PROBLEMATIKA DOLGOROČNE HRAMBE SPLETNIH STRANI

Mitja Dečman

Oddano: 30. 7. 2010 – Sprejeto: 19. 1. 2011

Pregledni znanstveni članek
UDK 005.921.1"746":004.774(497.4)

Izvleček

Svetovni splet je danes porazdeljena zbirka spletišč, ki so dostopna po internetu kjerkoli na svetu. Njegova vsebina se neprestano spreminja, nove informacije zamenjujejo stare, kar povzroča izgubljanje velike količine podatkov in posledično izgubo znanstvene, kulturne in druge dediščine. Neopazno je pogosto ogrožena tudi pravna varnost oseb. Kako rešiti pomembne podatke v svetovnem spletu in jih dolgoročno hraniti in ohraniti za prihodnost, je danes velik izziv. Kljub temu da so nekatere dobre prakse že razvite, je vprašanje celovite rešitve za celoten nacionalni okvir še vedno neodgovorjeno. Članek predstavlja problematiko hrambe spletnih strani tako s tehničnega kot tudi z organizacijskega vidika. Opredeljuje faze zajema spletnih strani in njihove hrambe, pri čemer prikaže rešitve in dobre prakse, ki so na voljo po svetu, ter strateške okvire, ki so jih nekatere razvite države vzpostavile za reševanje te problematike. Predlaga nekaj konceptualnih korakov, ki bi jih bilo treba opredeliti v Slovenji kot okvir za vse ustvarjalce dokumentov v spletnem okolju ter tako prispevati k ozaveščanju glede problematike in lajšanju težav vsem, ki se s to problematiko srečujejo danes ali pa se bodo v prihodnosti.

Ključne besede: digitalni arhivi, dolgoročna hramba, spletne strani, zajemanje

Review article
UDC 005.921.1"746":004.774(497.4)

Abstract

The World Wide Web is a distributed collection of web sites available on the Internet anywhere in the world. Its content is constantly changing: old data are being replaced which causes constant loss of a huge amount of information and consequently the loss

DEČMAN, Mitja. Problems of long-term preservation of web pages. Knjižnica, 55(2011)1, pp. 193–208

of scientific, cultural and other heritage. Often, unnoticeably even legal certainty is questioned. In what way the data on the web can be stored and how to preserve them for the long term is a great challenge. Even though some good practices have been developed, the question of final solution on the national level still remains. The paper presents the problems of long-term preservation of web pages from technical and organizational point of view. It includes phases such as capturing and preserving web pages, focusing on good solutions, world practices and strategies to find solutions in this area developed by different countries. The paper suggests some conceptual steps that have to be defined in Slovenia which would serve as a framework for all document creators in the web environment and therefore contributes to the consciousness in this field, mitigating problems of all dealing with these issues today and in the future.

Keywords: web archives, long-term preservation, web pages, harvesting

1 Spletne strani nekoč, danes in v prihodnosti

Splet je od svojega začetka konec osemdesetih na področju interneta zavzel najpomembnejše mesto in zasenčil druge storitve, kot sta oddaljen dostop in elektronska pošta, ter zaobjel storitev prenosa datotek. S svojo pomembnostjo je marsikje dosegel prenos papirne oblike v elektronsko (danes na primer praktično ni javnega medija, ki poleg primarne klasične oblike ne bi vključeval še spletne) in pogosto nad klasično obliko prevladal in jo izpodrinil (slovenski *Uradni list* npr. izhaja le še v elektronski obliki). Primere za to lahko iščemo v zasebnem ali javnem sektorju. Hkrati je splet prešel v drugo fazo, t. i. splet 2.0, ki je celotno družbo prestavil v novo sfero, katere blogi, wikiji, socialna omrežja in druge storitve so danes zanjo nepogrešljive. Spletne strani so postale tako pomembne, da nekaterim zmanjšujejo prihodek klasičnega poslovanja in porajajo odločitve, da se spletne vsebine lahko začne zaračunavati, saj sta trg in interes dovolj velika. Take odločitve so že sprejeli ali pa jih napovedujejo nekateri znani časopisi in medijske hiše, med drugim *The Times* in *The Sun* (Clark, 2009), v Sloveniji pa o tem že razmišljajo na spletnem portalu 24ur.com.

Ko je leta 1990 Tim Berners-Lee s svojim projektom ENQUIRE zagnal idejo spleta, je tega definiral kot »splet hiperbesedil, ki jih pregledujemo z brskalnikom ob uporabi arhitekture strežnik-odjemalec« (Berners-Lee in Cailliau, 1990). Spletne strani so bile takrat hiperbesedilo ter nič več in nič manj. Zaznamki, ki so jih vsebovale, so bili zgolj navodila interpretu – brskalniku, kako naj določeno besedilo, ki ga posreduje strežnik, prikaže oz. interpretira na zaslonu odjemalca oz. v oknu brskalnika. Kljub temu da je spletna stran poleg besedila lahko vsebovala tudi druge elemente, na začetku predvsem slike, sta bila vsebina in prikaz na strani strežnika statična in sta se zaporedno prenašala do odjemalca.

Današnje strani tako lahko poleg standardnega besedila vključujejo še druge vizualne elemente, ki so lahko nebesedilni (statične in animirane slike, avdio in video) ter interaktivni (interaktivno besedilo – DHTML, interaktivne slike, npr. Flash, hiperpovezave in obrazci). Poleg vizualnih pa lahko zaznamo še nevizualne oz. skrite elemente, kot so metapodatki, komentarji, skripte (npr. JavaScript), oblikovni podatki (npr. CSS) in drugi elementi. Ugotovimo lahko, da je spletna stran daleč od klasičnega fiksnega dokumenta, kot ga poznamo v poslovanju in upravljamo z ustreznimi elektronskimi sistemi. Razlog tiči v tem, da je bila spletna stran v osnovi oblikovana in razvita po konceptu neprestanega spreminjanja, nadgrajevanja in dopolnjevanja, ne pa kot statičen in fiksen dokument. Koehler (2004) na primer v svoji raziskavi ugotavlja, da je iz nabora strani, izbranih decembra 1996, maja leta 2003 v prvotni obliki in na prvotnem URL-naslovu ostalo le še 33,8 odstotka strani. Zato v okviru tega prispevka dokument obravnavamo kot katerokoli zaključeno obliko vsebine, ne pa zgolj kot klasično definirano dokumentarno ali arhivsko gradivo.

V prihodnosti bomo prav gotovo priče novemu razvoju in spremembam, ki bodo splet vključile v prav vsako poro našega življenja. Vprašanje pa je, kdo bo imel interes pregledovati spletne strani naše sedanjosti in preteklosti. Nedavna raziskava kaže, da so to lahko zelo različne skupine ljudi, od novinarjev, pravnikov, detektivov, razvijalcev spletnih strani in rešitev, javnih uslužbencev in raziskovalcev (Smith, 2009).

2 Dolgoročna hramba spletnih strani

2.1 Življenjski cikel spletnih strani in problematika hrambe

Spletne strani so eden od najpogostejših načinov ponujanja informacij v sodobni informacijski družbi. Danes ima praktično vsako podjetje dostop do svetovnega spleta in pogosto tudi svoje spletne strani, ki jih izdela samo ali s sodelovanjem zunanjega izvajalca in ponuja strani v svojih strežnikih ali pri najetih gostiteljih. V Sloveniji je bilo leta 2008 več kot 71 odstotkov podjetij z deset ali več zaposlenimi, ki so imela svojo spletno stran (Zupan, 2008). Njihov primarni interes je seveda doseči čim večjo gledanost, kar pomeni čim boljšo uvrstitev v največjih svetovnih spletnih iskalnikih. Dolgoročna hramba in kontinuiteta vsebine sta drugotnega pomena, saj so strani osredinjene predvsem na trenutno stanje in trenutnega uporabnika.

Tudi v javni upravi so z razvojem e-uprave spletne strani pridobile na veljavi, predvsem kot vir informacij, v zadnjem desetletju pa tudi kot vir e-storitev, ki se

jih države trudijo ponujati svojim državljanom. Hkrati se je z uveljavitvijo koncepta informacij javnega značaja povečala uporaba spletnih strani tudi za te namene, ponekod tudi zaradi zniževanja stroškov, saj je informacije javnega značaja ceneje in lažje objaviti na spletu kot pa posredovati glede na vsako zahtevo posebej. Hkrati se je spremenila zakonodaja, ki določene informacije javnega značaja zahteva (ali predvideva) v obliki spletnih strani.

Dolgoročno elektronsko hrambo lahko obravnavamo z več vidikov. Z arhivskega stališča pomeni hrambo dokumentarnega in arhivskega gradiva, daljšo od petih let in do preteka roka hrambe, če je le-ta določen. V knjižničnem okolju pa taka hramba pomeni hrambo za vse večne čase, kar lahko enačimo s hrambo arhivskega gradiva. Članek obravnava dolgoročno hrambo predvsem z arhivskega stališča, kjer se vsebina obravnava kot dokazna vrednost določene aktivnosti, procesa in temu primerno poudarja pomen ohranjanja verodostojnosti, avtentičnosti in nespremenljivosti.

Tako kot pri vsaki drugi (elektronski) hrambi je treba tudi pri dolgoročni hrambi spletnih strani vključiti aktivnosti, kot so selekcija, zajem, hramba, dostopanje in ohranjanje avtentičnosti. Dolgoročna hramba spletnih strani mora biti aktivnost od začeta do konca ter mora vključevati celoten življenjski cikel spletne strani (Farrell, 2010). Pri tem se srečujemo s specifiko spletnih strani, ki pomeni zahtevno dolgoročno hrambo, pri čemer lahko izpostavimo, da:

- se spletne strani v primerjavi s klasičnimi dokumenti pogosto spreminjajo, kar otežuje tudi kontinuiteto (lahko se spremeni spletni naslov strani – vsebina ostane, ni pa več dostopna z drugih povezav, ali pa vsebina – stran ostane dostopna, a je vsebina spremenjena). Barksdale in Berman (2007) navajata, da je povprečna življenjska doba spletne strani le 75 dni.
- so elementi spletnih strani lahko zelo raznoliki, kar pomeni, da je ob dostopu do spletne strani treba znati interpretirati prav vsakega od njih; pri tem se včasih zahtevajo še posebne aplikacije, ki interpretacijo elementov omogočajo (npr. Adobe Flash);
- elementi spletnih strani niso vedno na isti lokaciji (npr. slika z drugega spletnega strežnika, vključena v spletno stran na podlagi URL-naslova, del strani, vključen na podlagi ukaza (ang.) include page ipd.), kar povzroča dodatno težavo pri zajemu in hrambi;
- je vsebina spletne strani pogosto rezultat podatkovne poizvedbe zbirke podatkov, kar z vidika dolgoročne hrambe povzroča še dodatne težave, saj so zbirke podatkov dinamične narave.

Dolgoročno hrambo lahko gledamo z dveh vidikov – kot hrambo pri ponudniku in pri odjemalcu. Ponudnik je tisti, ki spletno stran izdela. Pri tem ima možnost izbrati obliko strani, njeno umeščenost v spletišče in ravnanje z njo. Pogosto je omejen ali zavezan zahtevam različnih pravnih aktov (glede informacij javnega

značaja, avtorskih pravic ipd.). Vendar pa lahko izbira ali določi, kakšna bo vsebina in oblika strani in kako bo z njo ravnal. Izbere lahko tudi možnost, da je stran zgolj duplikat dokumenta, ki je ustrezno obravnavan v procesu upravljanja z dokumenti. Za svojo pravno varnost lahko ponudnik izbere, katere strani so pomembne in katere ne, ter temu ustrezno prilagodi hrambo. Ponudnik ima možnost, da hrani izvorno obliko, programsko kodo in zbirke podatkov, ki tvorijo podatke in videz spletnih strani, vendar pa mora ohraniti tudi njihovo funkcionalnost, kar je z vidika dolgoročne hrambe zelo težavna naloga (mogoča je rešitev v obliki popolne emulacije). Ob teh aktivnostih lahko uporabi koncept upravljanja življenjskega cikla (ang. information lifecycle management – ILM), kar pomeni, da se določijo načini izdelovanja, upravljanja, hrambe in uničenja ter nabor in roki hrambe. Pri tem lahko uporabi določene koncepte upravljanja z dokumenti, čeprav je tam fiksnost dokumenta poglavitna prvina, ki pa ne ustreza spletnemu okolju (PoWR Team, JISC, 2008).

Na strani odjemalca pa je stvar nekoliko drugačna. V takem primeru na vsebino in obliko ne moremo vplivati, uporabimo lahko le svoje najboljše možnosti, da na izbranih naborih spletišč in strani pridobimo najustreznejše podatke in funkcionalnost. Pri tem smo odvisni še od interpretacije strežnika in odjemalca, saj vedno zajamemo samo okolju in trenutku prilagojeno stran na logični (interpretirani) ravni.

2.2 Tehnične ovire

Glede dolgoročne hrambe spletnih strani je treba razmišljati o hrambi na logični ravni zapisa. Logična raven strani je rezultat predstavitve fizične ravni, torej interpretacije fizičnega zapisa elementov v strežnikih. Interpretacijo pa lahko izvede tako spletni strežnik kot tudi odjemalec. To lahko predvsem na strani odjemalca glede na nedoločene standarde tako HTML- kot tudi CSS-formata, če se osredotočimo zgolj na besedilo, pomeni zelo različne načine logične oblike strani, kar lahko pomembno vpliva na razumevanje in verodostojnost predstavitve. Take interpretacije so odvisne od naprave odjemalca (odjemalcev v mobilnih telefonih, dlančnikih, interaktivni televiziji in računalnikih) in programske opreme, kot so brskalniki in njihovi dodatki (Internet Explorer, Mozzila Firefox, Safari, Opera, če naštejemo zgolj najbolj uveljavljene in množico njihovih različic).

Logična predstavitev dokumenta tako vključuje vsebino in videz ter dinamično in interaktivno komponento. Vsebina je običajno besedilna, pogosto z dodanimi slikami, ki se ob interpretaciji glede na zahteve za videz postavijo na ustrezno mesto. Besedilo je najpogosteje zapisano v HTML- ali XHTML-obliki in uporab-

ljeno tudi za indeksiranje na strani spletnih iskalnikov. Besedilo se včasih pojavlja kot del ugnezdenih elementov (objektov Flash ali slik), kar nekoliko oteži zajem. Del besedila predstavljajo zaznamki, ki so lahko element interpretacije videza ali pa po vzoru XML-koncepta predstavljajo opis določenega besedila (možnost metapodatkov). Najpomembnejši so deli besedila, ki predstavljajo neposredne ali posredne (npr. povezave na podlagi slike) hiperpovezave na druge strani. Ball (2010) tako ugotavlja, da lahko besedilo obravnavamo na različnih ravneh. Videz vključuje oblikovanje vsebine in je določen z zaznamki in (ali) dodatnimi oblikovnimi elementi, kot je npr. »Cascading Style Sheet« (CSS). CSS predstavlja od vsebine ločen način določitve oblikovanja vsebine, kar je bilo v preteklosti sicer neposredno vključeno v sam HTML dokument, kasneje pa se je z razvojem oblikovanja spletnih strani ločilo v ločen dokument, tj. CSS oblikovno datoteko. Težava pri CSS je, da je več različic in da različica ni nikjer eksplicitno zapisana, prav tako pa različni brskalniki uporabljajo različne načine interpretacije ali je stran sestavljena tako, da na podlagi podatka o brskalniku prilagodi interpretacijo videza. Interaktivnost (obrazci, povezave) se najpogosteje implementira na podlagi skript, elementov CSS, Flash ipd. Ball (2010) deli interaktivnost glede na lokacijo izvajanja (stran odjemalca, npr. Javascript, in stran strežnika), glede na distribucijo virov (vsebino v enem ali več strežnikih) in stopnjo interaktivnosti. Dinamična komponenta je najtežje obvladljiva, saj pri hrambi dokument kopiramo v prostor zunaj njegovega funkcionalnega okolja, kar pomeni, da se dinamična komponenta običajno izgubi. Tako kot trdi Tibbo (2003), pa lahko zgolj zaslonski posnetek take strani popolnoma izgubi svojo najznačilnejšo lastnost.

Posebna zgodba je Splet 2.0 saj uporablja za svoje delovanje rešitve različnih ponudnikov, je osredotočen na sodelovanje in komunikacijo, torej v kompleksne in raznovrstne storitve in pripadajoče tehnične rešitve, predvsem pa je izredno dinamičen (npr. Wikiji, blogi, socialna omrežja ipd.). Zato je potrebno ustrezno obvladovanje različnih tehnologij, preoblikovanje formatov, primernih za uporabnike, v formate, primerne za hrambo (npr. pretočni video v fiksni zapis), obvladovanje varstva osebnih podatkov (npr. avtor bloga in vsi avtorji komentarjev), avtorskih pravic (ena stran, veliko avtorjev), s čimer so se srečali tudi pri Nacionalni knjižnici Avstralije pri projektu PANDORA.

Ugotovimo lahko, kako je pomembno, da se pred zajemom odločimo, katere elemente spletnih strani želimo hraniti, in se zavedamo, da če bo takih elementov več, bosta zajem in hramba težja, raven tveganja pa bo višja. Zato moramo določiti tiste lastnosti dokumentov, ki so za nas najpomembnejše (ang. significant properties).

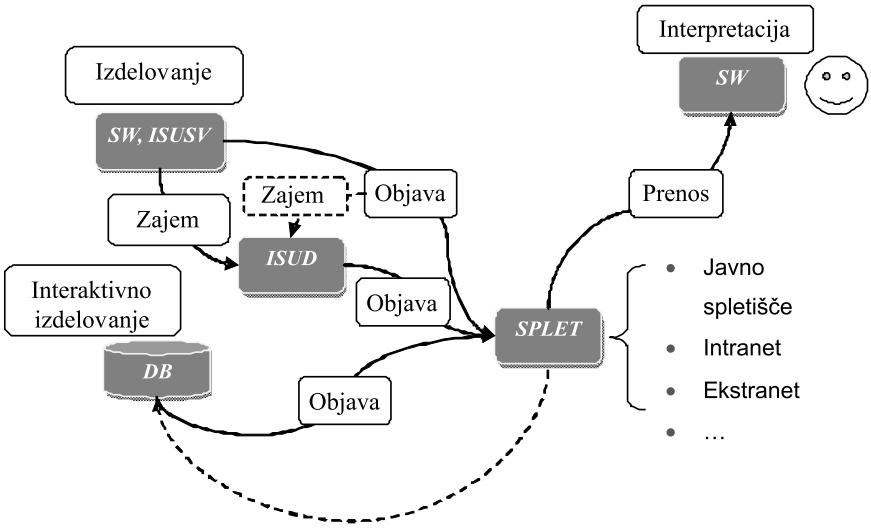
2.3 Priprava na dolgoročno hrambo in zajem

Zajem spletnih dokumentov zahteva, da se preučijo in zajamejo vsi potrebni spletni dokumenti, vključno s spletnimi stranmi, z vsebinskimi elementi, datotekami ipd. Če zajem izvaja ponudnik vsebine, poskusi zajeti še dokaze in evidence izvedenih aktivnosti izdelovanja in vzdrževanje spletnih dokumentov in spletnih mest, ob upoštevanju izvedenih transakcij (manualnih in samodejnih), konteksta in strukture. Zajem je mogoč na več različnih načinov, vendar dandanes nobeden od teh nima prevladujoče vloge zaradi omejitev in težav, ki jih povzročajo dinamična narava spletnih dokumentov, visoka stopnja povezovanja, frekventne spremembe vsebine in vizualna komponenta spletnih dokumentov. Sprejeto pa je, da so poglavitni element zajema prav politika in postopki, ki jih je treba razviti in ki določajo identifikacijo spletnih dokumentov, odbiranje, zajem, tehnične pogoje itd. Zajem na strani ponudnika je pogosto odvisen od načina izdelovanja spletnih strani in drugih dokumentov, objavljenih v spletu. Če za to uporabimo informacijski sistem za upravljanje spletnih vsebin (ISUSV) ali pa objavljamo dokumente, ki so prej evidentirani v informacijskem sistemu za upravljanje z dokumenti (ISUD), je mogoče zagotoviti evidenco zajema in sam zajem dokumentov v ISUD. Objava dokumenta lahko namreč poteka tako (Slika 1):

1. Od izdelovalca prek zajema v ISUD, nato na mesto spletne objave (npr. izdelovalec izdela spletni dokument, ga evidentira v ISUD in postavi v spletišče). Pri tem je treba poskrbeti, da ISUD zajame še dodatne metapodatke, kot sta npr. datum objave v spletišču, URI ipd.
2. Od izdelovalca neposredno na mesto spletne objave (npr. izdelovalec z ustrezno programsko opremo, kot je ISUSV, neposredno izdela spletno stran ali uredi njeno vsebino). Pri tem lahko ISUSV poskrbi za pripravo ustreznih metapodatkov. Nato je treba spletni dokument zajeti v ustreznem ISUD neposredno ali pa posredno (metapodatki) na podlagi ISUSV, če ta nadomešča, dopolnjuje ali se povezuje z ISUD.
3. Spletni dokument se izdela na podlagi interakcije uporabnika z virom podatkov, najpogosteje zbirko podatkov (npr. uporabnik na podlagi prijave z digitalnim potrdilom omogoči, da sistem izmed podatkov v zbirki podatkov pripravi podatke za uporabnika).

Omenjene možnosti se morajo zaradi zajema in hrambe realizirati tako, da se spletni dokument evidentira in zajame v ustrezni evidenci, kar pomeni:

- nastanek dokumenta v ISUD in objavo na spletnem mestu ali v ISUSV, pri čemer se lahko izvede ustrezna pretvorba v spletno obliko dokumenta;
- nastanek v ISUSV ali z drugo programsko opremo za izdelovanje spletnih dokumentov in nato zajem v ISUD; dandanes so že na voljo tako imenovani konektorji, kot so programski vmesniki (API), ki omogočajo komunikacijo med sistemi in prenos ustreznih podatkov na ustrezen način.



Slika 1: Načini objave spletnih dokumentov

Načini zajema spletnih strani tabelarično predstavlja tudi dvajset smernic avstralske zvezne države New South Wales (Guideline, 2009).

Kljub temu pa se danes večina raziskav osredinja na zajem (žetev) na strani uporabnika/odjemalca (ang. harvesting), pri čemer s posebnimi programi – pajki (ang. crawler, harvester) pregledujejo (žanjejo) svetovni splet (celoto ali njegov del) in zajemajo strani, ki jih ti programi najdejo. Pri tem ločimo selektivni princip, periodično žetev, tematski zajem in odlagalni princip. Zajem spletnih dokumentov je tako najpogosteje izveden za zagotavljanje dostopa brez povezave (ang. offline) in posredno zagotavljanja hrambe nekega trenutnega stanja dokumenta. Pri tem se seveda srečamo s pomembnimi vidiki, kot so:

- raznovrstnost strežnikov,
- podvajanje vsebin,
- velik obseg zajemanja,
- veliko formatov,
- slaba kakovost podatkov ipd.

Nedavno poročilo angleške organizacije Digital Curation Center izpostavlja tudi časovno konsistentnost zajema (Ball, 2010). Težava je predvsem v tem, da se zajem posamezne spletne strani j odvija v nekem trenutku t_j in traja do trenutka $t_j + i$. Temu sledi zajem naslednje spletne strani t_{j+1}. Kadar je obseg spletnih strani, ki jih želimo v nekem trenutku zajeti, velik in želimo za zajeto celoto ohraniti konsistentnost (tudi delovanje vseh povezav), se v celotnem času zajema nobena od

strani ne sme spremeniti. Nevarnost nekonsistentnosti je tako tem večja, čim večji je nabor strani in čim večja je njihova frekvenca spreminjanja.

Po svetu so že na voljo orodja, ki samodejno žanjejo in shranjujejo spletne strani v določene repozitorije. Znani primeri so Heritrix (http://crawler.archive.org), NEDLIB Harvester (http://www.csc.fi/sovellus/nedlib/), HTTrack Website Copier (http://www.httrack.com/), IBM WebCrawler (http://www-03.ibm.com/systems/i/software/http/services/webcrawler.html), MetaProduct Offline Explorer (http://www.metaproducts.com/) idr. (Diessen in Steenbakkers, 2002), med katerimi pa so se nekateri izkazali za neučinkovite in se ne uporabljajo več (npr. NEDLIB Harvester). Ta orodja vključujejo posebne programe – pajke, ki začnejo pregledovanje spletnih strani na točno določenem spletnem naslovu in ob uporabi hiperpovezav sami preiskujejo naslednje strani po vnaprej določenih pravilih. Težava se lahko pojavi, če je hiperpovezava umeščena v objekt, ki ni HTML-formata (npr. Flash), zaradi česar povezave pajek ne najde in dokumenta ne zajame. Zato je v primeru, ko so spletni dokumenti zajeti s pajkom (npr. znotraj nekega spletišča), nujna predhodna analiza hiperpovezav, npr. pretrganih hiperpovezav in, če je mogoče, osamelih spletnih strani (do njih ne vodi nobena hiperpovezava).

Ball (2010) poudarja, da med slabosti žetve spadajo tudi odpravljanje napak in nepravilnosti na strani, ne pa v zajetem arhivu (npr. avtor strani krši avtorske pravice, kar se pozneje popravi, medtem pa spletni arhiv neustrezno vsebino zajame in ponuja v spletu), sledenje spremembi dostopnosti (stran je javno objavljena, pozneje pa plačljiva, kar arhivski strežnik lahko krši), napačno rangiranje (arhivski strežnik dobi višji rang v spletnih iskalnikih kot prvotna stran, ker pomeni zmanjšanje zaslužka, ki se npr. nanaša na izvirno spletno stran).

2.4 Dolgoročna hramba in dostop

Dolgoročna hramba je posebna oblika hrambe, ki je primerna ali zahtevana za vso dokumentarno gradivo, katerega rok hrambe je več kot določeno število let (slovenska zakonodaja predpisuje pet let). Pri dolgoročni hrambi se je treba predvsem odločiti, kakšna bo strategija glede hrambe spletnih strani. Mogočih je več scenarijev (Strodl in Rauber, 2005):

- ohranitev izvirnika;
- migracija strani v katerega od uveljavljenih formatov za dolgoročno hrambo;
- standardizacija (npr. minimalni HTML-format).

Kako načrtovati dolgoročno hrambo, je odvisno predvsem od tega, komu je hramba namenjena, kakšne tipe spletnih strani želimo hraniti, katere so signifikantne

lastnosti, ki jih želimo ohraniti (glede vsebine, strukture, videza, interaktivnosti ipd.) in katere so druge zahteve (zanesljivost, pristnost, uporabnost ipd.). Pomemben parameter je seveda tudi strošek take hrambe.

Najpomembneje je, da ohranjamo berljivost in uporabnost gradiva, kar je pri spletnih straneh še posebej težka naloga. Tako mora hramba poskrbeti tudi za potrebne komponente, ki ohranijo zahtevano funkcionalnost spletnega dokumenta (skripte, programske dodatke, vtičnike in brskalnik), obvezno pa njihov opis, različico ipd. Nekatere funkcionalnosti je tudi smiselno ukiniti oziroma fiksirati (npr. števec dostopov ne bo prikazoval pravega stanja v času hrambe). Hiperpovezave se morajo ohraniti tako, da se tudi znotraj hrambe omogoči njihova uporaba. Mogočih je več strategij (Masanès, 2006):

- **Lokalizacija na datotečni sistem:** Koncept uporablja preslikavo iz izvirnega okolja spletnega dokumenta v okolje zajema in hrambe. Spletni dokumenti se prenesejo v enaki relativni strukturi, kot so bili na voljo na izvirnem spletnem mestu. Kot metapodatek zapiše pravilo preslikave hiperpovezave iz prvotne oblike v lokalizirano, v dokumentu pa se v skladu s tem pravilom preslikave izvede (npr. pravilo http://www.organ.si/* v file://hramba/20091108/*, kar pomeni, da se npr. http://www.organ.si/images/logo.png spremeni v file://hramba/20091108/images/logo.png). Slabosti tega sistema sta omejenost z datotečno strukturo sistema hrambe (skalabilnost) in neposredna odvisnost od okolja datotečnega sistema skozi čas (težave z migracijo). Hkrati spreminjamo izvirne spletne dokumente, kar povzroča težave pri zagotavljanju avtentičnosti in celovitosti.
- **Migracija oblike:** Uporablja preoblikovanje spletnih strani ali sklopa spletnih strani v drugo obliko, npr. PDF. Pri tem se seveda lahko ohranijo tudi določene hiperpovezave, vendar se izvirna struktura popolnoma spremeni. Tak način je primeren predvsem za posamezne spletne dokumente, ki so bili izdelani neodvisno od spletnega okolja.
- **WARC-datoteka:** Je ena priljubljenejših metod in hkrati v skladu s standardom ISO 28500 (2009), med drugim jo uporabljajo tudi v danskem nacionalnem arhivu (Kristiansen, 2006). WARC-datoteka vključuje nabor spletnih dokumentov, zajetih s spleta, s pripadajočimi metapodatki za vsak zajeti dokument. Taka datoteka je lahko precej velika, saj običajno vključuje spletne dokumente celotnega spletišča in zaradi učinkovitosti pogosto dodatno vključuje še ločen indeks vsebine, ki vzdržuje mapiranje med hiperpovezavami znotraj spletnih dokumentov, vključenih v WARC objekta, in drugimi objekti v hrambi. Hacox-Yu (2009) pa opozarja, da je implementacija zahtevna, neposrednega dostopa do shranjenih spletnih dokumentov ni, formata pa protivirusni programi za zdaj ne prepoznajo. Kristiansen (2006) v svojem eksperimentu ugotavlja, da je za učinkovito rabo WARC-datotek nujno pozneje zgraditi še indeksno datoteko.

3 Dolgoročna hramba spletnih strani po svetu in pri nas

Ohranjanje spletnih vsebin je bilo v preteklosti že opaženo kot pomembno, saj so raziskovalci že leta 1996 začeli prve resne poskuse v okviru projekta Internet Archive in projekta švedske nacionalne knjižnice. Do leta 2003 so v okviru projekta Royal Library's Kulturarw3 initiative večkrat izvedli žetev švedskega spleta in zbrali za 5,5 terabajta (TB) podatkov ter okoli 185 milijonov strani. Med pionirje lahko štejemo tudi Internet Archive, ki je od leta 1996 zbral že več 100 TB podatkov, količina pa mesečno naraste za 12 TB (Hakala, 2004). Pomembno vlogo je v okviru razvoja tega področja odigral International Internet Preservation Consortium (IIPC), ki je pripomogel k internacionalizaciji tega problema, sodelovanju med državami in predvsem nacionalnimi knjižnicami. Je tudi glavni akter pri razvoju in standardizaciji WARC formata (ISO 28500, 2009). Decembra 2009 pa je IIPC objavil register spletni arhivskih zbirk (21 zbirk) svojih članic kot enotno vstopno točko uporabnikom arhiviranih spletnih vsebin.

Danes lahko govorimo o štirinajstletnih izkušnjah, ki pa še vedno niso privedle do jasne situacije na tem področju. Čeprav nekatere države že uspešno izvajajo žetev spleta in hrambo, prav tako so ponekod uspešno združene različne ustanove na ravni države, pa v Sloveniji ni tako. Tudi na tem področju je nujno povezovanje in sodelovanje ustanov. Predvsem pri zajemu in hrambi spletnih dokumentov je pomembna tudi vloga nacionalne knjižnice, katere pogosti nalogi (npr. v Avstraliji ter na Novi Zelandiji in Nizozemskem) sta zajemanje in hramba spletnih dokumentov ali kar celotnih spletnih mest z vsebino, zanimivo tako za raziskovalce kot tudi za splošno populacijo. Sem pogosto spadajo tudi spletne publikacije, ki se pojavijo v spletni obliki kot edini obliki pojavitve. Tako je britanska nacionalna knjižnica vzpostavila program spletne hrambe, ki se izvaja od leta 2004 in vključuje zajem spletnih strani z okoli 5000 spletišč z vsebino raziskovalne vrednosti ter na podlagi dogovora s ponudniki teh vsebin (glede avtorskih pravic). Dostop zajetih vsebin je na voljo na podlagi storitve UK Web Archive. Predvideva se, da bodo leta 2011 dobili mandat hrambe vseh prosto dostopnih spletnih publikacij v Veliki Britaniji. Kongresna knjižnica v ZDA je na podlagi projekta MINERVA vzpostavila nekaj spletnih arhivskih zbirk izbranih spletnih strani (tematski pristop k hrambi), ki so jim dodani kataloški podatki v obliki metapodatkovnega standarda MODS (angl. metadata object description standard). Delo nadaljujejo na področju eksperimentiranja z formatom METS, metapodatkovnimi opisi PREMIS ipd. Nacionalna knjižnica v Avstraliji v sistemu PANDORA (iniciativa je bila začeta leta 1996) hrani spletne strani glavnih spletnih časopisov in spletne publikacije, ki se nanašajo na Avstralijo (izbirni pristop hrambe). PANDORA danes vključuje okoli 90.000 strani in več kot štiri terabajte podatkov. Nacionalna knjižnica Nizozemske (Koninklijke Bibliotheek) je leta

2006 razvila sistem e-depot (odlagalni princip) in začela selektivno žetev spleta. Danes ta zbirka vključuje okoli 2500 spletišč, leta 2013 pa naj bi jih 10.000.

Poleg velike pozornosti za zajem in hrambo spletnih dokumentov domen, ki so pomembne za zagotavljanje kulturne in zgodovinske komponente nekega okolja, se marsikje ukvarjajo tudi s problemom zagotavljanja zajema in hrambe spletnih dokumentov kot dela dokumentarnega gradiva. Primer so recimo zahteve ameriškega arhiva NARA (2005), ki se bolj osredotoča na tveganje in koristi upravljanja s spletnimi dokumenti. Taka je implementacija U.S. Department of Health & Human Services, kjer se osredotočajo na določanje rokov hrambe različnih oblik spletnih dokumentov (Web, 2007). Smernice o tem, kaj in kako zajemamo in hranimo spletne dokumente, pa je določil tudi nacionalni arhiv Avstralije (Archiving, 2001). Prav na Novi Zelandiji so v navodilih za upravljanje spletnih dokumentov izpostavili razliko med spletnim dokumentom in spletno publikacijo, v katerem dokument obravnavajo z vidika upravljanja z dokumenti ter v navezavi z ESUD in ESUSV (Guide, 2009).

4 Slovenija in njena spletna dediščina v prihodnosti

V Sloveniji trenutno še nimamo nobenega strateškega dokumenta o dolgoročni hrambi spletnih strani. *Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih* – ZVDAGA (2006) kot arhivsko gradivo šteje izvode spletnih objav (40. člen), predvideva spletni dostop in uporabo javnega arhivskega gradiva (63. člen) ter zahteva tudi, da vlada predpiše obseg in način posredovanja arhivskega gradiva v svetovni splet. Težava je seveda natančno ugotoviti, kaj vse se šteje kot spletna objava. Podobne težave drugod po svetu omenja tudi Chi-Shiu Lin (2010), ki opozarja, da ima ponekod oseba, ki določa, kaj šteje kot spletna objava, pri tem diskrecijsko pravico.

Uredba o varstvu dokumentarnega in arhivskega gradiva – UVDAG (2006) zahteva, da se pri zajemu izvirnega dokumentarnega gradiva v digitalni obliki zajema in izdela metapodatke, kar v okviru spletnih strani pomeni vse povezave, ter da se pri zajemu zajame izvorna oblika zapisa posamezne spletne strani, vključno s povezavami in z drugimi podatki, nujnimi za njen prikaz ter zajeto sliko zaslona spletne strani v obliki zapisa za dolgoročno hrambo (11. člen). Hkrati določa nabor informacij in gradiva, ki ga morajo arhivi brezplačno objaviti v spletu, in omogoča prilagajanje oblike shranjenega gradiva v obliko, primerno za objavo v spletu (125. člen).

Z vidika hrambe spletnih publikacij *Zakon o obveznem izvodu publikacij* – ZOIPub (2006) v 2. členu določa, da so tudi spletne strani elektronske publikacije, tj. *elektronske knjige, elektronski časopisi in časniki, spletne strani in podobno, ki so objavljeni na fizičnih nosilcih (npr. na magnetnih trakovih, kasetah, disketah, CD-jih ipd.) ali so dostopni v računalniških omrežjih ali svetovnem spletu)*, predmet predaje obveznega izvoda nacionalni depozitarni organizaciji, tj. NUK-u (podobne so zahteve npr. na Novi Zelandiji, v Avstraliji in na Nizozemskem). Načine, pogostost in obseg predaje natančneje določa *Pravilnik o vrstah in naboru elektronskih publikacij* (2007), ki pod definicijo spletne publikacije (3. člen) vključuje spletne elemente spleta 2.0, s čimer si postavlja zahtevno vlogo, kajti prav taki viri povzročajo vsem svetovnim spletnim arhivom največje težave. Kontradiktorno pa se tem vrstam publikacije odpoveduje v 6. členu. Kot tehniko zajema omenjeni pravilnik opredeljuje princip žetve domene *.si* in še nekaterih drugih domen, podobno kot to opredeljuje npr. avstrijska nacionalna knjižnica.

Verjetno bo treba za dostop do zajetih spletnih strani vzpostaviti nov sistem, ki bo poleg zdajšnjega kooperativnega online bibliografskega sistema in servisi COBISS, v katerem so npr. diplomske in magistrske naloge, in Digitalne knjižnice Slovenije (dLib.si), v katerem so mnoge druge publikacije, ponujal znanstvenoraziskovalni in drugi javnosti dostop do zajetih spletnih vsebin internetne preteklosti in sedanjosti. Ker pa vsak tak sistem dolgoročne hrambe zahteva prav vse, kar npr. določa OAIS, bo treba razmisliti o optimizaciji stroškov, človeških virov na podlagi sodelovanja vseh zainteresiranih ustanov v državi ter pripravljenosti in podpore (finančne in strateške) ministrstva za kulturo.

5 Zaključek

Ugotovimo lahko, da je splet kompleksno in dinamično okolje za dolgoročno elektronsko hrambo. Kljub desetletni zgodovini dejavnosti in mnogim rešitvam za zajem in hrambo spletnih strani pa zaradi nenehnega razvoja spleta in nedorečenosti glede pomena take hrambe dela še zdaleč ni konec. Ugotavljamo, da popolne spletne hrambe ni in da je treba določiti ključne elemente in funkcionalnosti, ki jih želimo ohraniti pri dolgoročni hrambi.

V Sloveniji bomo morali razmisliti, kakšna je vloga Arhiva RS ter Narodne in univerzitetne knjižnice pri zajemu in hrambi spletnih dokumentov. Strategijo hrambe bo treba razdeliti na dva dela, pri čemer bi moral Arhiv RS pokrivati hrambo, ki jo zahteva ZVDAGA, torej hrambo na strani ponudnikov spletnih vsebin, ter vzpostaviti ustrezen red tudi na tem področju. Menimo namreč, da se organizacije ne zavedajo, da mnogi spletni dokumenti in publikacije spadajo med dokumentarno in arhivsko gradivo. NUK bi moral skrbeti za tiste vsebine in

spletišča, ki so zanimiva za celotno javnost in niti ne izključno kot arhivsko gradivo. Medsebojno sodelovanje pa bi odpravilo podvajanje dejavnosti, prekrivanje dela in povečalo transparentnost delovanja ter znižalo stroške. Ker je orodij v spletu že kar nekaj, mnoga od njih pa so na voljo kot odprta koda in brezplačno, bi bilo hkrati smiselno uporabiti dozdajšnje izkušnje in se učiti iz dobrih izkušenj drugih.

Navedeni viri

1. *Archiving web resources: guidelines for keeping records of web-based activity in the Commonwealth Government*. (2001). Canberra: National archives of Australia. Pridobljeno 12. 10. 2010 s spletne strani: http://www.naa.gov.au/Images/archweb_guide_tcm2-903.pdf
2. Ball, A. (2010). *Web archiving (version 1.1).* Edinburgh: Digital Curation Centre.
3. Barksdale, J. in Berman, F. (2007, 16. maj). Saving our digital heritage. *The Washington Post*. Pridobljeno 21. 7. 2010 s spletne strani: http://www.washingtonpost.com/wp-dyn/content/article/2007/05/15/AR2007051501873.html
4. Berners-Lee, T. & Cailliau, R. (1990). *Worldwideweb: proposal for a hypertext project.* Pridobljeno 7. 7. 2010 s spletne strani: http://www.w3.org/Proposal.html
5. Chi-Shiou, L. (2010). Librarian-initiated publications discovery: how do digital depository librarians discover and select web-based government publications for state digital depositories? *Government Information Quarterly, 27* (3), 292–304.
6. Clark, A. (2009, 6. avgust). Rupert Murdoch plans charge for all news websites by next summer. *The Guardian*. Pridobljeno 7. 7 2010 s spletne strani: http://www.guardian.co.uk/media/2009/aug/06/rupert-murdoch-website-charges
7. Diessen, R. J. van in Steenbakkers, J. F. (2002). *The long-term preservation study of the DNEP project: an overview of the results*. Amsterdam: IBM; The Hague: Koninklijke Bibliothek.
8. Farrell, S. (2010). *A guide to web preservation.* London: UKOLN / ULCC.
9. *Guide to managing web records.* (2009). Wellington: Archives New Zealand. Pridobljeno 21. 7. 2010 s spletne strani: http://archives.govt.nz/sites/default/files/e_to_Managing_Web_Records__Publication_Copy__3.pdf
10. *Guideline 20 – Keeping web records.* (2009). Kingswoods: State of New South Wales, State Records Authority. Pridobljeno 1. 7. 2010 s spletne strani: http://www.records.nsw.gov.au/recordkeeping/government-recordkeeping-manual/guidance/guidelines/guideline-20-in-this-guideline

11. Hakala, J. (2004). Archiving the web: European experiences. *Program, 38* (3), 176–183.

12. Hockx-Yu, H. (2009). *Web archiving tools: an overview.* V *JISC, the DPC and the UK Web Archiving Consortium Workshop Missing links: the enduring web.* Pridobljeno 27. 11. 2009 s spletne strani: http://www.dpconline.org/docs/events/090721MissingLinksHockxYu.pdf

13. *ISO 28500:2009, Information and documentation – WARC file format.* (2009). Geneva: ISO.

14. Koehler, W. (2004). A longitudinal study of web pages continued: a report after six years. *Information Research, 9* (2). Pridobljeno 7. 7. 2010 s spletne strani: http://informationr.net/ir/9-2/paper174.html

15. Kristiansen, M. A. (2006). *Digital preservation using the WARC file format.* Copenhagen: Department of Computer Science, University of Copenhagen (DIKU). Pridobljeno 27. 11. 2009 s spletne strani: ftp://130.225.96.5/pub/diku/semantics/papers/D-548.pdf

16. Masanès, J. (2006). Web archiving: issues and methods. V J. Masanès (Ur.), *Web archiving* (str. 234). Berlin: Springer.

17. *NARA guidance on managing web records*. (2005). College Park (MD): The National Archives and Records Administration. Pridobljeno 10. 9. 2010 s spletne strani: http://www.archives.gov/records-mgmt/pdf/managing-web-records-index.pdf

18. Pravilnik o vrstah in izboru elektronskih publikacij za obvezni izvod. (2007). *Uradni list RS, št.* 90.

19. *PoWR – The preservation of web resources handbook.* (2008). Bath: PoWR Team, JISC. Pridobljeno 6. 6. 2010 s spletne strani: http://www.jisc.ac.uk/media/documents/programmes/preservation/powrhandbookv1.pdf

20. Smith, C. (2009). *Context and content: delivering coordinated UK web.* Pridobljeno 7. 7. 2010 s spletne strani: www.dpconline.org/vendor-reports/…/398-0907smithmissinglinks.html

21. Strodl, S. in Rauber, A. (2005). *Selecting preservation strategies for web archives*. V *5th International web archiving workshop (IWAW05), Vienna, Austria.* Pridobljeno 7. 8. 2010 s spletne strani: http://www.ifs.tuwien.ac.at/dp/presentation/iwaw_2005_Strodl.pdf

22. Tibbo, H. R. (2003). On the nature and importance of archiving in the digital age. *Advances in Computers*, *57*, 1–67.

23. Uredba o varstvu dokumentarnega in arhivskega gradiva. (2006). *Uradni list RS,* št. 86.

24. *Web records policy & guidance*. (2007). Washington: U.S. Department of Health & Human Services. Pridobljeno 15. 10. 2009 s spletne strani: http://www.dhhs.gov/web/policies/webrecords.html

25. Zakon o obveznem izvodu publikacij. (2006). *Uradni list RS,* št. 69.

26. Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (ZVDAGA). (2006). *Uradni list RS,* št. 23.

27. Zupan, G. (2008, 1. oktober). Uporaba interneta v podjetjih z 10 ali več zaposlenimi osebami, Slovenija, 1. četrtletje 2008. *Statistični urad RS*. Pridobljeno 6. 6. 2010 s spletne strani: http://www.stat.si/novica_prikazi.aspx?id=1912

Dr. Mitja Dečman je zaposlen na Fakulteti za upravo, Univerza v Ljubljani.
Naslov: Gosarjeva 5, 1000 Ljubljana
Naslov elektronske pošte: mitja.decman@fu.uni-lj.si