacije serijskih publikacija: 1997.-2009. godina

2.2 *Tiskana inačica*

Novi parametar analize iz 2009. je postojanje tiskanog izdanja za sve vrste mrežne građe. Rezultati su pokazali da serijske publikacije u daleko većem broju uz mrežno izdanje imaju i tiskanu inačicu – gotovo polovica svih serijskih publikacija u arhivu. Slijede omeđene publikacije te integrirajuća građa koja tiskanu inačicu ima u daleko najmanjem broju. Ovi rezultati zapravo i nisu neočekivani s obzirom da velik broj nakladnika serijskih publikacija čine fakulteti i znanstveni instituti, udruge i državne službe koji njeguju tradicionalno izdavaštvo. Integrirajuća građa jednako tako po svojoj prirodi nema tiskanog ekvivalenta (koji bi u tiskanom izdanju odgovarao publikaciji koja ima uvez sa slobodnim listovima), a podaci

[7] Klarin, Sofija; Pigac, Sonja. Hrvatske daljinski dostupne elektroničke serijske publikacije // Vjesnik bibliotekara Hrvatske 4, 43(2000.[i.e. 2001.]). Str.163

dobiveni analizom odnose se na starije publikacije koje su u početku bile serijske, a s vremenom postale integrirajuće.



Slika 4. Postojanje tiskane inačice mrežne građe

2.3 Sadržajna obrada mrežne građe

2008. godine, u suradnji s Odsjekom za sadržajnu obradu NSK započelo se sa sadržajnom obradom mrežne građe. Jedinicama svih vrsta građe dodjeljuje se UDK klasifikacijske oznaka u polje 080 formata MARC 21,dok se predmetnice dodjeljuju samo omeđenim publikacijama i integrirajućoj građi u polja 6XX. U svrhu što kvalitetnijeg pregledavanja i pretraživanja sadržaja digitalnog arhiva, uspostavljene su kategorije prema sadržaju/predmetu temeljene na UDK tablicama i praksi sadržajne obrade u Nacionalnoj i sveučilišnoj knjižnici u Zagrebu te su usklađene s kategorijama korištenim u drugim arhivima, poput PANDORA-e[8] i UK WebArchive-a[9]. Tablica 2 prikazuje zastupljenost pojedinih kategorija koja za sada obuhvaća oko trećinu arhiva
Zanimljivo je primijetiti da se najviše građe odnosi na područje ekonomije i poslovanja (23%), umjetnosti (19%) te prava i uprave (17%), dok se tek oko 9% odnosi na građu vezanu za obrazovanje i prirodne znanosti.
S obzirom da jedna jedinica može imati i više od jedne UDK oznake za potrebe ove analize odlučeno je da se uzme samo ona prva.

SADRŽAJNE KATEGORIJE	RUJAN 2009.
Ekonomija i poslovanje	23%
Umjetnost	19%
Pravo i uprava	17%
Sport i rekreacija	11%
Povijest i geografija	10%
Obrazovanje	9%
Priroda i okoliš	9%
Politika	2%

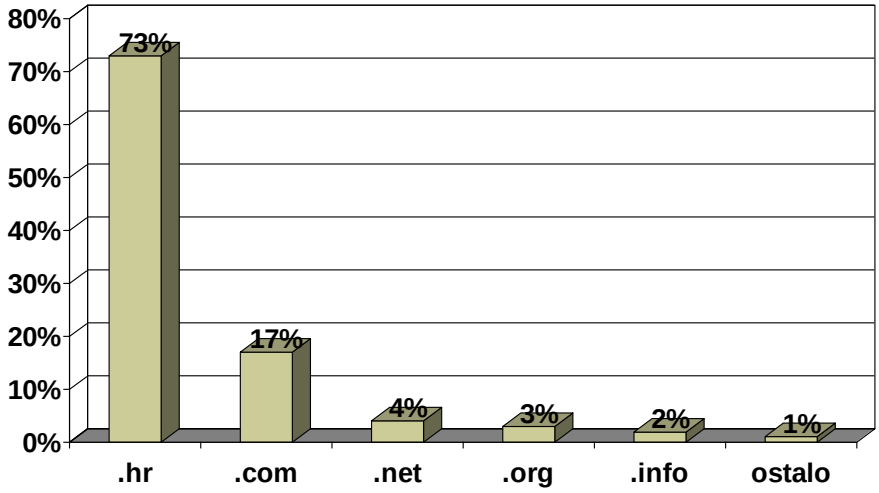
[8] PANDORA. [citirano: 2010-01-10]. Dostupno na: http://pandora.nla.gov.au/

[9] UK Web Archive. [citirano: 2010-01-10]. Dostupno na: http://www.webarchive.org.uk/ukwa/

Tablica 2. Zastupljenost sadržajnih kategorija

2.4 Distribucija krovnih domena

Jedan od kriterija odabira mrežne građe za dugoročnu pohranu u arhiv DAMP je i domena mrežne adrese na kojoj je publikacija objavljena. Iako se prednost daje publikacijama objavljenim na nacionalnoj krovnoj domeni .hr, pobiru se i one s vrijednim sadržajem na drugim domenama.
Analiza krovnih domena publikacija zastupljenih u arhivu pokazala je da je daleko najveći udio publikacija s nacionalne krovne domene (73%). Odmah nakon .hr domene slijedi .com (17,%) sa znatno manjim udjelom, zatim .net sa 4%, .org s 3%, .info s 2%. Ukupno je registrirano 19 različitih domena (uz već spomenute tu su još .eu, .biz, .ch, .at, .tk, .cz, .ba, .uk, .de, .us, .dk, tv, .in, .it). Iz navedenog se vidi da je mrežna građa od značaja za hrvatsku nacionalnu baštinu obuhvatila i domene nekih drugih zemalja poput Austrije, Švicarske, Češke, Bosne i Hercegovine, Njemačke itd. (npr. .at, .ch, .cz, .ba, .de), iako zastupljene samo s jednom jedinicom građe. Prema podacima analize iz 2005. u arhivu je bilo zastupljeno 12 domena te je njihov broj porastao za novih sedam. Slika 5 prikazuje prvih 5 domena prema zastupljenosti (domene at, .tk, .cz, .ba, .uk, .de, .us, .dk, tv, .in, .it zastupljene su samo s 1, .eu s 4, .biz s 3, a .ch s 2 jedinice građe)

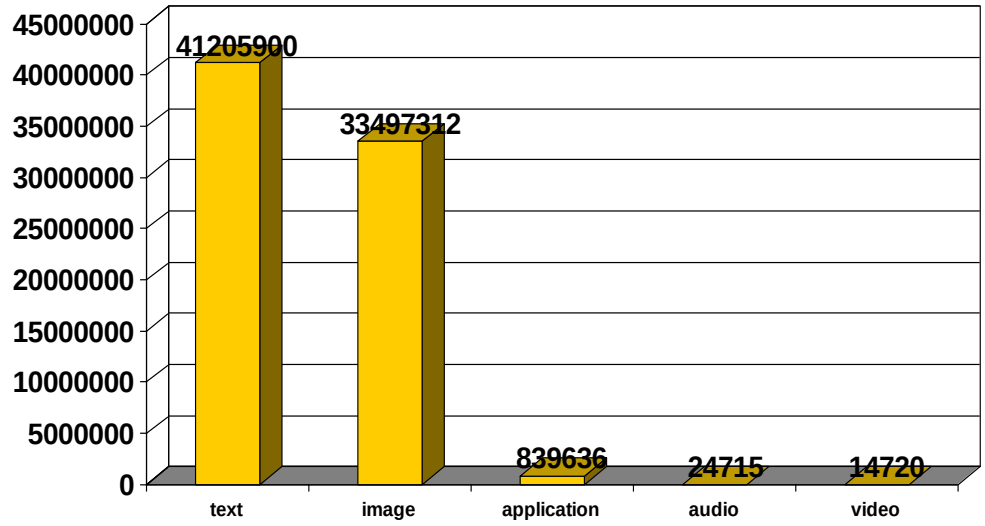


Slika 5. Zastupljenost domena do rujna 2009. godine

2.5 Vrste podataka

Sustav DAMP pri svakom pojedinačnom pobiranju mrežne publikacije bilježi i formate datoteka koje se nalaze na tom web-sjedištu. Vrste medija koje se još nazivaju i MIME vrste ili vrste sadržaja, zapravo su klasifikacijski sustav koji se koristi za identifikaciju datoteka koje se obično nalaze na web-sjedištima, a presudne su za uspješno funkcioniranje weba jer omogućuju komunikaciju između klijentskog računala i poslužitelja na kojem se određeno web-sjedište nalazi. u pogledu prihvaćanja određenih vrsta datoteka.
Najčešće vrste su text/html datoteke koje se koriste za identifikaciju HTML datoteka i image/jpg koje se koriste za identifikaciju slikovnih datoteka u formatu JPEG. Korist bilježenja ovakve vrste podataka nije samo u slučaju razjašnjenja neuspješnih pobiranja, već i za potrebe dugoročne zaštite koja još uvijek predstavlja prilično neistraženo područje i na razini međunarodne knjižničarske zajednice. Kao što prikazuje slika 6 najzastupljenija vrsta

sadržaja je tekst, nakon kojeg odmah slijede slike. Iako su zastupljene i druge vrste podataka (sadržaja), njih u cjelini ima znatno manje.

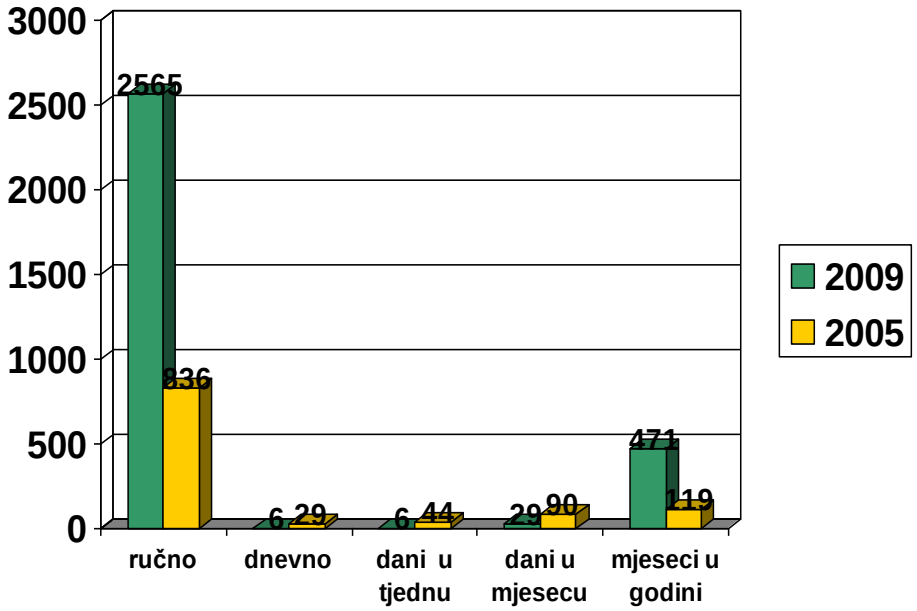


Slika 6. Zastupljenost vrsta sadržaja

2.6 Učestalost pobiranja

Mrežna se građa u arhivu pobire određenom učestalošću. Prema vrstama građe to znači da se omeđene publikacije pobiru jednom, a serijske publikacije i integrirajuća građa nekom određenom učestalošću, koja može i ne mora biti redovita. Sustav DAMP ima zadane sljedeće opcije učestalosti pobiranja: *ručno, dnevno, dani u tjednu, dani u mjesecu, mjeseci u godini*. Pri odabiru opcije učestalosti pobiranja, knjižničar mora voditi računa ne samo o vrsti građu za koju zadaje parametar, već i o vrijednosti sadržaja, odnosno područja koje publikacija pokriva, ali i o izgledu te obavezno o veličini[10] same publikacije. Ukoliko se radi o neomeđenoj građi (serijskim publikacijama i integrirajućoj građi) koja će se nužno pobrati više nego jednom, knjižničar se ne može u potpunosti osloniti na učestalost izlaženja/osuvremenjivanja navedenu na samoj građi s obzirom da je se ni izdavači uvijek ne pridržavaju. Tek praćenjem redovitosti izlaženja /osuvremenjivanja može se zadati odgovarajuća opcija učestalosti pobiranja kako bi se izbjegla mogućnost da isti broj bude pobran dvaput, a novi niti jednom (npr. u slučaju serijske publikacije).
Opcija *ručno* koristi se za pobiranje omeđenih publikacija ili integrirajuće građe koja se ne arhivira zadanim automatizmom, već samo jednom ili u nepredvidivim vremenskim razmacima te za publikacije čija veličina prelazi 500 MB. Slika 7 prikazuje zastupljenost učestalosti pobiranja građe u arhivu za 2005. i 2009. godinu. Vidljiv je porast publikacija kojima je učestalost pobiranja postavljena na ručno što ukazuje na to da se publikacije u arhivu ili ne osuvremenjuju redovito ili su prevelike.

[10] Ukoliko veličina publikacije prelazi 200 MB smanjuje se učestalost pobiranja zbog ograničenog prostora na disku.

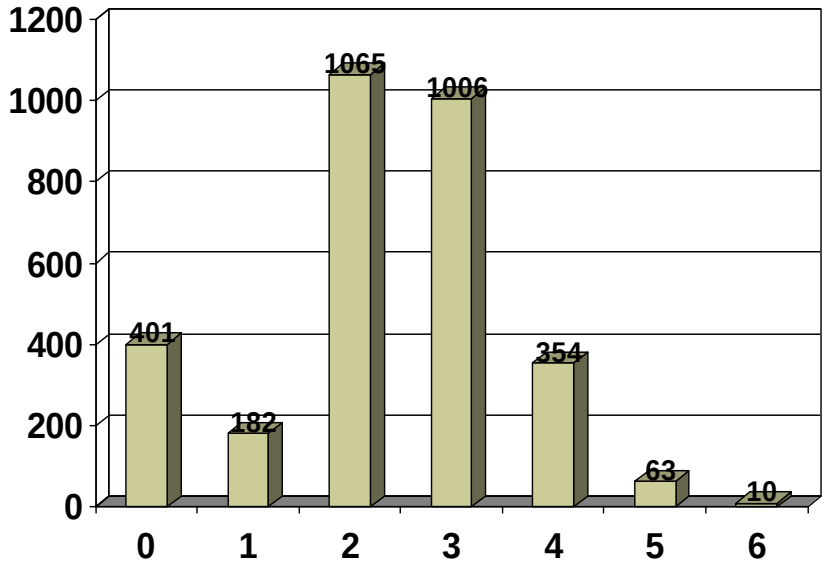


Slika 7. Učestalost pobiranja – usporedba 2005. i 2009.

2.7 Parametri pobiranja

Dubina pobiranja jedini je obvezni parametar koji se ujedno ne može ponavljati. Izravno utječe na uspješnost pobiranja željenog web-sjedišta jer se njime određuju granice publikacije (građe) koja se želi pobrati. Dubina pobiranja također različito utječe na resurs (npr. slika) ovisno o tome je li on uložen (eng. embedded) u mrežnu stranicu ili nije. U slučaju neuloženog resursa, on će biti prikupljen ukoliko je dubina na kojoj se nalazi manja ili jednaka zadanoj vrijednosti parametra dubine pobiranja. U suprotnom (resurs je uložen), prikupit će se ukoliko se nalazi na dubini manjoj ili jednakoj vrijednosti parametra dubine pobiranja uvećanoj za 2[11]. Općenito, dubina 0 znači da će se pobrati samo naslovnica, dubina 1 naslovnica i poveznice na njoj i sl. Sukladno tomu, veća dubina znači veću količinu podataka koji se prikupljaju, ali i usporeniji rad pobirača te bržu potrošnju diskovnog prostora za pohranu podataka . Slika 8 prikazuje zastupljenost vrijednosti parametra dubine pobiranja u cjelokupnom arhiv iz čega je vidljivo da se najviše se pobiru publikacije s dubinom 2 i 3.

[11] Milinović, Miroslav; Topolšćak, Nebojša. DAMP II : Digitalni arhiv mrežnih publikacija : nova funkcionalnost – novi planovi. // 9. seminar Arhivi, knjižnice, muzeji : mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova / uredile Mirna Willer i Ivana Zenić. Zagreb : Hrvatsko knjižničarsko društvo, 2006. Str. 40-55.



Slika 8. Dubina pobiranja

Osim parametra *dubina pobiranja*, za uspješno i precizno određivanje granica mrežne publikacije sustav DAMP nudi i druge parametre.
Prije svega tu je parametar *neželjeni dio stabla* kojim se iz pobiranja isključuje staza u URL-u, odnosno dijelovi stabla mrežne publikacije koju se ne želi pobrati (suprotan ovome je parametar *željeni dio stabla*). Ovaj se parametar često koristi kada je mrežna stranica strukturirana na način da se njegovom primjenom iz svakog sljedećeg pobiranja mogu ukloniti određeni dijelovi koje ne želimo čuvati (poput foruma, chatova i sl). Ostali parametri služe za upućivanje sustava na drugi URL na koji se publikacija proteže i slijeđenje vanjskih poveznica, navođenje sinonima naziva domaćina, specifikaciju vrste podataka (formate datoteka) koja se želi pobrati unutar nekog mrežnog mjesta, naređivanje pobiraču da uvijek prikupi uložene dijelove mrežnih stranica, vremenski poček prije početka sljedećeg pobiranja i sl.
Tablica 3 prikazuje usporedbu zastupljenosti svakog pojedinačnog parametra.
Zbog potrebe prilagodbe novim web-tehnologijama u razdoblju od 2005. do 2009. godine u sustav DAMP uvedeni su novi parametri: *poček prije dohvata, rabi sadržaj zaglavlja, autentikacijski mehanizam, vrijednost korisničke oznake i lozinke te najnoviji dohvati flash.*

PARAMETRI POBIRANJA	PROSINAC 2005.	RUJAN 2009.
Neželjeni dio stabla	84	799
Željeni dio stabla	63	212
Alternativno računalo	38	73
Uklanjanje iz URL-a	15	77
Sinonim	8	38
Uloženi dio	6	442

Dohvat vanjskih	5	2
Poček prije dohvata[12]	-	10
Rabi sadržaj zaglavlja	-	1
Autentikacijski mehanizam	-	4
Vrijednost korisničke oznake i zaporke	-	5

Tablica 3. Parametri pobiranja- usporedba 2005. i 2009. godina

Razvojni poslovi

Razvojem i praćenjem tehnologije izrade mrežnih stranica radi se i na unapređivanju tehnologija pobiranja istih. Formati kao što su Java script i Flash koji su donedavno bili nepoberivi, sada se uspješno pobiru, a radi se i na razvoju za sakupljanje novih formata i aplikacija. Unapređenje rada pobirača nužno je zbog prilagodbe novim tehnologijama izrade mrežnih sadržaja (web-sjedišta), tj. nestandardnim načinima uporabe tih tehnologija. Krajem studenog 2009. godine dovršena je verzija 3.2 pobirača koji je poboljšao pobiranje mrežnih mjesta izrađenih u Java scriptu, Flashu, poveznica s hrvatskim dijakriticima te ispravno prikazivanje dijakritika u metapodacima. Osim navedenog, unapređenje jezgre sustava nužno je i zbog učinkovitijeg pobiranja građe koja je zaštićena nekim mehanizmom autentikacije i autorizacije (u prilog tomu govori čak 30% serijskih mrežnih publikacija koje, iz navedenih razloga, još uvijek nisu pobrane).

Kako bi sadržaj arhiva bio što dostupniji korisnicima, osim pretraživanja prema naslovu publikacije i URL-u, implementirana je tražilica za pretraživanje indeksa njegovog cjelovitog tekstualnog sadržaja. Rezultati pretraživanja daju najviše 2 rezultata iz jedne publikacije i grupirani su prema publikacijama. Tražilica je lagana je za korištenje čemu pridonosi i struktura zaglavlja rezultata pretraživanja koja je slična korisnicima već poznatih tražilica, poput Googlea. Uključuje ukupan broj rezultata (datoteka) prema zadanom upitu, broj datoteka dostupnih putem rezultata pretraživanja, postavljeni upit i vrijeme trajanja pretraživanja.
Prikazani rezultati pretraživanja sadrže naslov iz naslovne oznake (eng. tag) na izvornoj mrežnoj stranici (ukoliko on postoji), u protivnom samo nekoliko prvih riječi iz sadržaja; napomenu da se radi o inačici iz digitalnog arhiva (a ne izvornom web-sjedištu); datum i vrijeme arhiviranja; naslov publikacije (pod kojim je identificirana i bibliografski obrađena u katalogu) i jedan do dva ulomka iz teksta publikacije koji sadrže pojam (pojmove) iz upita.
Zbog indeksiranja javila se i potreba za nabavom prostora na poslužitelju. Tijekom prosinca 2009. godine nabavljen je dodatni diskovni prostor za pohranu (?) te su sustavi za arhiviranje i indeksiranje razdvojeni. U indeksu se, za sada, nalazi 22.000.000 dokumenata, a koliko diskovnog prostora zahtijeva takav jedan indeks vidi se i po njegovoj veličini koja iznosi 1 TB što je gotovo polovica cijelog sustava za arhiviranje (2.4 TB.)

Razvijene su i dodatne usluge i funkcije u svrhu učinkovitijeg rada i upravljanja procesom arhiviranja pri čemu valja spomenuti sljedeće:

- automatska provjera dostupnosti mrežne građe – svakog prvoga u mjesecu putem elektroničke pošte sustav dojavljuje popis neaktivnih URL-ova;

[12] Parametri pobiranja *poček prije dohvata, rabi sadržaj zaglavlja, autentikacijski mehanizam, vrijednost korisničke oznake i zaporke te dohvati flash* nisu predmet analize iz 2005. godine.

- usporedba sličnosti pobiranja – dojava o pobiranju sličnih datoteka (instanci) radi izbjegavanja višestrukog arhiviranja iste instance;
- velika pobiranja – dojava o pobiranjima koja su prekoračila maksimalnu zadanu veličinu (200 MB);
- zauzeće diskovnog prostora – dojava o postotku zauzeća diskovnog prostora kako bi se na vrijeme stiglo dobaviti dodatni (odnosno smanjila učestalost pobiranja neke mrežne građe koju možda nije potrebno pobirati redovitom učestalošću);
- warning DAMP/fatal error - dojava o neuspjelom ishodu pobiranja koje se eventualno može popraviti s promijenjenim parametrima.

U tijeku je izrada novih web stranica arhiva kao i plan prikupljanja nacionalne .hr domene kako su napravile i drugi arhivi temeljeni na selektivnom i/ili tematskom arhiviranju mrežne građe.

Zaključak

Rezultati usporedne analize prema vrstama građe pokazali su da je najveći porast zabilježen kod integrirajuće građe, dok je najveći pad kod serijskih publikacija. Također, usporedba prema kategorizaciji serijskih publikacija pokazala je da su najbrojniji časopisi državnih službi, dok je najveći pad uočen kod novina i magazina/revija. Tiskanu inačicu najviše sadrže serijske publikacije. Prema sadržajnim kategorijama u arhivu je najzastupljenija kategorija *ekonomija i poslovanje*.

Rezultati analize sadržaja u arhivu pokazali su da se najviše publikacija pobire *ručno* s dubinom pobiranja postavljenom na 2, dok je najčešće korišten parametar *neželjeni dio stabla*. Najviše publikacija objavljeno je na nacionalnoj .hr domeni.

Selektivno arhiviranje hrvatskih mrežnih publikacija koje imaju dugoročnu kulturološku ili istraživačku vrijednost značajno je kako za knjižničnu i informacijsku djelatnost nacionalne knjižnice Republike Hrvatske, tako i za njenu znanstveno-istraživačku i razvojnu djelatnost. Nastavak suradnje NSK i SRCA na daljnjem razvoju sustava u svrhu proširenja njenih funkcionalnosti i ostvarenja postavljenih zadataka na razvojnih poslovima zasigurno će pomoći u nastojanjima trajnog očuvanja mrežnih sadržaja kao dijela kulturne nacionalne baštine.

Literatura

Klarin, Sofija; Pigac, Sonja. Hrvatske daljinski dostupne elektroničke serijske publikacije // Vjesnik bibliotekara Hrvatske 4, 43(2000.[i.e. 2001.]). Str. 156-167.

ISBD(CR) : međunarodni standardni bibliografski opis serijskih publikacija i druge neomeđene građe. Zagreb : Hrvatsko knjižničarsko društvo, 2005.

Milinović, Miroslav; Topolšćak, Nebojša. DAMP II : Digitalni arhiv mrežnih publikacija : nova funkcionalnost – novi planovi. // 9. seminar Arhivi, knjižnice, muzeji : mogućnosti suradnje u okruženju globalne informacijske infrastrukture: zbornik radova / uredile Mirna Willer i Ivana Zenić. Zagreb : Hrvatsko knjižničarsko društvo, 2006. Str. 40-55.

Pigac, Sonja; Buzina, Tanja. Selektivno arhiviranje hrvatskog weba : rezultati i otvorena pitanja // 9. seminar Arhivi, knjižnice, muzeji : mogućnosti suradnje u okruženju globalne

informacijske infrastrukture: zbornik radova / uredile Mirna Willer i Ivana Zenić. Zagreb : Hrvatsko knjižničarsko društvo, 2006. Str. 28-39.