Arhiviranje weba

Lučić, Lucija

SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

SMJER ARHIVISTIKA

Ak. god. 2019./2020.

Lucija Lučić

Arhiviranje weba

Diplomski rad

Mentor: prof. dr. sc. Hrvoje Stančić

Zagreb, 2020.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

Sadržaj

1. UVOD

Arhiviranje weba relativno je nova pojava u našem svijetu, ali isto bi se moglo reći i za sam web: u posljednjih nekoliko desetljeća online sadržaji zagospodarili su svim aspektima ljudskih privatnih i poslovnih života. Bilježenje i pohrana prošlih i sadašnjih web sadržaja neophodno je za budućnost u kojoj će ti materijali moći poslužiti kao kulturni i povijesni izvori, poslovni spisi, dokazi u pravnim procesima i sl. Ovaj se rad uglavnom fokusira na teorijski dio arhiviranja weba i najprije donosi kratak povijesni pregled samog weba i arhivske djelatnosti zasebno. Drugo poglavlje objašnjava osnovnu terminologiju, vrste i alate arhiviranja weba. U trećem poglavlju ističu se pravni okviri arhiviranja weba. Četvrto poglavlje nabraja najznačajnije organizacije i usluge koji se bave ovom djelatnošću. Peto govori o nagloj evoluciji arhiviranja weba u vremenskom razdoblju od samo nekoliko godina. Šesto opisuje sve veću važnost arhiviranja weba za prošlost, sadašnjost i budućnost, a sedmo raspravlja o brojnim tehničkim i društvenim izazovima s kojima se suočavaju arhivi weba. Osmo i deveto poglavlje ističu primjere arhiva weba u Hrvatskoj, SAD-u, UK, Danskoj i Francuskoj uz opise njihovih funkcija, opsega pohranjenog sadržaja i najčešćih izazova s kojima se suočavaju. Zaključno se opisuju inicijative arhiviranja društvenog weba.

Andrews (2013.) donosi kratak pregled povijesti interneta: krajem 60-tih godina prošloga stoljeća u SAD-u je stvoren Advanced Research Projects Agency Network (ARPANET), odnosno prvi funkcionalni prototip interneta koji je omogućio komunikaciju više računala unutar jedne mreže. Robert Kahn i Vinton Cerf u sedamdesetima su stvorili Transmission Control Protocol and Internet Protocol (TCP/IP), komunikacijski model koji je postavio standarde za prijenos podataka između različitih mreža. U osamdesetima ARPANET usvaja TCP/IP, a u devedesetima Tim Berners-Lee stvara World Wide Web. Andrews upozorava da internet i WWW nisu sinonimi: web je najčešći način pristupanja sadržajima na internetu kroz web mjesta i hiperlinkove/poveznice.

Što se tiče arhivske djelatnosti, može se reći da je u nekom obliku postojala od samih početaka ljudskih civilizacija. Prema definiciji arhiva Hrvatske enciklopedije, čovječanstvo je čak i tisućama godina prije nove ere prepoznavalo važnost čuvanja službenih dokumenata. Smatra se da su prvi arhivi nastali za vrijeme drevnih civilizacija (Egipat, Babilon, Perzija, Stara Grčka, Rim itd.) kada su se važni spisi čuvali u hramovima i palačama. Stvaranje i djelovanje modernih arhiva započinje nakon završetka Francuske revolucije: narodni arhiv u

Parizu osnovan je 1794. Tada su se počeli stvarati pravni i znanstveni okviri rada arhiva i zemlje diljem svijeta počele su prepoznavati potrebu za otvaranjem vlastitih državnih arhiva.

U suvremenom svijetu dolazi do jedinstvenog spoja jedne veoma stare i jedne veoma nove djelatnosti: arhiviranja i stvaranja sadržaja na webu. Internet je u posljednjih nekoliko desetljeća postao glavni izvor informacija, vijesti i razonode za moderno društvo: prema podacima iz 2018. koje prenosi World Economic Forum, na internetu se svaki dan stvara preko 2,5 kvintilijuna (10^{18}) bajtova podataka. Arhiviranje weba od presudne je važnosti za očuvanje današnje kulture, znanja i događaja jer informacije na webu često imaju vrlo kratak životni vijek i zauvijek će nestati ako se ne zabilježe na vrijeme. Web izvori smatraju se digitalnom baštinom i jednako su važeći, vrijedni očuvanja i potrebni čovječanstvu kao i bilo kakav drugi oblik baštine. Devedesetih godina prošloga stoljeća različite su institucije postale svjesne važnosti očuvanja digitalnih informacija i počele arhivirati svoj web sadržaj (tj. gotovo odmah nakon osnivanja WWW-a). Sve je počelo 1996. s arhivom interneta (Internet Archive, IA) o kojemu će više biti riječi u četvrtom poglavlju. Prije toga je potrebno objasniti osnovne termine s područja arhiviranja weba.

2. ŠTO JE ARHIVIRANJE WEBA?

Konzorcij za očuvanje internetskoga sadržaja (International Internet Preservation Consortium, IIPC) definira arhiviranje weba kao proces prikupljanja dijelova World Wide Weba i očuvanja zbirki u arhivskom formatu uz omogućavanje pristupa arhivima i uporabu njihovih sadržaja. Prema ISO/TR 14873:2013 standardu koji definira statistiku, terminologiju i kriterije kvalitete u arhiviranju weba, arhiv weba je potpuni skup web resursa arhiviranih tokom vremena koji sadrži jednu ili više zbirki. Arhivi weba mogu prikupljati sadržaje na tradicionalan način, tj. prihvaćanjem dokumenata vladinih agencija i izdavača u sklopu zakona o obveznom primjerku (više o tom zakonu u trećem poglavlju), a mogu i samostalno prikupljati sadržaje korištenjem pobirača. (Niu, 2012.) ISO standard definira pobiranje (engl. *harvesting*) kao proces pregledavanja i kopiranja resursa koji se može provoditi na razini domene ili selektivno korištenjem automatskih alata (robota tj. pobirača, odnosno alata za indeksiranje weba). Domensko crawlanje/pobiranje teži prikupljanju cjelokupnog sadržaja s jedne ili više vršnih domena te njihovih poddomena. Vršne domene su najviše razine domena u sustavu, uključujući državne (.fr, .uk, .hr) i generične (.com, .net, .org). Ovakvo pobiranje uglavnom rezultira stvaranjem arhiva weba širokog opsega. Opseg arhiva weba je veličina arhiva ili zbirke koju određuju pravni okviri ili politika nabave ustanove koja provodi arhiviranje. Selektivno crawlanje/pobiranje teži prikupljanju resursa odabranih prema određenim kriterijima poput znanstvene važnosti, relevantnosti za temu ili učestalosti ažuriranja resursa. Ovdje su obuhvaćena i selektivna pobiranja događaja koja su vremenski ograničena (tj. imaju određen datum završetka) i teže prikupljanju resursa povezanih s jedinstvenim događajima poput političkih izbora, sportskih natjecanja ili prirodnih katastrofa. U usporedbi, domenska pobiranja imaju širi opseg i provode rjeđe (najčešće jedanput godišnje) od selektivnih pobiranja koja imaju uži opseg i provode se češće (od nekoliko puta godišnje do više puta dnevno).

Prema ISO/TR 14873:2013, metapodaci su podaci koji opisuju kontekst, sadržaj i strukturu digitalnih objekata te njihovo upravljanje tijekom vremena. Metapodaci mogu biti deskriptivni, strukturalni i administrativni. Deskriptivni metapodaci opisuju intelektualan sadržaj digitalnog objekta, strukturalni opisuju kako objekti zajednički stvaraju logičke jedinice, a administrativni su informacije potrebne za ispravno upravljanje objektima. Metapodaci se mogu dodatno podijeliti na tri kategorije: metapodaci konteksta ili provenijencije (opisuju životni ciklus resursa, uključujući povezane entitete i procese), tehnički (opisuju tehničke karakteristike digitalnog objekta, npr. format) i metapodaci o

pravima (definiraju vlasništvo i pravne uvjete korištenja objekta). Selekcija je proces odlučivanja o značaju određenog skupa resursa za arhiv weba prema politici razvijanja zbirke. Niu (2012.) ističe razliku u vrednovanju gradiva pri ručnom i automatskom prikupljanju web sadržaja. Vrednovanje prema objektivnim kriterijima kao što su tip medija ili domene može se lako provoditi pobiračem. Kvalitetni i/ili popularni sadržaji mogu se otkriti praćenjem dolaznih poveznica i posjetitelja, brojem pregleda i ocjenama korisnika. Vrednovanje prema temi i sadržaju pak zahtijeva ručni odabir koji je vremenski i financijski zahtjevan, zbog čega se provodi samo kod arhiviranja weba malog opsega. Prema Niu, načela provenijencije (informacije o izvoru spisa) i prvobitnog reda (redoslijed prema kojemu je stvaratelj organizirao spise) postoje i u arhiviranju weba: provenijencija opisuje URL-ove, stvaratelje (vlasnike web mjesta), poslovne transakcije ili namjenu web mjesta, a prvobitni red opisuje vanjsku i unutarnju strukturu arhiviranog web mjesta. Vanjska struktura prikazuje kako je web mjesto posloženo u odnosu na druga mjesta, a unutarnja prikazuje kako su komponente web mjesta posložene u odnosu jedna na drugu.

Arhiviranje weba nije djelatnost koja bi trebala biti ograničena samo na arhive i knjižnice: bilo tko može stvoriti osobne zbirke za bilježenje svih izvora koji su im važni ili zanimljivi. Najjednostavniji i najmanjkaviji oblik ovakve pohrane jest stvaranje snimke zaslona (engl. *screenshot*). Time se bilježi točno ono što je korisnik vidio u određenom trenutku, ali gubi se svaka funkcionalnost web stranice koja je pretvorena u statičnu sliku. Weigle (2018.) uspoređuje stvaranje snimke zaslona s arhiviranjem web stranica. Navodi da snimke zaslona mogu poslužiti kao brzi podsjetnik na izvorni izgled web stranice, ali njihova funkcionalnost završava na tome jer su potpuno statične i interakcija sa zabilježenom web stranicom nije moguća ni na koji način (klikovi, klizanje, otvaranje poveznica itd.) Arhivi weba bilježe potpuni sadržaj web stranica (koliko im to dopuštaju njihove mogućnosti) koji uključuje izvorni HTML, umetnute slike, stilove i JavaScript.

2.1. Pobirači

Cloudflare definira pobirače kao programe (robote) koji preuzimaju i indeksiraju sadržaj cijelog interneta. Cilj im je naučiti sadržaj gotovo svake stranice na webu kako bi se informacije mogle dohvatiti po potrebi. Oni se najviše koriste u arhiviranju weba i indeksiranju za svrhe pretraživanja. Kako bi pronašao sve relevantne informacije, taj alat počinje s popisom poznatih web stranica (ishodišni URL-ovi, engl. *seed*) i slijedi poveznice s

tog popisa na druge stranice. Zatim dodaje nove URL-ove na svoj popis, prikuplja njihove sadržaje i počinje slijediti nove otkrivene poveznice. Ovaj se proces može nastaviti unedogled. Pobirači odlučuju koje će stranice najprije pobirati prema relativnoj važnosti web stranica (broju ostalih stranica koje su povezane s početnom), broju posjeta i ostalim čimbenicima koji ukazuju na vjerojatnost da ta stranica sadrži važne informacije. Ako se web stranica često posjećuje i citira, može se pretpostaviti da sadrži informacije visoke kvalitete i autoriteta. Budući da se sadržaj na webu konstantno ažurira, uklanja ili premješta, pobirači periodički posjećuju već indeksirane stranice kako bi prikupili najsvježiju verziju sadržaja.

Prema ISO/TR 14873:2013, opseg pobiranja određen je postavkama i parametrima koje za svaki skup ishodišnih URL-ova određuju resurse koje pobirač treba prikupiti (tj. broj web mjesta i dubinu pobiranja) te potrebnu učestalost ponovnog posjećivanja mjesta. Opseg može biti na širini čitave vršne domene ili pak ograničen na samo jednu datoteku. Također uključuje i pristojnost pobirača (broj zahtjeva koji se šalju poslužitelju u jednoj sekundi ili minuti), suglasnost s robots.txt datotekama i filtere za izbjegavanje zamki. Zamka za pobirače je jedna ili više web stanica koja može uzrokovati njegovo rušenje ili beskonačno praćenje referenci na druge resurse s malo ili nimalo vrijednosti. Zamke se mogu postaviti namjerno kako bi se spriječilo pobiranje određenih resursa ili slučajno, npr. kad pobirač slijedi datume kalendara u beskonačnost. Kako bi prikupili sadržaj, pobirači poslužitelju šalju zahtjev za dopuštenjem na koji on možda iz različitih razloga neće odgovoriti ili će dopustiti pobiranje samo pod određenim uvjetima. Vlasnici web mjesta možda neće dozvoliti prečesto pobiranje svojih sadržaja radi potencijalnog preopterećenja poslužitelja i/ili povećanja troškova zbog potražnje za većom propusnošću. Još jedan razlog može biti taj što vlasnik ne želi da određene web stranice budu pronađene bez da se prvo plati naknada ili stvori korisnički račun. Prema Bouard (2014.), brzina takvih alata može uvelike varirati: neki posjećuju i preko 200 web stranica u sekundi, što može preopteretiti poslužitelj ili izazvati probleme kod ljudskih korisnika kojima se to web mjesto neće pravilno učitavati na preglednicima. Pobirači i korisnici različito opterećuju poslužitelje jer se različito ponašaju, primjerice pobirač unutar jednog pobiranja svaku stranicu posjećuje samo jednom (iako mogu uslijediti ponovna pobiranja istih mjesta kako bi se ažurirale izmjene), ali zato nastoji posjetiti svaku pojedinu stranicu unutar ciljanog web mjesta. Suprotno tome, ljudski korisnici posjećuju samo ono što im je relevantno i često posjećuju iste stranice više puta. Pobirači predstavljaju veći „napor“ za poslužitelje web mjesta jer stvaraju novu sesiju za svaku stranicu koju posjete i ne koriste kolačiće (engl. *cookies*), što može uvelike usporiti rad poslužitelja i učitavanje web

stranica. Bouard napominje kako bi arhivi weba trebali uzeti u obzir potencijalne probleme s pojasnom širinom (engl. *bandwidth)* različitih poslužitelja i prilagoditi ponašanje pobirača svojim klijentima.

2.2. Robots.txt

ISO/TR 14873:2013 definira robots.txt kao standard/protokol koji sprečava pobirače da pristupe određenim dijelovima nekog web mjesta ili mjestu u cjelini. Robots.txt datoteke također se mogu koristiti za davanje zahtjeva ili uputa pobiračima npr. obvezan minimalni vremenski razmak između dva uzastopna pobiranja ili umetanje poveznice na mapu web mjesta kako bi se pobiračima olakšalo kretanje. Vrlo je važno istaknuti da robots.txt datoteke nisu pravno obvezujuće, što znači da ih pobirači mogu zanemariti bez posljedica. IIPC upozorava da i zlonamjerni softveri također mogu zanemariti robots.txt ograničenja. Također, robots.txt datoteke su javno dostupne, što znači da bilo tko može točno vidjeti koje dijelove web mjesta vlasnik namjerava zaštititi od pobirača. IIPC napominje da robots.txt datoteke često mogu dovesti do smetnji pri točnom i potpunom arhiviranju sadržaja zbog čega mnogi arhivi weba odlučuju djelomično ili potpuno zanemariti te datoteke. Članovi Konzorcija različito postupaju s robots.txt datotekama: neki ih poštuju u svim slučajevima osim u onima gdje ograničenja sprečavaju bilježenje podataka potrebnih za vjernu reprezentaciju web mjesta (slike, stilovi). Drugi pak posve zanemaruju postojanje robots.txt ograničenja kako bi arhivirali web mjesta što je potpunije moguće. IIPC apelira na vlasnike web mjesta koji žele dozvoliti pobiranje svojih sadržaja da prilagode postavke svojih robots.txt datoteka kako bi se pobiračima dopustio neometan pristup.

Cloudflare objašnjava kako se robots.txt datoteke mogu sastojati od više komponenti (slika 1). *User-agent* je ime dodijeljeno svakoj osobi ili programu koji posjećuje web mjesto. Za ljudske korisnike to uključuje podatke o pregledniku i sustavu kako bi se prikazale kompatibilne verzije web mjesta, a za programe *User-agent* informira administratore o tome kakvi roboti posjećuju i pobiru njihova mjesta. U robots.txt datoteci administratori mogu različitim korisnicima pružiti različite specifične upute, npr. može se postići da se određena stranica pojavi u rezultatima pretraživanja Google-a, ali ne i u Bingu. U primjeru na slici 1, *User-agent* s asteriskom znači da slijedeće upute vrijede za svakog robota koji posjeti web mjesto. Naredba *Disallow* (ne dozvoli) je najčešća i zabranjuje robotima pristup određenim web stranicama ili skupovima stranica. Može se koristiti za blokiranje jedne datoteke (tj.

jedne web stranice, npr. Disallow: /learning/bots/what-is-a-bot/) ili jednog direktorija (npr. Disallow: /__mesa/), a prazan prostor nakon naredbe znači da nema ograničenja tj. pristup je dozvoljen cijelom web mjestu. Skrivanje cijelog web mjesta provodi se stavljanjem znaka „/“ nakon naredbe. Druge naredbe uključuju *Allow* (dozvoli) i *Crawl-delay* (razmak između pobiranja): *Allow* dopušta pristup određenoj stranici ili direktoriju (dok je ostatak web mjesta blokiran), a *Crawl-delay* određuje koliko milisekundi pobirači trebaju čekati između svakog zahtjeva kako ne bi preopteretili poslužitelja.

```
User-agent: *
Disallow: /__esa
Disallow: /__mesa/
Disallow: /__xesa/
Disallow: /__csup/
Disallow: /__xsla/
Disallow: /__xcusp/
Disallow: /__xesa/
Disallow: /__xsla/
Disallow: /lp
Disallow: /feedback
Disallow: /langtest

Sitemap: https://www.cloudflare.com/sitemap.xml
Sitemap: https://www.cloudflare.com/fr-fr/sitemap.xml
Sitemap: https://www.cloudflare.com/de-de/sitemap.xml
Sitemap: https://www.cloudflare.com/es-es/sitemap.xml
Sitemap: https://www.cloudflare.com/pt-br/sitemap.xml
```

Slika 1. Primjer robots.txt datoteke. Izvor: Cloudflare.

Protokol *Sitemaps* (mape web mjesta) pomaže robotima da znaju što uključiti u svoje pobiranje web mjesta. Mapa web mjesta je XML datoteka koja je strojno čitljiv popis svih web stranica unutar mjesta (primjer prikazan na slici 2). Poveznice na mape mogu se uključiti u robots.txt datoteke pomoću ovog protokola. Iako protokol *Sitemaps* pomaže pri osiguravanju da robotima ništa ne promakne dok pobiru web mjesta, oni će nastaviti slijediti svoj tipični proces pobiranja tj. protokol ih ne može natjerati da drugačije prioritiziraju web stranice.

Slika 2. Primjer mape web mjesta (XML datoteka). Izvor: Cloudflare.

3. PRAVNI OKVIRI ARHIVIRANJA WEBA

3.1. UNESCO

UNESCO je 2003. godine prepoznao digitalne materijale kao oblik kulturne baštine i podigao svijest o potrebi njihovog očuvanja objavljivanjem povelje o očuvanju digitalne baštine. Prema povelji, digitalni materijali uključuju tekstove, baze podataka, pokretne i nepokretne slike, audio, grafiku, softver te web stranice. Ističe se prolazna i kratkotrajna priroda digitalnih materijala, kao i značajna potreba za njihovim održavanjem i očuvanjem. Povelja napominje da je svrha očuvanja digitalne baštine njezina dostupnost javnosti, zbog čega je potrebno uspostaviti ravnotežu između korisničkog pristupa sadržaju i istovremene zaštite osjetljivih/osobnih podataka i autorskih prava. Od presudne je važnosti kontinuitet digitalne baštine tj. bilježenje cijelog životnog ciklusa i svih eventualnih ažuriranja i izmjena. Povelja nalaže da pri odabiru materijala za očuvanje prednost treba dati objektima koji su izvorno nastali u digitalnome obliku. UNESCO napominje da digitalni svijet prebrzo evoluira za razvoj adekvatnih strategija očuvanja, što predstavlja prijetnju od gubitka dragocjenih tehničkih, pravnih, znanstvenih, kulturnih, obrazovnih, medicinskih i administrativnih podataka. Ističe važnost podizanja globalne svijesti o važnosti očuvanja digitalnih podataka i stvaranja pravnih okvira, zajedničkih standarda i procedura za ovaj tip djelatnosti.

3.2. Zakon o obveznom primjerku

Dr. Jean Lunn je 1981. kroz UNESCO objavila smjernice za pravnu regulaciju obveznog primjerka. Lunn definira obvezni primjerak kao „državnu obvezu koja zahtijeva da svaka organizacija, komercijalna ili javna, te svaki pojedinac koji proizvodi bilo koji oblik dokumentacije u više kopija, bude dužan odložiti jednu ili više kopija u prepoznatu nacionalnu ustanovu.“ Nakon izdavanja ovih smjernica, mnoge su pravne nadležnosti preuredile svoje postojeće zakone o obveznim primjercima ili počele graditi svoje vlastite. Larivière 2000. godine objavljuje prošireno i ažurirano izdanje smjernica kako bi se uveli novi oblici izdavaštva poput elektroničkih publikacija. Namjera proširene verzije jest pomoći državama pri oblikovanju vlastitih zakona i propisa o novim i konstantno evoluirajućim formatima digitalnog sadržaja. Potreba za zapisivanjem i čuvanjem odloženog materijala jednako je prisutna i u digitalnom i u tiskanom okolišu. Smjernice ističu da online i offline elektroničke publikacije spadaju pod obvezne primjerke i trebaju se odložiti zajedno sa svim

pridruženim materijalom, uključujući i prikladni softver za njihovo reproduciranje. Također se navodi da i dinamične online publikacije trebaju biti uključene u obvezni primjerak.

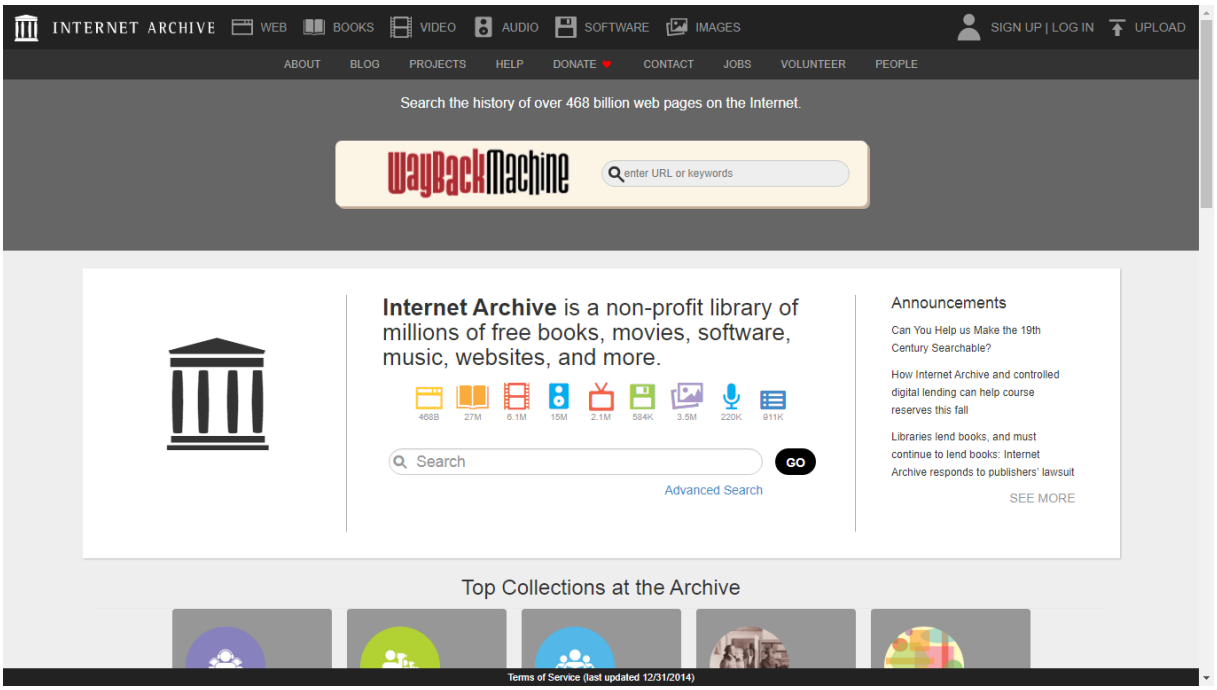
Zakon o obveznom primjerku u Hrvatskoj je obrađen unutar Zakona o knjižnicama i knjižničnoj djelatnosti. Prema članku 39., nakladnici sa sjedištem u Republici Hrvatskoj koji objavljuju ili proizvode publikacije u tiskanom, digitalnom ili drugom obliku dužni su o svom trošku dostaviti nekoliko primjeraka svake publikacije određenim knjižnicama u Hrvatskoj (dva primjerka Nacionalnoj i sveučilišnoj knjižnici, po jedan primjerak sveučilišnim knjižnicama u Splitu, Osijeku, Puli, Rijeci i Mostaru te po jedan primjerak knjižnicama u Zadru i Dubrovniku) najkasnije 30 dana nakon njihova izdavanja. Prema članku 40., nakladnici sa sjedištem u Republici Hrvatskoj koji objavljuju online publikacije (e-knjige, web stranice i sl.) dužni su obavijestiti Nacionalnu i sveučilišnu knjižnicu u Zagrebu o postojanju i objavljivanju publikacije te joj dostaviti građu i pripadajuće metapodatke.

4. ORGANIZACIJE, USLUGE I ALATI

4.1. Internet Archive

Internet Archive (IA) je osnovan 1996. sa sjedištem u San Franciscu (Kalifornija) i naziva se neprofitnom digitalnom knjižnicom web mjesta i ostalih digitalnih kulturnih artefakata. Prema podacima sa službenog web mjesta (slika 3), misija IA je stvaranje univerzalnog pristupa svom znanju (tj. knjižnica i svi njeni sadržaji trebaju biti dostupni svima) i knjižnica u skladu s time svim korisnicima pruža besplatan pristup bez potrebe za registracijom (bez obzira na stručnost ili razlog pretraživanja). IA omogućava besplatno preuzimanje sadržaja te stvaranje korisničkog računa za korištenje usluga digitalne knjižnice. Organizacija danas surađuje s tisućama partnera diljem svijeta i teži prikupljanju svih javno dostupnih podataka s weba, neovisno o jeziku, domeni, temi i sl. U IA je trenutno pohranjeno preko 330 milijardi web stranica, 20 milijuna knjiga i tekstova, 4,5 milijuna zvučnih zapisa, 4 milijuna video zapisa, 3 milijuna slika, 200 tisuća softverskih programa i preko 150 milijardi web zabilješki na više od 40 jezika. Čitava zbirka IA sveukupno zauzima preko 45 petabajta[1] prostora za pohranu (stvaran broj još je veći uzevši u obzir da IA pohranjuje barem po dvije kopije svake datoteke). IA provodi vlastita pobiranja, ali i prihvaća sadržaje koje su samostalno prikupili njegovi partneri. IA je u skladu sa svojom misijom pružanja slobodnog pristupa znanju cijelom čovječanstvu stvorio i projekt Offline Archive kako bi zbirke mogle biti dostupne javnosti usprkos nedostatku internetske veze.

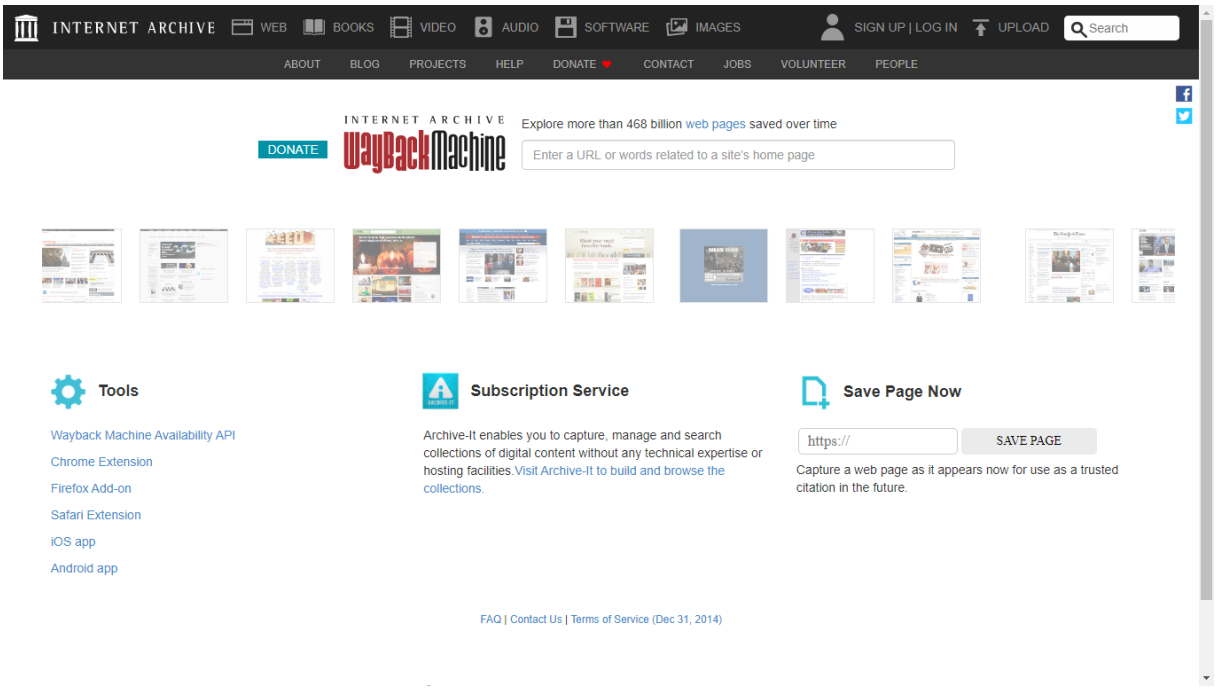
[1] Jedan petabajt (PB) sadrži 10^{15} bajtova odnosno 1.000 terabajta (TB) ili 1.000.000 gigabajta (GB) Beal, V. Webopedia. URL: https://www.webopedia.com/TERM/P/petabyte.html) (28.4.2020.)



Slika 3. Web mjesto IA. Izvor: Internet Archive.

4.1.1. Wayback Machine

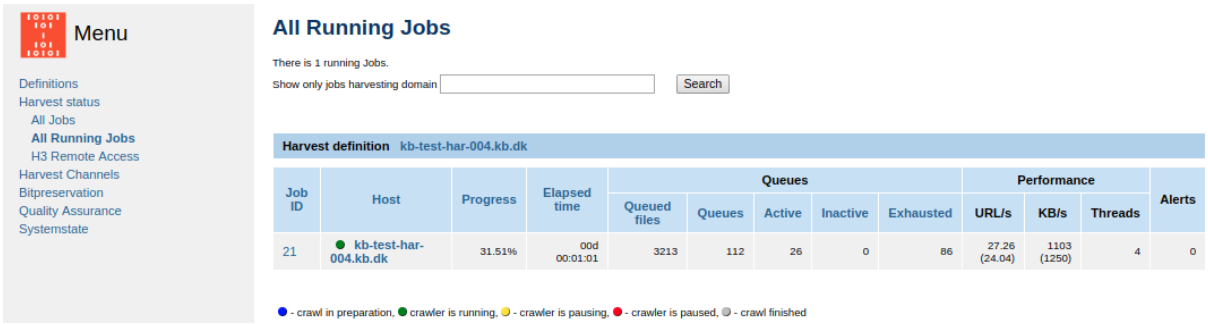
IA je 2001. stvorio arhiv weba pod nazivom Wayback Machine (WM) koji omogućava pretraživanje arhiviranih verzija web stranica prema URL-u. Kroz WM je moguće pretraživati imena web mjesta unutar IA (URL-ove) i odrediti željeni vremenski raspon (u datumima) za pretraživanje. IA u budućnosti namjerava implementirati potpuno tekstualni pretraživač. WM može poslužiti kao legitimna referenca za citiranje u slučaju da originalan sadržaj više ne postoji na živom webu ili ako neki autor želi svojim čitateljima omogućiti pregled iste verzije stranice koju je koristio za svoj materijal. Sučelje WM-a (slika 4) također nudi opciju Save Page Now za spremanje specifične web stranice samo jedanput. Time se taj URL ne dodaje na nikakve buduće popise za pobiranje i ne sprema se ništa više od te jedne stranice. Ako žele da se njihovi sadržaji pohrane u WM, korisnici mogu samostalno pobirati svoje sadržaje i potom ih poslati u IA ili dozvoliti IA pobiračima da prikupe dotične materijale. Kako bi se osiguralo da će pobirači uspjeti pronaći sve resurse, potrebno je osigurati dobru unutarnju i vanjsku povezanost web mjesta i prilagoditi robots.txt datoteke kako ne bi došlo do odbijanja pobirača.



Slika 4. Web mjesto WM. Izvor: Internet Archive.

4.1.2. Heritrix

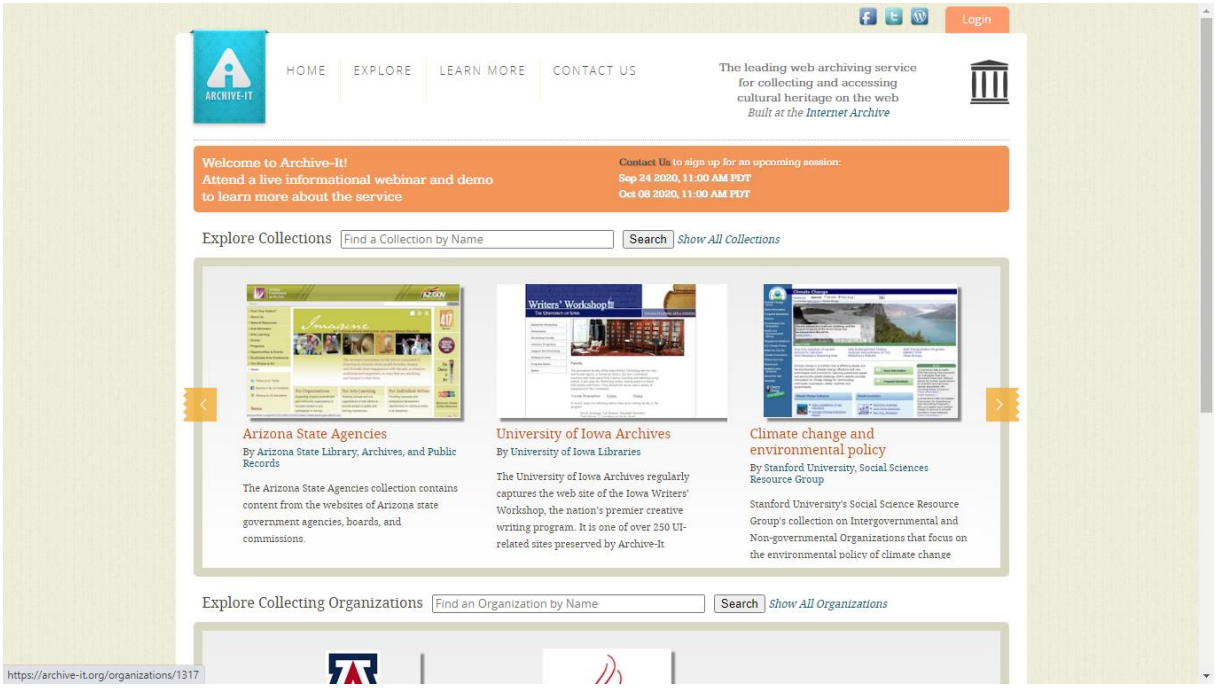
Heritrix je razvijen u IA 2003. kao proširivi pobirač otvorenog koda. Dizajniran je za poštivanje uputa robots.txt datoteka i prikupljanje materijala prema prilagodljivom tempu koji ne bi trebao ometati normalne aktivnosti web mjesta. Heritrix danas koriste IA, arhivi weba u UK, SAD-u, Danskoj, Francuskoj, Hrvatskoj i mnoge druge ustanove. Slika 5 prikazuje korisničko sučelje trenutne verzije Heritrixa.



Slika 5. Korisničko sučelje Heritrixa. Izvor: Sbforge.

4.1.3. Archive-It

Archive-It (A-I) osnovan je 2006. godine u sklopu IA kao fleksibilan sustav arhiviranja weba baziran na pretplatama (slika 6). Partneri ove usluge mogu stvarati vlastite arhive weba, odnosno zbirke koje sadrže samo onaj sadržaj koji je njima osobno važan i potreban. Kroz web aplikaciju prilagođenu korisnicima, A-I partneri mogu samostalno pobirati, katalogizirati, upravljati i pregledavati digitalno rođene sadržaje. Zbirke se udomljavaju na IA podatkovnom centru i dostupne su javnosti kroz potpuno tekstualno pretraživanje. A-I usluga drži barem po dvije kopije svake online zbirke, primarnu i sigurnosnu. Partnerima se pruža mogućnost pobiranja web materijala na deset različitih frekvencija, od svakodnevne do godišnje (svaki URL može imati zasebnu prilagođenu frekvenciju). Dodatno, ustanove mogu započeti pobiranje „na zahtjev“ u slučaju nepredviđenog događaja od velike društvene važnosti. Prema IA, državni arhivi, knjižnice, muzeji i vladine institucije diljem svijeta koriste A-I uslugu.



Slika 6. Web mjesto A-I. Izvor: Archive-It.

4.1.4. Specifičnosti IA

Leetaru (2016.) u članku posvećenom obilježavanju dvadesete obljetnice osnivanja WM-a nastoji razotkriti pozadinske procese iza arhiva weba. Uspoređuje aktivnosti IA s

tradicionalnim arhivima i internetskim pretraživačima: pobirači IA ne funkcioniraju kao pobirači internetskih pretraživača već su sličniji tradicionalnim arhivima. Ističe da današnje operacije pobiranja weba upravljaju velikim brojem standardiziranih pobirača koji dijele zajednički sustav pravila i ponašanja, dok IA prikuplja podatke iz mnogo različitih izvora tj. različite organizacije s različitim pristupima pobiranju prikupljaju materijale i šalju ih u IA. Nadalje, autor ističe da se pristup IA prema robots.txt datotekama promijenio tijekom vremena. Arhiv je prije tvrdio da izbjegava sva web mjesta na kojima su postavljena ograničenja, ali sada to više nije slučaj. Zaključuje da se takva web mjesta pobiru i potom pohranjuju u arhiv kojemu javnost nema pristup (engl. *dark archive*) IA iako se pri pobiranju pojavljuje obavijest da se web mjesto zbog robots.txt datoteka ne može arhivirati. Webopedia definira arhive bez pristupa javnosti kao „arhive kojima ne mogu pristupiti nikakvi korisnici. Pristup podacima je ograničen na nekoliko odabranih pojedinaca ili potpuno zabranjen. Svrha takvog arhiva jest funkcionirati kao repozitorij informacija koja se može koristiti kao sigurnosna mjera za oporavak datoteka u slučaju katastrofe.“ Još jedan prigovor odnosi se na arhiviranje portala sa svjetskim vijestima. Iako IA provodi više aktivnosti posvećenih pobiranju internetskih vijesti, Leetaru ističe da nisu svi portali jednako zastupljeni. Postoji očita sklonost prema zapadnjačkim medijima, odnosno sadržajima na engleskom jeziku u odnosu na ostatak svijeta. Ostali novinski portali nisu ni približno zabilježeni u tolikim količinama i frekvencijama. Autor ovdje navodi primjer 303 arhivske zabilješke američkog portala The Washington Post u usporedbi sa samo 34 zabilješke njemačkog portala Der Spiegel. Leetaru ističe potrebu za dodatnim doprinosom zajednice kako bi se povećao broj samostalnih zbirki i arhiva weba stranih portala.

4.2. IIPC

Konzorcij za dugoročnu pohranu internetskoga sadržaja (International Internet Preservation Consortium, IIPC) osnovan je 2003. godine u Francuskoj gdje se narodna knjižnica udružila s još dvanaest ustanova kako bi zajednički financirale i sudjelovale u projektima arhiviranja weba. Danas su članovi IIPC-a knjižnice, arhivi i baštinske ustanove iz preko 45 zemalja diljem svijeta. Misija IIPC-a je nabaviti i očuvati znanje i informacije s interneta te ih učiniti dostupnim za sadašnje i buduće generacije diljem svijeta, istovremeno promovirajući globalnu razmjenu znanja i međunarodne odnose. IIPC iznosi tri glavna cilja za postizanje svoje misije. Prvi je omogućiti očuvanje zbirke weba kako bi se ona mogla arhivirati, osigurati i biti dostupna na dugi vremenski rok. Drugi je cilj poticanje razvoja i

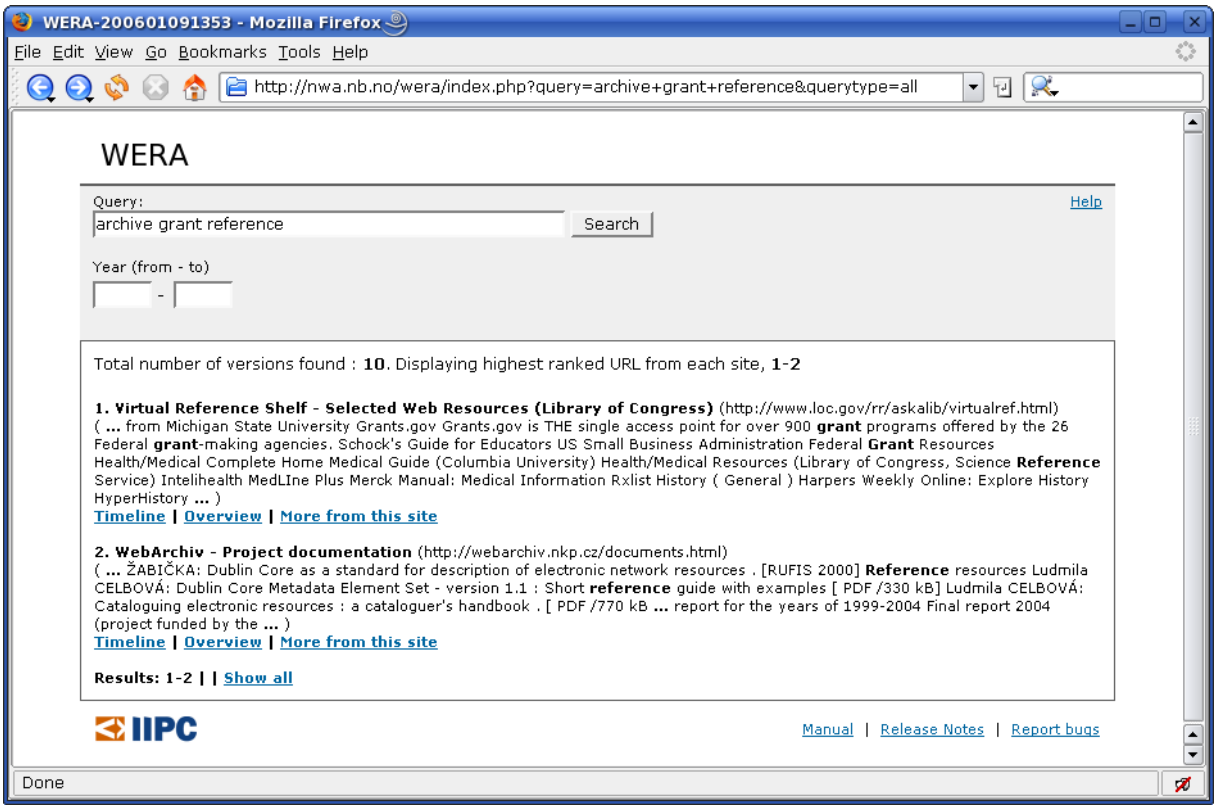
korištenja alata, tehnika i standarda koji omogućavaju stvaranje međunarodnih arhiva weba. Treći je poticanje knjižnica, arhiva i sličnih organizacija diljem svijeta da se počnu ozbiljno baviti arhiviranjem i očuvanjem weba.

4.2.1. NutchWax

NutchWax je pretraživač koji je IIPC dizajnirao 2005. za tekstualno pretraživanje zbirki arhiva weba. Može se koristiti s WM, WERA-om (WEb ARchive Access) ili sličnim aplikacijama i pruža potpuni pristup malim ili srednje velikim zbirkama arhiva weba (500 milijuna dokumenata ili oko 150 tisuća ARC datoteka). Prilikom indeksiranja većih zbirki preporuča se podjela popisa ukupnih arhiviranih datoteka na manje dijelove koji se mogu procesirati jedan po jedan.

4.2.2. WERA

WERA (WEb ARchive Access) je aplikacija za pregledavanje arhiva weba koja pruža pristup zbirkama uz mogućnost punog tekstualnog pretraživanja i navigacije među različitim verzijama web stranice. IIPC je sponzorirao razvoj WERA-e. Slika 7 prikazuje sučelje NutchWaxa i WERA-e.



Slika 7. Sučelje WERA-e s rezultatima NutchWax pretraživača. Izvor: Archive access.

4.3. Formati ARC i WARC

Mike Burner i Brewster Kahle (članovi IA) razvili su 1996. godine ARC format za pohranu arhiviranih web podataka. Format je dizajniran za stvaranje datoteka koje su nezavisne (prikupljenim se objektima može rukovati bez posredovanja nekih drugih datoteka), proširive (za dohvaćanje podataka s različitih mrežnih protokola) i održive (integritet podataka ne smije ovisiti o narednom stvaranju indeksa). Format također mora biti sposoban ulančati više objekata u jednu datoteku: pobrani web podaci prikupljaju se u zbirnim datotekama od po 100 MB radi lakše pohrane i upravljanja. Prema IA, najbolji način dohvata pojedinih objekata iz ARC datoteke jest čuvanje vanjske baze podataka s imenima, veličinom i lokacijom objekata. ARC je dizajniran za čuvanje više web izvora u jednoj zbirnoj datoteci, ali nedostajala su mu neka ključna svojstva. Informacija o lokaciji sadržaja dostupna je u *crawl log* datotekama (automatski stvoreni zapisi aktivnosti pobirača), što znači da se oni ne mogu pronaći ako ta datoteka nedostaje. Također, sadržaji potrebni za prikazivanje jedne web stranice mogu biti raspršeni kroz više različitih ARC datoteka. Potrebno je uključiti ARC i log

datoteke u proces stvaranja indeksa, što je problematično je ARC datoteke ovise o log datotekama na nestandardiziran način. Bilo je nužno riješiti se zavisnosti koje ARC datoteke imaju o vanjskim izvorima kako bi postale uistinu samostalne.

Pojavila se potreba za formatom koji dopušta jednoj datoteci da na jednostavan i siguran način prikupi veliki broj podatkovnih objekata neograničenog tipa za svrhe pohrane, upravljanja i razmjene. Članovi IIPC-a razvili su format WARC (Web ARChive format) za popravak nedostataka ARC formata. WARC je prvi put stvoren 2009. kao međunarodni standard ISO 28500:2009. Druga verzija (koja je zamijenila prvu) stvorena je 2017. kao ISO 28500:2017. Time je WARC postao prvi jedinstveni standard za arhive weba, ali koriste ga i druge organizacije za upravljanje različitim digitalnim objektima. Standardiziranje arhiva weba bilo je nužno iz dva glavna razloga. Prvo, nastala je prijeka potreba za stvaranjem zajedničke terminologije budući da su arhivske ustanove često koristile različite termine za iste aktivnosti ili iste termine za različite aktivnosti. Drugi je razlog bilo i postojanje brojnih tehničkih poteškoća koje je bilo potrebno jasno adresirati: dokumenti su na webu raspršeni preko različitih izvora, imaju nejasne granice (npr. vrlo velika web mjesta ili potreba za arhiviranjem samo pojedinih dijelova određenog mjesta) i komponente koje se mogu promatrati kao zasebni dokumenti. Kroz standardiziranje su se mogle pojasniti semantičke definicije te razviti sporazumi o statistici, pokazateljima, najboljim praksama vrednovanja i boljem razumijevanju inicijativa arhiviranja weba.

4.4. WS-DL

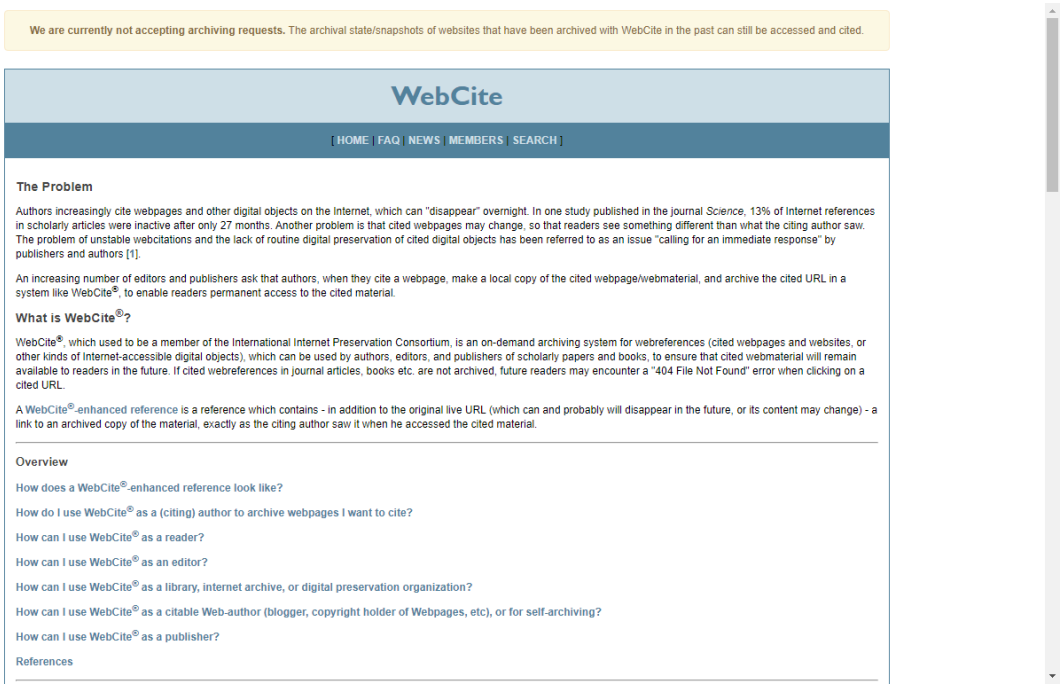
Grupa Web znanosti i digitalnih knjižnica (Web Science and Digital Libraries, WS-DL) sa sveučilišta Old Dominion (ODU, Virginia) proučavala je izazove s kojima su se suočavali znanstvenici pri stvaranju i dijeljenju vlastitih arhive weba. Projekt Archives Unleashed (AU) razvijen je kao suradnja između povjesničara, knjižničara i informatičara. Projekt razvija alate pretraživanja i analize podataka koji istraživačima mogu omogućiti pristup i dijeljenje podataka unutar arhiva weba. WS-DL grupa također razvija alate za korisnike: moguće je dozvoliti lokalno arhiviranje web stranice za vrijeme pregledavanja weba i dostavljanje URL-ova javnim arhivima za pohranu. Jedan problem s dostavljanjem URL-ova za arhiviranje (za razliku od stvaranja lokalnih arhiva) jest da ono što se prikazuje u web pregledniku vjerojatno neće biti isto ono što će arhiv zabilježiti. Kad se URL dostavi, pobirač će prema svojim parametrima prikupiti sadržaje te stranice iz svoje perspektive, bez

geolokacije ili kolačića (engl. *cookie*) pojedinca koji je predao URL. Još jedan problem predstavlja činjenica da neki web pobirači (kao što je Heritrix kojega koristi IA) ne provode JavaScript pri posjećivanju web stranica i zbog toga im mogu promaknuti arhivski izvori koji se učitavaju samo kad preglednik provede JavaScript.

WS-DL grupa stvorila je WARCreate ekstenziju i WAIL (Web Archiving Integration Layer) aplikaciju za arhiviranje weba. WARCreate ekstenzija za Google Chrome omogućava stvaranje lokalnog arhiva web stranice koja se trenutno prikazuje u pregledniku. To može biti stranica koja je učitana nakon interakcije (npr. klizanje/*skrolanje* koje uzrokuje učitavanje dodatnog sadržaja) ili stranica koja se prikazuje samo nakon provjere identiteta korisnika (npr. korisnički računi društvenih mreža). WARCreate stvara datoteke u standardnom WARC formatu koji se sprema na osobno računalo korisnika. WS-DL grupa također je razvila WAIL (Web Archiving Integration Layer) aplikaciju koja dozvoljava korisnicima da ponovno reproduciraju lokalne arhive (WARC datoteke) te da samostalno provode pobiranje weba. Umjesto arhiviranja samo jedne stranice kao što je slučaj sa WARCreate-om, WAIL može stvoriti arhiv web stranice sa svim njezinim poveznicama ili čak cijelog web mjesta. Najnovija verzija aplikacije koristi softver baziran na Pythonu i pobirače bazirane na preglednicima, koji provode JavaScript prije stvaranja arhiva.

4.5. WebCite

WebCite (slika 8) je sustav arhiviranja web referenci na zahtjev tj. služi za citiranje web stranica, web mjesta i drugih oblika internetskih objekata. Osnovan je 2005. u Torontu (Kanada) iako se svijest o potrebi takve vrste usluge prvi put pojavila već u 1997. Autori, urednici i izdavači znanstvenih publikacija mogu koristiti WebCite kako bi se osiguralo da će citirani web materijal u budućnosti ostati dostupan čitateljima. Korištenje ove usluge osigurava da je zabilješka citiranog rada prisutna unutar WebCitea. Ako autor želi da njegovi čitatelji citiraju određenu verziju njegove publikacije, može samostalno arhivirati svoj rad unutar WebCitea. Time se osigurava da službeno neobjavljeni materijali ostanu dostupni za citiranje. Autori trebaju objaviti te materijale na webu i potom ih samostalno arhivirati. Usluga je besplatna i ne zahtijeva korisničku registraciju. WebCite je bivši član IIPC-a.



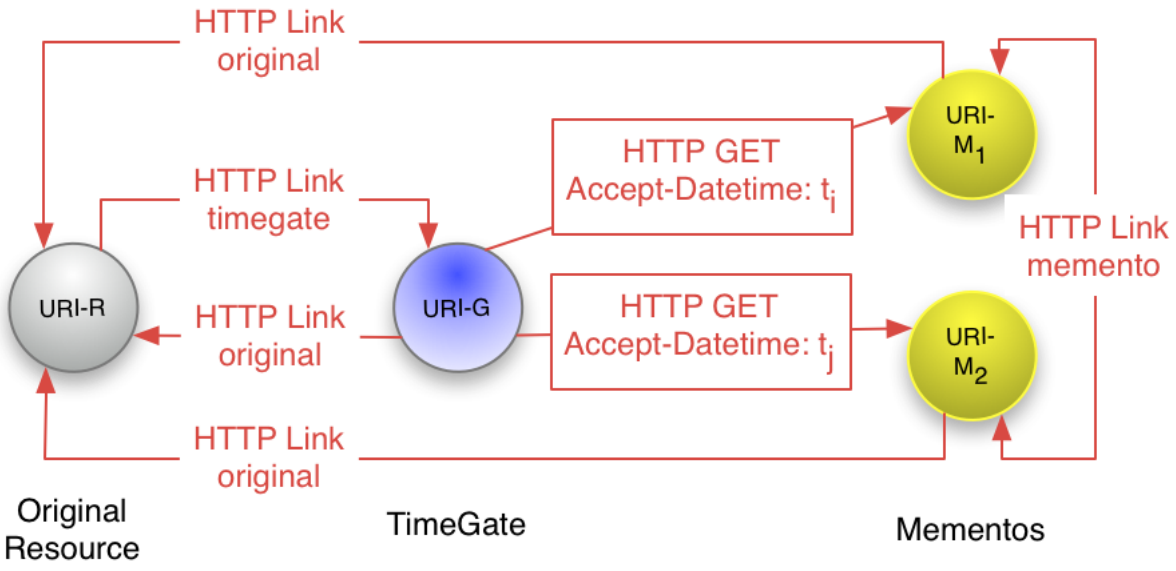
Slika 8. Web mjesto WebCite-a. Izvor: WebCite.

4.6. Projekt Memento

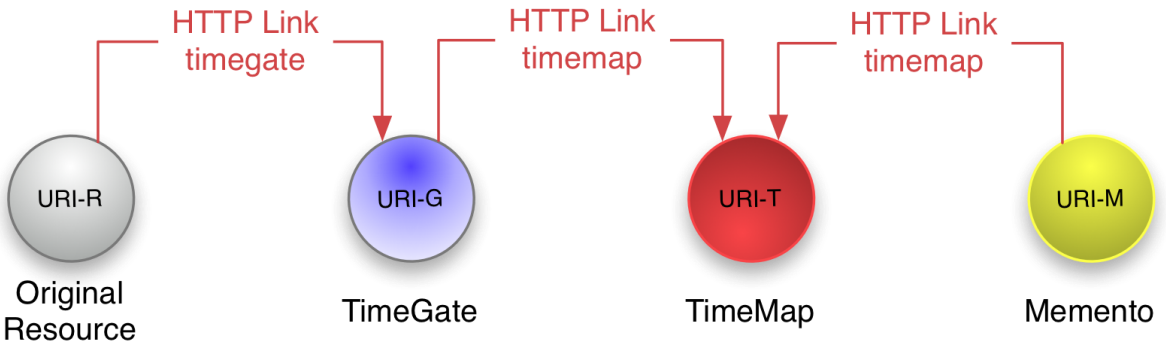
Projekt Memento osnovan je 2009. godine kao suradnja između istraživačkog laboratorija u Los Alamosu (New Mexico) i odjela informatike sveučilišta Old Dominion (Virginia). Cilj projekta jest učiniti dohvaćanje zastarjelih web resursa jednako lakim kao i dohvaćanje sadašnjeg weba. Na web mjestu Mementa (slika 9), kao i pomoću ekstenzije za preglednik Google Chrome može se pristupiti prošloj verziji web mjesta pod uvjetom da je ta verzija javno dostupna unutar nekog arhiva weba koji podržava Mementov protokol. Osnivači projekta ističu da većina javnih arhiva već podržava ovaj protokol i nadaju se da će ga jednog dana moći podržavati i preglednici (tada više ne bi bilo potrebno instalirati ekstenziju). Mementov protokol HTTP-u dodaje vremensku dimenziju, odnosno dohvaća sadržaje prema željenom datumu i vremenu.

Slika 9. Web mjesto Menta. Izvor: Mementoweb.

Mentov okvir sastavljen je od dvije komponente. Prvu čine tri tipa resursa: originalni resurs, Memento i TimeGate. Originalni resurs je web resurs koji već postoji ili je nekoć postojao na webu čiju prijašnju verziju korisnik želi pronaći. Memento je prethodna verzija originalnog web resursa koja prikazuje kako je on izgledao u određenom trenutku u prošlosti. TimeGate je web resurs koji na temelju traženog datuma i vremena dohvaća Memento koji se najbliže podudara s datumom i vremenom koji je unio korisnik. Sva tri tipa resursa mogu postojati na istom (npr. kod sustava upravljanja sadržajem) ili različitom (npr. kod arhiva weba) poslužitelju. Jedan Memento može ukazivati na druge Mente, a TimeGate i Memento mogu ukazivati na prvi i posljednji Memento za originalni resurs (slika 10). Drugu komponentu Mentovog okvira čine TimeMaps: resursi koji pružaju popise URL-ova. TimeMap je strojno čitljiv dokument koji u popisu nabraja originalni resurs, njegov TimeGate, njegove Mente i pripadajuće metapodatke kao što su datum i vrijeme arhiviranja svakog Menta (slika 11).



Slika 10. Prikaz Mementovog procesa pristupanju bivšim verzijama resursa. Izvor: Mementoweb.



Slika 11. Prikaz otkrivanja Mementa pomoću TimeMaps. Izvor: Mementoweb.

4.6.1. Mink

Mink je ekstenzija za Chrome koja koristi Mementov protokol za informiranje korisnika o postojanju arhivirane verzije web stranice koja se trenutno pregledava te pružanje pristupa toj verziji. Ako ne postoji nijedna arhivirana verzija, ekstenzija korisniku pruža opciju za pohranu te stranice unutar različitih arhiva weba i pregled stranice nakon što se ručno arhivira. Svaki put kad korisnik posjeti novu stranicu, Mink šalje zahtjev Mementu i prikazuje popis arhiviranih verzija tog određenog URL-a.

5. EVOLUCIJA ARHIVIRANJA WEBA

Costa, Gomes i Silva (2016.) proveli su dva istraživanja o različitim inicijativama arhiviranja weba diljem svijeta kako bi pratili njihov razvoj. Najprije su kontaktirali 33 različite inicijative s pitanjima o količini i formatu arhiviranih podataka, broju osoblja i sl. Zatim su analizirali njihove odgovore i 2011. objavili rezultate na Wikipediji kao popis inicijativa arhiviranja weba[2] kako bi se podaci mogli zajednički ažurirati. Godine 2014. opet je provedena ista analiza kako bi se najnovija ažuriranja usporedila s rezultatima prvog istraživanja. Između 2010. i 2014. godine zabilježen je golem porast broja samih inicijativa, količine arhiviranih podataka te funkcionalnosti i alata uvedenih u arhive weba. U četiri godine broj inicijativa arhiviranja weba porastao je s 42 na 68. Konkretnije, u 2010. su analizirane 42 inicijative u 26 zemalja: 23 u Europi, 10 u Sjevernoj Americi, 6 u Aziji i 3 u Oceaniji. Godine 2014. se radilo o 68 inicijativa u 33 zemlje: 38 u Europi, 22 u Sjevernoj Americi, 8 u Aziji, 3 u Oceaniji i 1 u Africi. Autori ističu da su za ovaj porast velikim dijelom zaslužne akademske institucije jer su razna sveučilišta započela vlastite programe arhiviranja weba.

Korištenje vanjskih usluga (trećih stranki) za arhiviranje web sadržaja poraslo je sa 16 na 19%. Autori ističu da ova činjenica može biti razlog tome što mnoge inicijative nisu mogle precizno odgovoriti na pitanja o broju osoblja koje se bavi tim zadacima (tj. ustanove ne upravljaju samostalno svojim inicijativama arhiviranja weba već to prepuštaju svojim partnerima u čiji rad nemaju potpun uvid). Prema autorima, još jedan razlog nemogućnosti odgovora na pitanje o osoblju jest to što su timovi koji rade na arhiviranju weba često mali i podložni variranju. Iako se broj samih inicijativa drastično povećao, broj članova osoblja povećao se tek neznatno, a timovi su se dodatno smanjili. Godine 2010. je 112 osoba radilo u arhivima weba na puno radno vrijeme, a njih 166 na pola radnog vremena. Prosjek je otprilike 3,5 osoba na punom radnom vremenu i dvoje na pola. Prema podacima iz 2014., u inicijativama je 108 osoba bilo zaposleno na puno radno vrijeme (prosječno dvoje u svakom arhivu), a 197 na pola (također dvoje u prosjeku). Osoblje se uglavnom sastojalo od knjižničara i informatičara čiji su zadaci prikupljanje podataka i kontrola kvalitete arhiviranja. Što se tiče opsega, u 2010. je 50% arhiva čuvalo zbirke manje od 10 TB, 31% između 10 i 100 TB, a 19% preko 100 TB. U 2014. je 42% arhiva čuvalo manje od 10 TB, 42% između 10 i 100 TB, a postotak inicijativa sa preko 100 TB sadržaja ostao je na 19. Prema broju datoteka,

[2] List of Web Archiving Initiatives, Wikipedia. URL: https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives (13.5.2020.)

postotak inicijativa sa zbirkama između 100 milijuna i milijardu sadržaja smanjio se sa 43 na 33% budući da se postotak inicijativa sa preko milijardu sadržaja povećao sa 22 na 33%. Godine 2010. je bilo očuvano gotovo 182 milijarde sadržaja (6,6 PB), od čega se 150 milijardi (5,5. PB) nalazilo unutar IA. Godine 2014. je ukupno bilo očuvano oko 535 milijardi sadržaja (17 PB). Rezultati prikazuju porast od 294% u sadržaju i 258% u volumenu podataka u samo četiri godine. Podaci iz 2010. ukazuju da je 38% inicijativa postavilo nekakav oblik ograničenja pristupa svojim zbirkama, ali podaci iz 2014. nisu bili dovoljni za usporedbu.

ARC i WARC su očekivano najpopularniji formati za pohranu arhiviranih podataka. U 2010. godini 26% inicijativa koristilo je isključivo ARC format, 10% isključivo WARC, a 54% oba formata. U 2014. je 13% inicijativa koristilo isključivo ARC, 13% isključivo WARC, a 47% oba. Postotak inicijativa koje koriste samo ARC se prepolovio i autori nagađaju da je razlog tome masovno prebacivanje na novi WARC standard. U 2010., 62% inicijativa koristilo je alate Heritrix (za pobiranje), NutchWax (za tekstualno pretraživanje) i WaybackMachine (za pretraživanje URL-ova). Postotak korištenja ovih alata smanjio se na 57% u 2014. Autori smatraju da je razlog tome povećanje korištenja različitih alata koje razvijaju usluge arhiviranja weba trećih stranaka. Što se tiče metoda za pretraživanje, u 2010. je 89% inicijativa podržavalo pretraživanje arhiviranog sadržaja preko URL-a, 79% putem metapodataka i 67% potpuno tekstualno pretraživanje. Autori ističu pretraživanje po tekstu i URL-u kao metode koje korisnici najradije koriste. Rezultati iz 2014. su gotovo isti: glavne metode za otkrivanje arhiviranog sadržaja još uvijek su pretraživanje prema URL-u, metapodacima i punom tekstu. Međutim, ispitanici su napomenuli da postojeće tehnologije pružaju nezadovoljavajuće rezultate pretraživanja te da je tekstualno pretraživanje (metoda koju korisnici preferiraju) teško implementirati u arhive weba. Autori upozoravaju na presudnu važnost razvitka učinkovite tehnologije pretraživanja za pristup golemim količinama podataka koji su već pohranjeni unutar arhiva weba.

6. VAŽNOST ARHIVIRANJA WEBA

Službena stranica IIPC-a navodi razloge za arhiviranje i očuvanje weba. Ističe da današnji svijet (u kojemu se informacije zahvaljujući internetu dijele brže i više nego ikada prije) stvara brojne nove izazove za ustanove kojima je zadaća dokumentacija i očuvanje modernoga znanja i kulture. Određeni sadržaji koje su arhivske ustanove dužne prikupiti trajno su se „preselili“ na internet, uključujući i brojne znanstvene publikacije, vijesti, materijale političkih kampanja itd. Dodatnu problematiku predstavlja dinamična priroda web stranica, odnosno njihova podložnost vrlo čestim promjenama: potrebno je zabilježiti sadržaj u stvarnom vremenu, uključujući i sve njegove prethodne i buduće verzije.

6.1. Arhivi weba kao dokazi na sudu

Članak Kierena McCarthyja iz 2018. opisuje slučaj u američkom zakonodavstvu u kojem su se materijali prikupljeni unutar IA i WM koristili kao legitimni dokazi u sudskom procesu. Riječ je o slučaju hakera Fabia Gasperinija koji je 2014. razvio virus kojim je zarazio preko 150 tisuća računala (uglavnom unutar SAD-a). Virus bi ušao kroz rupu u softveru, ukrao korisnička imena i lozinke te počeo posjećivati online reklame i tražiti nova računala za širenje napada. Tužitelji su pronašli testnu kopiju virusa na Gasperinijevoj osobnoj adresi elektroničke pošte te povezali poslužitelje na kojima je Gasperini pokrenuo *malware*. Nakon njegova uhićenja, Gasperinijevi korisnički računi na Googleu i Facebooku su izbrisani, kao i podaci s njegovih čvrstih diskova kod kuće. Tužiteljstvo je kroz uslugu WM uspjelo naći arhivirane verzije web mjesta na kojima su bile udomljene reklame koje je posjećivao njegov *malware*. Otkriveno je da su ta web mjesta također bila registrirana pod Gasperinijevim imenom. Gasperini je uložio žalbu na odluku suda (odslužio je godinu dana zatvora) tvrdeći da se arhivirane verzije web stranica nisu mogle koristiti kao legitiman dokaz protiv njega. Član osoblja IA svjedočio je u slučaju, objasnio kako funkcionira arhiviranje weba i potvrdio da se verzije koje je prikupilo tužiteljstvo poklapaju s verzijama unutar samog IA. Gasperinijeva je žalba odbijena i time je postavljen presedan za buduće pravne procese u kojima bi se arhivi weba mogli koristiti kao izvori dokaza. McCarthy komentira kako bi se mogle pojaviti situacije u kojima svjedočanstva osoblja arhiva weba više neće biti potrebna jer će se materijali sami po sebi smatrati važećima i legitimnima. Također valja spomenuti da službena stranica IA navodi da WM nije stvoren za pravnu uporabu. Postoji segment s odgovorima na česta pitanja odvjetnika i smjernice za odgovaranje na pravne zahtjeve, ali

organizacija ističe da nema vlastiti odvjetnički tim te da ih pravni zahtjevi odvraćaju od redovitih dužnosti.

7. IZAZOVI PRI ARHIVIRANJU WEBA

Arhiviranje weba suočava se s jedinstvenim izazovima i poteškoćama zbog kompleksnosti same djelatnosti, veličine weba, različitosti formata datoteka te potencijalnih pravnih i društvenih problema. Arhivi weba trebaju očuvati autentičnost i integritet arhiviranog sadržaja, ali uvjeti za to variraju sa svrhom zbirke. (Niu, 2012.) U nekim je slučajevima dovoljno očuvati samo intelektualni sadržaj (primjerice samo HTML stranice), a u drugima je potrebno očuvati cjelokupnu strukturu i kontekst web izvora.

7.1. Tehnički izazovi

7.1.1. Vremenska dosljednost

Ball (2010.) opisuje vremensku dosljednost (koherenciju, koheziju) kao svojstvo skupa web stranica koje opisuje trenutak u kojemu su sve te stranice istovremeno postojale na webu. Ističe da je vremensku dosljednost relativno lako postići za manji skup stranica koje se rijetko ažuriraju, ali postaje sve teže s većim brojem stranica i ažuriranja. Neke od strategija postizanja vremenske dosljednosti mogu biti pobiranje web mjesta prema učestalosti njihova ažuriranja (na taj način stranice s najkraćim životnim vijekom imaju najkraći razmak između pobiranja, a one s najdužim najduži), korištenje uzorkovanja (npr. korištenje istog redoslijeda pobiranja za nekoliko stranica koje imaju približno jednak životni vijek i učestalost ažuriranja) ili pokretanje više pobirača istovremeno (iako time može doći do preopterećenja poslužitelja).

7.1.2. Propadanje poveznica

Digitalni se mediji suočavaju sa zastarijevanjem hardvera i softvera te kvarenjem podataka, a arhivirani web sadržaji podložni su propadanju poveznica. Prema definiciji Techopedije, propadanje poveznica označava neispravne hipertekstualne poveznice do kojih dolazi kad se web stranica pomakne na novu adresu ili domenu, ukloni ili reorganizira. Klik na propalu poveznicu uglavnom rezultira prikazivanjem prazne stranice (greška 404) s porukom da se željena stranica ne može pronaći. Neke se poveznice mogu „popraviti“ preusmjeravanjem na ispravne stranice, ali neka su propadanja neizbježna zbog uklanjanja čitavih web mjesta na koje poveznice ukazuju. Blog objava organizacije Decentralised Public Library (DPL) u sklopu projekta Arweave iz 2019. ističe da mnogi

korisnici imaju pogrešnu predodžbu o webu kao vječnom i beskonačnom, tj. vjeruje se da sadržaji koji se jednom postave na web nikad ne mogu nestati. Prema DPL-u, web je zapravo „šokantno krhak“: u vremenskom roku od 20 godina propada u prosjeku 98,4% aktivnih poveznica, čime dijelovi weba postaju potpuno nedostupni budućim generacijama. Kako se sve više akademskih, pravnih i poslovnih procesa odvija isključivo putem weba, propadanje poveznica može uvelike ugroziti te procese. DPL ističe da su uzroci propadanja poveznica uglavnom „banalni“, primjerice vlasnik web mjesta može prekinuti svoju registraciju na domeni ili ostati bez vremenskih i financijskih resursa za održavanje web mjesta. Međutim, razlozi mogu biti i politički (npr. cenzura) ili poslovni (npr. zatvaranje velikih web mjesta koja više nisu dovoljno profitabilna). Osim korištenja usluga IA i WM za pohranu i pristup zastarjelim poveznicama, postoje i posebne ekstenzije koje mogu pohraniti poveznice na *permaweb* koji je razvila organizacija Arweave. Permaweb se od tradicionalnog weba razlikuje po tome što se sadržaji unutar njega ne mogu izgubiti, mijenjati niti namjerno uklanjati.

Arhivi weba više su se puta dokazali kao „spasitelji“ u situacijama propalih poveznica. IA je razvio dvije ekstenzije za web preglednike Google Chrome (The Wayback Machine for Chrome, razvijena 2017.) i Mozilla Firefox (No More 404s, razvijena 2016.) koje će u slučaju klika na neispravnu poveznicu obavijestiti korisnika o postojanju arhivirane verzije željenog URL-a (ukoliko ta verzija postoji). Dans (2019.) opisuje svoje iskustvo korištenja IA i WM za pohranu poveznica s popisa literature iz svoje objavljene knjige: želio je osigurati da će sve poveznice biti funkcionalne njegovim čitateljima u bilo koje doba u budućnosti. Ističe IA i WM kao izrazito važne alate za svakoga tko želi „spasiti“ bilo kakve poveznice i komentira kako bi praksa pohranjivanja poveznica unutar IA i njihovo pregledavanje putem WM mogla postati norma za svaku publikaciju koja citira vanjske izvore.

7.1.3. Funkcionalnost

Prema Niu (2012.), arhivi i knjižnice trebaju početi obraćati više pažnje na iskoristivost i funkcionalnost arhiva weba kako se sve više upoznaju s tim načinom rada (ističe da je mogućnost budućeg korištenja arhiviranih izvora krajnja svrha njihovog arhiviranja). Autorica je sastavila popis funkcija i prema tim stavkama vrednovala funkcionalnosti deset arhiva weba koji su članovi IIPC-a, uključujući WM, UKGWA, LC itd. (više o tim arhivima u idućem poglavlju). Otkrila je da arhivi redovito nude osnovne funkcije

poput pretraživanja po URL-u i/ili ključnim riječima te sužavanja rezultata prema datumu, domeni i tipu medija. Međutim, niti jedan arhiv iz istraživanja nije podržavao složenije funkcije poput rudarenja podataka, personaliziranih usluga za korisnike ili rekonstrukcije izgubljenih web stranica. Neki arhivi iz istraživanja imali su probleme s funkcijama kao što su skrivene ili „nevidljive“ poveznice za pomoć, alati naprednog pretraživanja te segmentirano pretraživanje. Niu ističe da pojavljivanje ovih problema ne znači da arhivima nije stalo do poboljšanja funkcionalnosti svojih sučelja niti da su nesposobni sami izaći na kraj s takvim poteškoćama, već da jednostavno nemaju vremena za razvoj naprednijih funkcija radi prevelikog opterećenja izgradnjom i održavanjem svojih zbirki. Nekoliko mjeseci nakon provedbe istraživanja, Niu je u nekim arhivima primijetila promjene na bolje, zbog čega vjeruje da arhivi weba s vremenom mogu postati funkcionalniji. Također naglašava da je moguće postojanje nekih mnogo naprednijih arhiva koji nisu bili uključeni u istraživanje te da su neke napredne funkcije možda otpočetka bile dostupne na web mjestima dotičnih arhiva weba, ali nisu bile dovoljno vidljive.

7.1.4. Skriveni web

ISO/TR 14873:2013 definira duboki/skriveni/nevidljivi web kao dio weba koji se ne može pobirati niti indeksirati kroz pretraživače i sastoji se od resursa koji su dinamično generirani ili zaštićeni lozinkama. Prema Masanèsu (2006.), glavna metoda prikupljanja web sadržaja je praćenje poveznica s drugih stranica, što znači da na svaki dokument mora voditi barem jedna poveznica kako bi ga pobirači mogli pronaći. Iz ovog se razloga velik dio weba ne može arhivirati automatskim alatima. Autor definira strukturalni skriveni web kao dio weba koji zahtijeva složenu korisničku interakciju (tj. više od jednog klika) koju je pobiračima teško imitirati. Ova vrsta skrivenog weba također sadrži veliku količinu sadržaja jer se koristi za objavljivanje strukturiranih (baze podataka) i nestrukturiranih (zbirke) repozitorija dokumenata. Masanès opisuje različite načine moguće suradnje arhiva i vlasnika web mjesta. Vlasnici mogu stvoriti popis svih skrivenih sadržaja unutar svojeg web mjesta (najjednostavnija metoda) kako bi se osiguralo da će pobirači sve posjetiti ili razviti komunikacijske protokole. Implementacija protokola je korak dalje koji omogućava izravnu komunikaciju između pobirača i poslužitelja kako bi se dobio popis dokumenata s pridružujućim metapodacima. Autor napominje da se ove metode ipak ne mogu smatrati zadovoljavajućima te da će biti potrebna dodatna istraživanja za kvalitetno i jednostavno očuvanje skrivenog weba. Masanès također ističe da se skriveni web ne bi trebao zanemariti

pri arhiviranju zbog njegove veličine i opsega sadržaja koji će potencijalno biti od interesa mnogim baštinskim ustanovama.

7.1.5. Premještanje arhiva weba

Premještanje zbirki arhiva weba koje mogu sadržavati na milijune ARC/WARC datoteka (od kojih se svaka sastoji od 100 MB zbirnih resursa) može se doimati kao vrlo zahtjevan posao. Newing (2017.) opisuje proces prebacivanja arhiviranih podataka unutar arhiva vlade UK (UKGWA). Podaci su se najprije trebali prebaciti iz podatkovnog centra Internet Memory Research (IMR) u sjedište arhiva u Kewu (London). Najprije se pokušalo kodirati datoteke (ARC format) i poslati ih putem interneta, ali taj se proces pokazao kao nepouzdan, spor i sklon greškama zbog kojih je bilo potrebno više puta počinjati ispočetka. Odlučeno je da bi sljedeći najbolji potez bilo prebacivanje podataka na fizičkom mediju (USB mediji kapaciteta 2 TB) koristeći kurira. Ovom je metodom između 2015. i 2017. uspješno prebačeno otprilike 120 TB podataka na 70 USB diskova. Međutim, 2016. je došlo do izmjene ugovora s partnerskom institucijom MirrorWeb i svi su se podaci iz Kewa morali prebaciti na pohranu u računalni oblak (tvrtka Amazon). Korištena su dva primjerka posebnog računala zvanog Amazon Snowball i po osam USB diskova za svako računalo koji bi ih „hranili“ podacima. Nakon manjih tehnoloških poteškoća riješenih prebacivanjem Snowballa na drugačiju verziju operativnog sustava Linux, UKGWA se uspješno prebacio na pohranu u oblaku i proces koji je prije trajao dvije godine dovršen je u dva tjedna. Ako arhivi weba nemaju partnerstva s institucijama koja im mogu uvelike olakšati ovaj proces posuđivanjem posebnog hardvera, premještanje sadržaja može biti veoma dugotrajan i rizičan proces.

7.2. Društveni izazovi

Društveni izazovi za arhiviranje weba mogu uključivati probleme oko nadležnosti, odabira strategije, dinamike, sadržaja trećih strana, strategija transfera i izlaza, pravnih uvjeta, zloporabe, upravljanja, arhiviranja obimnog sadržaja i sl. (Ball, 2010.)

7.2.1. Pravni problemi

Prema podacima sa službenog web mjesta IIPC-a, nijedan se član Konzorcija ne suočava s jednakim pravnim izazovima i situacijama pri arhiviranju weba budući da svaki

djeluje unutar različitih pravnih okvira. U nekim zemljama još nisu doneseni zakoni koji bi regulirali arhiviranje weba. Ako zakoni ne postoje ili su pravni okviri arhiviranja weba nejasni, arhivi moraju zatražiti dopuštenje vlasnika web mjesta za pobiranje njihovih sadržaja. IIPC naglašava pristup sadržaju kao još jedan izazov: u nekim je zemljama korištenje arhiva weba dostupno samo unutar prostora knjižnice u kojoj arhiv djeluje zbog zakonskih ograničenja ili potencijalnog kršenja privatnosti. Za širi bi se pristup potencijalno mogle uključiti druge knjižnice ili partnerske ustanove. Također, neki zakoni koji propisuju postupanje s digitalnim sadržajima ne uključuju i web sadržaje u svoje propise. Internetske domene mogu predstavljati još jedan potencijalni problem zbog svojeg opsega, teritorija i pravnih ograničenja. Primjerice, ako je zakonski propisano da određeni arhiv smije prikupljati samo sadržaje unutar glavne državne domene, mogu se propustiti sadržaji koje su stvorili građani te države, a koji se nalaze na drugim općenitijim domenama kao što su .com, .org, .edu itd. IIPC navodi primjere arhiva weba Francuske i Danske koji prikupljaju sadržaje unutar domena .fr i .dk, ali i sve ostale sadržaje koji se odnose na njihove državljane (tj. čiji su stvaratelji državljani Francuske ili Danske ili se izdavači fizički nalaze na prostoru tih zemalja).

Zakon u nekim slučajevima dozvoljava arhivima i knjižnicama da od vlasnika web mjesta traže lozinke i ostale tehničke podatke kako bi prikupili materijale koji se ne mogu prikupiti automatskim pobiranjem (npr. sadržaji koji zahtijevaju pretplatu). Unutar zemalja koje nemaju definirane pravne okvire arhiviranja weba, arhivi su primorani osobno tražiti dopuštenje vlasnika web mjesta za prikupljanjem njihovih sadržaja. Podaci IIPC-a navode da arhivi na takve upite u 70 do 50% slučajeva ne dobivaju nikakav odgovor, čak ni odbijanje. Osoblje arhiva weba može biti preopterećeno jer je ponekad potreban izuzetan napor za kontaktiranje vlasnika mjesta i dobivanje dopuštenja za arhiviranje, što ih odvraća od njihovih redovitih dužnosti. Rezultati ovakvih situacija mogu biti neuravnotežene zbirke pune praznina.

7.2.2. Uklanjanje sadržaja

Korisnici i vlasnici web mjesta iz različitih razloga mogu od arhiva weba zatražiti uklanjanje određenih sadržaja iz zbirke. IA u takvim situacijama djeluje prema smjernicama o upravljanju zahtjevima o uklanjanju i očuvanju arhivskog integriteta koje su razvijene 2002. na sveučilištu u Berkeleyju (Kalifornija). Postoji više kategorija zahtjeva za uklanjanjem te reakcija koje bi arhivi weba trebali poduzeti kao odgovor na njih. U slučaju da vlasnici privatnih (tj. ne vladinih) web mjesta zatraže uklanjanje sadržaja zbog problema privatnosti,

potencijalne klevete ili sramoćenja, arhivisti bi trebali omogućiti korisnicima da samostalno uklone svoje materijale prema uputama robots.txt standarda. Ovime se osigurava da pobirači više neće prikupljati taj materijal niti ga činiti dostupnim unutar arhiva. Zahtjevi ove prirode ne bi se trebali javno objavljivati, ali arhivisti bi trebali zadržati njihove kopije. Ako treće strane zahtijevaju uklanjanje sadržaja zbog povrede autorskih prava ili intelektualnog vlasništva, arhivi bi najprije trebali provjeriti je li originalno sporno web mjesto uklonjeno i preusmjeriti žalbu na vlasnika tog mjesta ako je još aktualno. Ako je mjesto uklonjeno, tada bi arhivi trebali provjeriti valjanost žalbe i ukloniti sporne sadržaje ako se ispostavi da krše nečija prava. Arhivi su također dužni javno objaviti zahtjeve o uklanjanju sadržaja na temelju autorskih prava ili intelektualnog vlasništva, obavijestiti svoje korisnike o uklonjenom sadržaju i kontaktirati vlasnika spornog web mjesta. Zahtjevi uklanjanja zbog kontroverznih političkih, religioznih i sličnih razloga uglavnom se ne bi trebali uvažavati zbog očuvanja točne i potpune slike stanja društva u tom trenutku. Žalbe trećih stranaka zbog privatnosti osobnih podataka tretiraju se kao slučajevi u kojima se žale originalni autori ili izdavači spornih podataka. Zahtjevi od strane vladajućih tijela uvijek se moraju uvažiti u skladu sa sudskim odredbama, a ostale vrste zahtjeva (npr. ispravke grešaka, promjena vlasništva nad web mjestom) trebaju se rješavati na osnovi svakog slučaja posebno.

Opća uredba o zaštiti podataka (General Data Protection Regulation, GDPR) definira pravo na zaborav koje daje pojedincima pravo da zatraže brisanje svojih osobnih podataka od organizacija koje te podatke pohranjuju ili obrađuju. Ističe se da organizacije nisu obvezne uvijek se odazvati na takve zahtjeve zbog brojnih kompliciranih svojstava podataka na webu. Pružatelj zahtjeva najprije treba ispuniti jedan od uvjeta za opravdano brisanje podataka i potrebno je potvrditi njegov identitet, nakon čega se zahtjev uvažava i podaci se brišu unutar roka od mjesec dana. Pravo na zaborav primjenjuje se u situacijama u kojima osobni podaci više nisu nužni za svrhu za koju su prvobitno prikupljeni ili obrađeni, pojedinac povuče svoj pristanak na dijeljenje svojih osobnih podataka, organizacija je nezakonito prikupila ili obrađivala podatke, brisanje je potrebno iz pravnih razloga i sl. Organizacije mogu odbiti zahtjev iz različitih razloga koji mogu uključivati korištenje podataka u svrhu ostvarivanja prava na slobodu izražavanja i informacija ili u skladu s pravnim obvezama i odlukama, javni interes za podatke, njihova vrijednost za medicinska/povijesna/statistička istraživanja, pravne procese i sl. Svaka je situacija jedinstvena i svaki se zahtjev mora pojedinačno razmotriti. Pronalaženje svih mjesta na kojima su sporni podaci pohranjeni i procesirani te njihovo

uklanjanje zahtjevno je i vremenski i tehnički, zbog čega ovo pravo može biti veliko opterećenje organizacijama.

7.3. Savjeti za vlasnike web mjesta

Taylor (2012.) vlasnicima web mjesta nudi osam savjeta za osiguravanje što lakšeg arhiviranja njihovih sadržaja. Prvo, potrebno je slijediti relevantne web standarde i smjernice. Na taj se način uvelike olakšava arhiviranje, ponovno reproduciranje i korištenje web mjesta. Također se smanjuje broj iznimaka kojima se sustav za ponovnu reprodukciju treba prilagoditi prilikom učitavanja arhivirane verzije web sadržaja. Nadalje, potrebna je pažljivost pri postavljanju ograničenja unutar robots.txt datoteka budući da one mogu spriječiti pobirače u bilježenju elemenata koji su od presudne važnosti za vjerno očuvanje originalne verzije web mjesta. Taylor navodi primjer zanemarivanja CSS ili JavaScript karakteristika koje bi učinilo veliku razliku između originalne i arhivirane verzije sadržaja. Također bi trebala postojati detaljna i transparentna mapa web mjesta kroz koju se svaka pojedina stranica može lako posjetiti i zabilježiti. Pobirači se kreću webom isključivo pomoću praćenja poveznica, što znači da će im u potpunosti promaknuti ona mjesta ili stranice koja nisu povezana s njihovom početnom točkom ili ostatkom već pobranih stranica. Korisnik koji pregledava arhivirane web sadržaje također njima može navigirati jedino kroz praćenje poveznica jer određeni alati (npr. pretraživanje) više ne mogu funkcionirati u arhiviranoj verziji. Nadovezujući se na strukturu poveznica, Taylor ističe važnost svijesti o njihovu propadanju. Kad se URL promijeni bez preusmjeravanja na novu lokaciju, smanjuje se vjerojatnost arhiviranja novog URL-a i time je gotovo zagarantirano da će arhivirane verzije web mjesta prije i poslije promjene biti odvojene. Taylor predlaže vlasnicima web mjesta da promisle o postavljanju Creative Commons (CC) licence za svoje sadržaje. Time bi web mjesto moglo „samostalno“ pružiti arhivima i pobiračima dopuštenje za pobiranje i pohranu bez potrebe za traženjem dopuštenja od vlasnika. Taylor također ističe održive formate podataka, odnosno otvorene standarde i formate datoteka kao najbolji izbor za dugotrajno čuvanje. Vlasnicima preporuča korištenje HTML-a ili XML-a kako bi se jasno indiciralo kakvo se kodiranje treba koristiti za vjerno prikazivanje web stranice. Posljednji Taylorov savjet nalaže vlasnicima da koriste pružatelje platformi i sustave upravljanja sadržajem koji su prilagođeni arhiviranju weba. Potrebno je promotriti robots.txt datoteke platforme na kojoj su web mjesta udomljena ili se raspitati o njezinim pravilima oko pobirača prije obvezivanja na korištenje. Taylor upozorava da se predlošci web mjesta ili sustavi upravljanja sadržajem koje platforma koristi možda neće

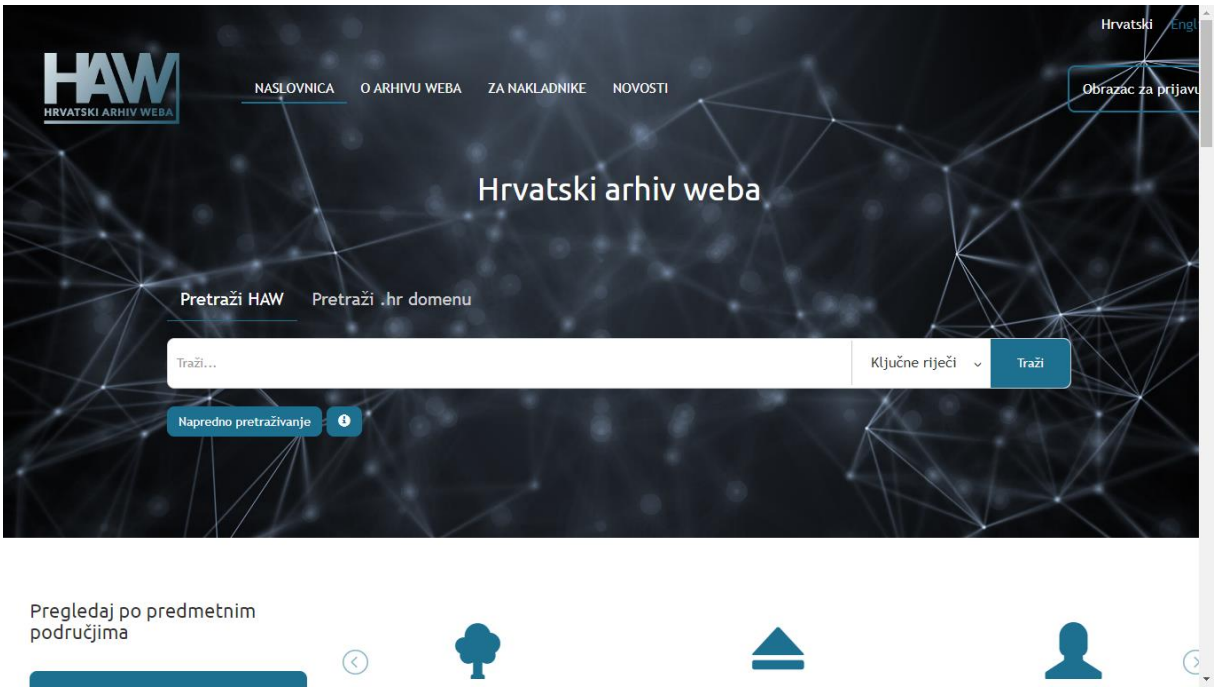
dobro arhivirati čak i ako platforma ne blokira pobirače. Preporuča vlasnicima da prouče kako se web mjesta koja koriste istu platformu prikazuju na WM, kao i konfiguraciju svih robots.txt datoteka na otvorenom sustavu upravljanja sadržajem. Taylor napominje da praćenje njegovih uputa ne garantira visokokvalitetno bilježenje web sadržaja niti njihovo besprijekorno očuvanje, ali i da će nepridržavanje tih uputa zasigurno dovesti do dodatnih poteškoća i izazova pri arhiviranju i očuvanju web sadržaja.

8. HRVATSKI ARHIV WEBA

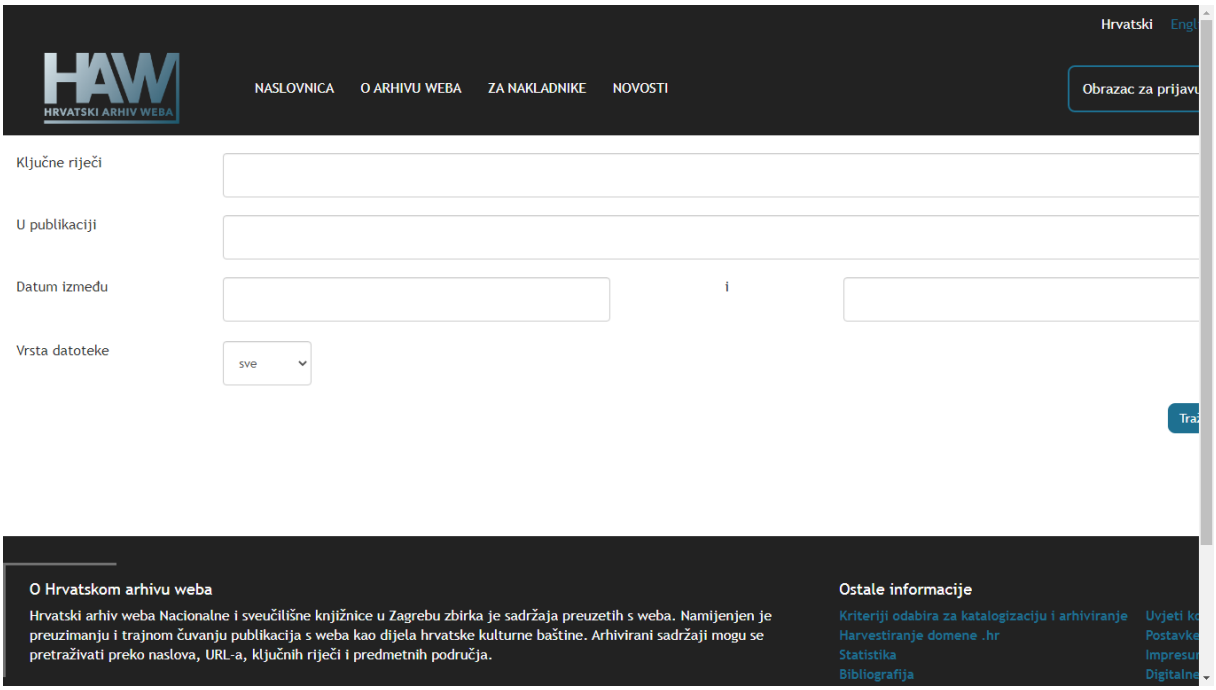
Nacionalna i sveučilišna knjižnica u Zagrebu (NSK) osnovala je Hrvatski arhiv weba (HAW) 2004. godine u sklopu zakona o obveznom primjerku (slika 12). U Hrvatskoj je 1997. u Zakon o knjižnicama dodan propis o uključivanju online publikacija u obuhvat obveznog primjerka. NSK 1998. počinje provoditi katalogizaciju online publikacija, a 2004. NSK i Sveučilišni računalni centar Sveučilišta u Zagrebu (Srce) zajednički stvaraju Digitalni arhiv hrvatskih mrežnih publikacija (DAMP). HAW je 2008. postao članom IIPC-a, a 2011. je započeo godišnja pobiranja .hr domene i izgradnju tematskih zbirki. Od 2012. metapodaci HAW-a postaju dostupni u Europeani i HAW je zasad jedini europski arhiv weba koji je to postigao. (Holub, K., osobna komunikacija, 23.6.2020.) Od 2019. novi zakon o knjižnicama i knjižničnoj djelatnosti propisuje da su web stranice uključene u obuhvat obveznog primjerka online publikacija. Nakladnici su dužni knjižnici dostaviti obvezne primjerke svih svojih publikacija objavljenih na webu, a knjižnica mora prihvatiti i obraditi te sadržaje te omogućiti njihovu pohranu i pristup. Uz razvoj DAMP sustava, Srce je također prilagodilo pobirač Heritrix za godišnja i tematska pobiranja hrvatskog weba. HAW prikuplja web sadržaje različitih ustanova, udruga, događaja, projekata, upravnih tijela, blogova, časopisa i knjiga. Do 2018. je u HAW-u prikupljeno preko 50 TB sadržaja.

HAW provodi selektivno pobiranje javno dostupnih publikacija na webu, godišnje pobiranje .hr domene i tematska pobiranja sadržaja od nacionalnog značaja. Javno dostupni sadržaji na .hr domeni pobiru se jedanput godišnje. Tematska pobiranja provode se povremeno, ovisno o aktualnim temama ili događajima koji su od značajne važnosti npr. prirodne katastrofe, predsjednički i parlamentarni izbori itd. Selektivno arhiviranje provodi se od početka postojanja HAW-a (2004.) korištenjem DAMP-a, a godišnja i tematska pobiranja provode se od 2011. pomoću Heritrixa. Cilj selektivnog arhiviranja je ponuditi općenitu (ne samo povijesnu) sliku hrvatskog weba, zbog čega se odabiru kriteriji za prikupljanje kvalitetne i reprezentativne građe. Knjižnica i arhiv za vrijeme procesa selektivnog arhiviranja svakodnevno komuniciraju i međusobno razmjenjuju podatke. Pri selektivnom arhiviranju .hr domena ima prednost pred ostalima, ali prikupljaju se i relevantni sadržaji s generičkih domena (.com, .org itd.) Učestalost pobiranja određuje se prema procijenjenom značaju određenog web mjesta za hrvatsku zajednicu, njegovoj strukturi, tehničkim i sadržajnim promjenama te veličini (datoteke veće od 500 MB pobiru se rjeđe). HAW pri selektivnom pobiranju u potpunosti zanemaruje robots.txt datoteke, a pri domenskom i tematskom poštuje njihova ograničenja i apelira na vlasnike web mjesta da dopuste pristup pobiračima.

Pretraživanje selektivno arhiviranih sadržaja moguće je prema naslovu, URL-u i ključnim riječima, a dostupno je i napredno pretraživanje (slika 13) prema publikaciji, datumu i vrsti datoteke (HTML, PDF, .doc, .ppt, .xls, .txt, .rtf). Pretraživanje sadržaja iz pobiranja prema domeni provodi se kroz unos potpunog URL-a željenih web stranica, a tematske zbirke mogu se pregledavati prema naslovu.



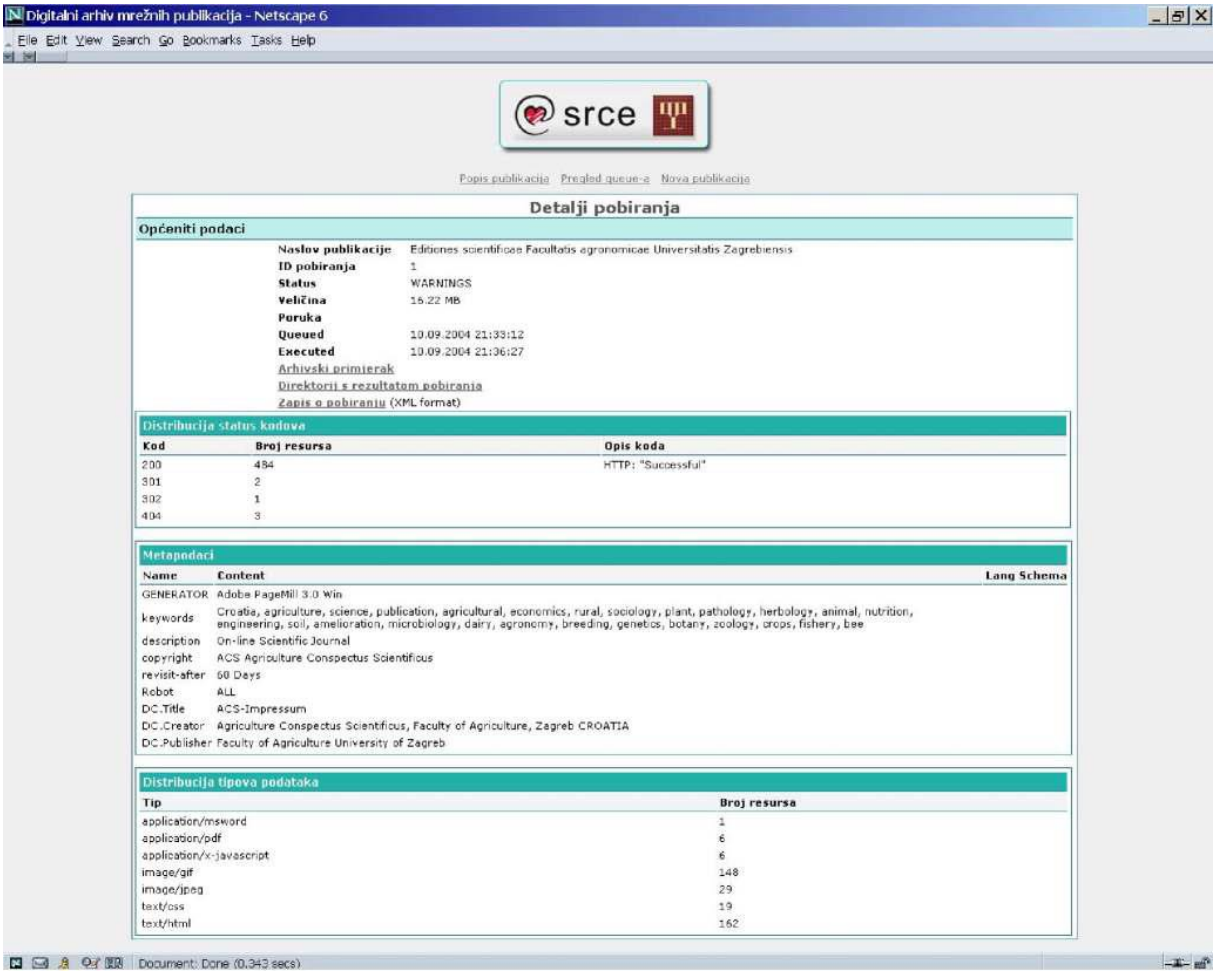
Slika 12. Web mjesto HAW-a. Izvor: Hrvatski arhiv weba.



Slika 13. Napredno pretraživanje HAW-a. Izvor: Hrvatski arhiv weba.

HAW za pobiranje koristi dva različita alata: DAMP (slika 14) i Heritrix. DAMP se koristi od 2004. za arhiviranje obveznog primjerka online publikacija, a Heritrix je stvoren u IA 2003. i koriste ga brojne arhivske organizacije diljem svijeta. Heritrix se unutar HAW-a koristi od 2011. za domensko i tematsko pobiranje. HAW od početka korištenja Heritrixa pohranjuje sadržaje u WARC formatu tj. ARC nikad nije korišten. Na početku se za prikaz pobranog sadržaja koristio WM, a od 2017. se koristi OpenWayback. U HAW-u nisu prikupljeni sadržaji stvoreni prije 2004. godine (tj. prije postojanja samog arhiva), a također se ne prikupljaju ni sadržaji elektroničke pošte, računalne igre, radne verzije publikacija, chat, reklame, sadržaji iza korisničke registracije te građa koju je zbog tehničkih poteškoća nemoguće arhivirati (u takvim se slučajevima autorima/nakladnicima preporuča promjena formata) ili koja već jest dio neke druge digitalne zbirke ili arhiva weba. Što se tiče zahtjeva za brisanjem sadržaja, dosad nije bilo slučajeva u kojima se neki sadržaj morao u potpunosti odstraniti. Ukoliko se arhivirani sadržaj iz nekog razloga mora ukloniti iz arhiva, pristup mu se ograničava na samo jedno računalo unutar NSK. Postoji i mogućnost da se pristup određenom sadržaju dozvoli samo administratorima. HAW je udomljen u Srcu na dvije lokacije i nova se pobiranja svaki dan sigurnosno kopiraju na njih. Starija pobiranja pohranjena su kao sigurnosne i produkcijske kopije u Srcu te kao arhivirani primjerci u sefovima na obje lokacije. U HAW-u su na puno radno vrijeme zaposleni jedan diplomirani

knjižničar i jedan viši knjižničar čiji su zadaci identifikacija, odabir, prihvat, katalogizacija i arhiviranje sadržaja te komunikacija s nakladnicima. U NSK je na dio radnog vremena zaposlen jedan knjižničarski savjetnik čiji su zadaci razvoj i koordinacija HAW-a. Planovi za budućnost HAW-a uključuju djelovanje unutar Hrvatske digitalne knjižnice koja bi trebala uključivati centralno pretraživanje sve digitalne građe (pa tako i arhivirane sadržaje iz HAW-a), uključivanje u Memento, rad na vizualizaciji pohranjenog sadržaja te početak pobiranja Twittera.



Slika 14. Sučelje DAMP-a. Izvor: Willer, M. (2017.) Arhiviranje hrvatskih mrežnih publikacija: od projekta do programa arhiviranja u Nacionalnoj i sveučilišnoj knjižnici u Zagrebu (NSK).

8.1. Usporedba DAMP-a i Heritrixa u HAW-u

Celjak i Milinović (2012.) opisuju dva alata otvorenoga programskog koda koja su na početku bila razmatrana za provođenje pobiranja u HAW-u: DAMP i Heritrix. Provedeno je probno pobiranje korištenjem oba pobirača, a zatim su se vizualno usporedile arhivirane verzije posjećenih web mjesta. Iako je utvrđena otprilike jednaka kvaliteta prikaza arhivskih kopija, otkrivene su i određene razlike između dva alata. Rezultati su pokazali da Heritrix može pohraniti arhivirane verzije u ARC/WARC formatu kao i u zrcalnom (engl. *mirror*) obliku, a DAMP samo u zrcalnom obliku. Zrcalni oblik pohranjivanja podrazumijeva velik broj malih datoteka (jedan web resurs zauzima jednu datoteku), a ARC/WARC manji broj velikih datoteka. Heritrix za ponovnu reprodukciju arhiviranih sadržaja treba dodatnu programsku podršku (npr. WM), što nije slučaj kod DAMP-a. Heritrix je fleksibilniji s većim brojem opcija za prilagođavanje dok je DAMP jednostavniji za korištenje. Heritrix je jedinstven prema tome što sprema zaglavlja HTTP upita i odgovora, a DAMP po svojoj skalabilnosti (posao se može lako raspodijeliti na dodatne poslužitelje).

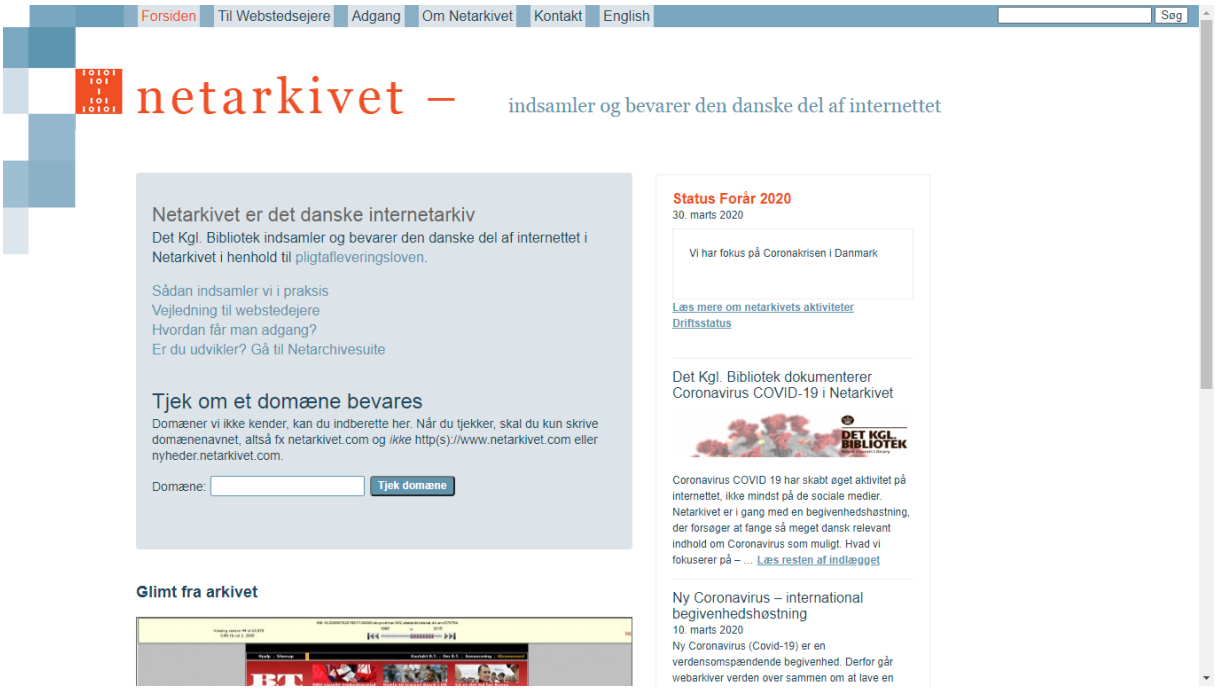
Heritrix je odabran za pobiranje unutar HAW-a nakon analize rezultata. Glavni razlozi za tu odluku bili su njegova mogućnost spremanja resursa u WARC formatu i fleksibilnost tj. mogućnost finog prilagođavanja raznih aspekata pobiranja koja je posebno važna pri pobiranju domena. Autori ističu da „veća fleksibilnost Heritrixa ujedno znači i da se od korisnika očekuje veća razina tehničkih znanja, tj. moglo bi se reći da je Heritrix više namijenjen informatičarima nego knjižničarima, dok za DAMP vrijedi obrnuto“.

9. PRIMJERI IZ INOZEMSTVA

9.1. Danska

Danski arhiv weba (Netarkivet, slika 15) djeluje od 2005. godine pod nadležnošću danske kraljevske knjižnice (Det Kongelige Bibliotek). Provodi domenska, selektivna i događajna pobiranja danskog weba. Domensko pobiranje bilježi sve sadržaje na .dk domeni četiri puta godišnje (vrlo velika web mjesta posjećuju se dvaput godišnje) prema popisu koji arhivu pruža administrator domene. Netarkivet je otkrio oko 45.000 web mjesta koja nisu na .dk domeni, ali se svejedno uključuju u ovaj proces zbog svojih sadržaja koji su na danskom jeziku i/ili usmjereni danskoj publici ili zbog prebivališta vlasnika. Neka mjesta unutar .dk domene mogu se zanemariti ako se otkrije da sadržaji nisu na danskom niti su usmjereni danskoj publici. Selektivno pobiranje provodi se za web mjesta koja se često ažuriraju, npr. portale s vijestima, često korištena dinamična web mjesta i sl. Netarkivet provodi selektivno pobiranje između šest puta dnevno i jedanput mjesečno, ovisno o učestalosti ažuriranja web mjesta. Pobiranje prema događajima prikuplja sadržaje s web mjesta posvećenih posebnim događajima čiji se nestanak predviđa nakon završetka toga događaja. Ovakva se pobiranja prosječno provode tri puta godišnje.

Netarkivet navodi da potpuno zanemaruje svako postojanje robots.txt datoteka jer se u protivnom ne bi mogli prikupiti svi relevantni sadržaji. Otkriveno je da mnoga važna web mjesta (vijesti, politika) imaju vrlo stroga robots.txt ograničenja: kad bi ih se pridržavao, arhiv bi s tih mjesta prikupio vrlo malo ili nimalo sadržaja. Arhiv priznaje da postoje žalbe na ovakav način poslovanja. Još jedna žalba odnosi se na ograničenost pristupa sadržaju: zbirci mogu pristupiti samo istraživači koji imaju doktorsku razinu obrazovanja ili su kandidati za doktorat. Razlog tome je zaštita osjetljivih osobnih podataka koji se mogu nalaziti čak i unutar sadržaja prikupljenih s javno dostupnih web mjesta. Arhiv se zbog toga ne može otvoriti širokoj javnosti te još nije riješen problem o tome kako identificirati osjetljive podatke i sakriti ih od javnosti uz istovremeno oslobađanje pristupa ostatku arhiva. Drugi problemi s kojima se Netarkivet suočava su zamke za pobirače, sadržaji skriveni iza korisničke prijave (arhiv procjenjuje da otprilike 16% takvih web mjesta sadrži materijale koji su pod obuhvatom zakona o obveznom primjerku i stoga nužni za prikupljanje), preopterećenja internetskih poslužitelja te otežano pobiranje audio i video sadržaja. U arhivu je zaposleno 20 kustosa i informatičara.

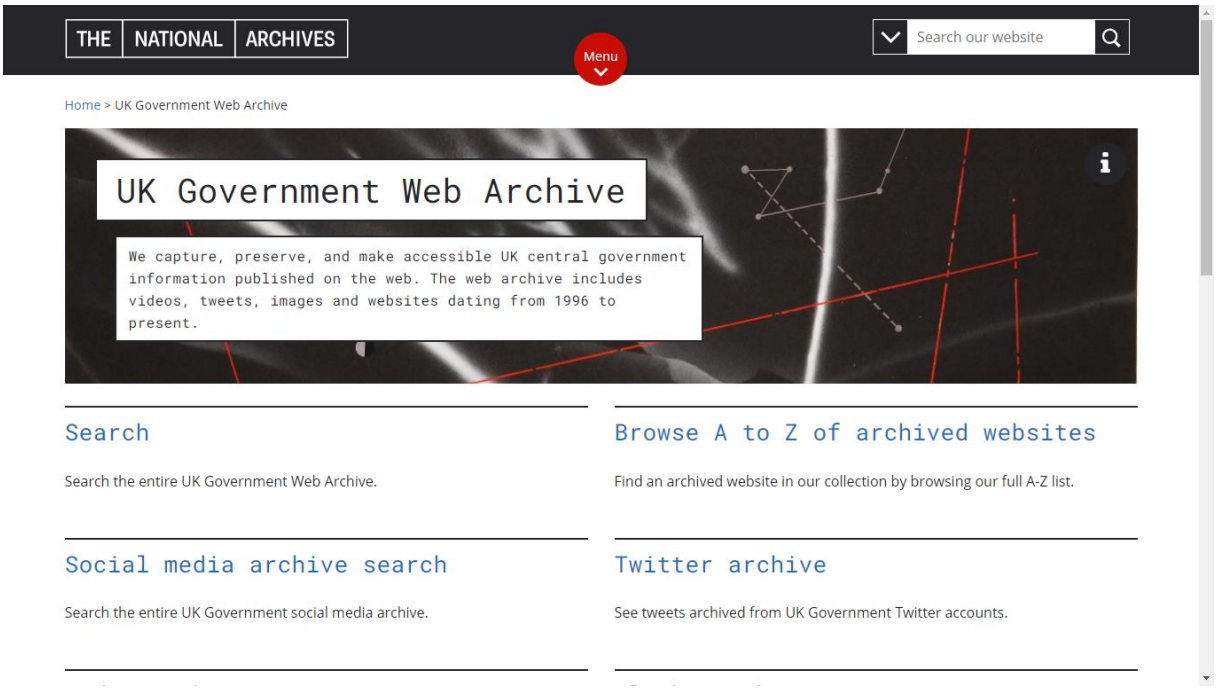


Slika 15. Web mjesto Netarkiveta. Izvor: Netarkivet.

9.2. Ujedinjeno Kraljevstvo

Dva arhiva weba djeluju unutar UK: arhiv vlade Ujedinjenog Kraljevstva (United Kingdom Government Web Archive, UKGWA) posvećen arhiviranju web sadržaja vlade (slika 16) i arhiv weba UK (UK Web Archive, UKWA) za općenito prikupljanje sadržaja britanskog weba (slika 17). UKGWA djeluje pod nadležnošću Nacionalnog arhiva UK (The National Archives, TNA), a UKWA prikuplja sadržaje u sklopu zakona o obveznom primjerku u ime šest knjižnica diljem UK: narodnih knjižnica Škotske i Walesa, Britanske knjižnice (The British Library) te sveučilišnih knjižnica Oxforda, Cambridgea i Dublina.

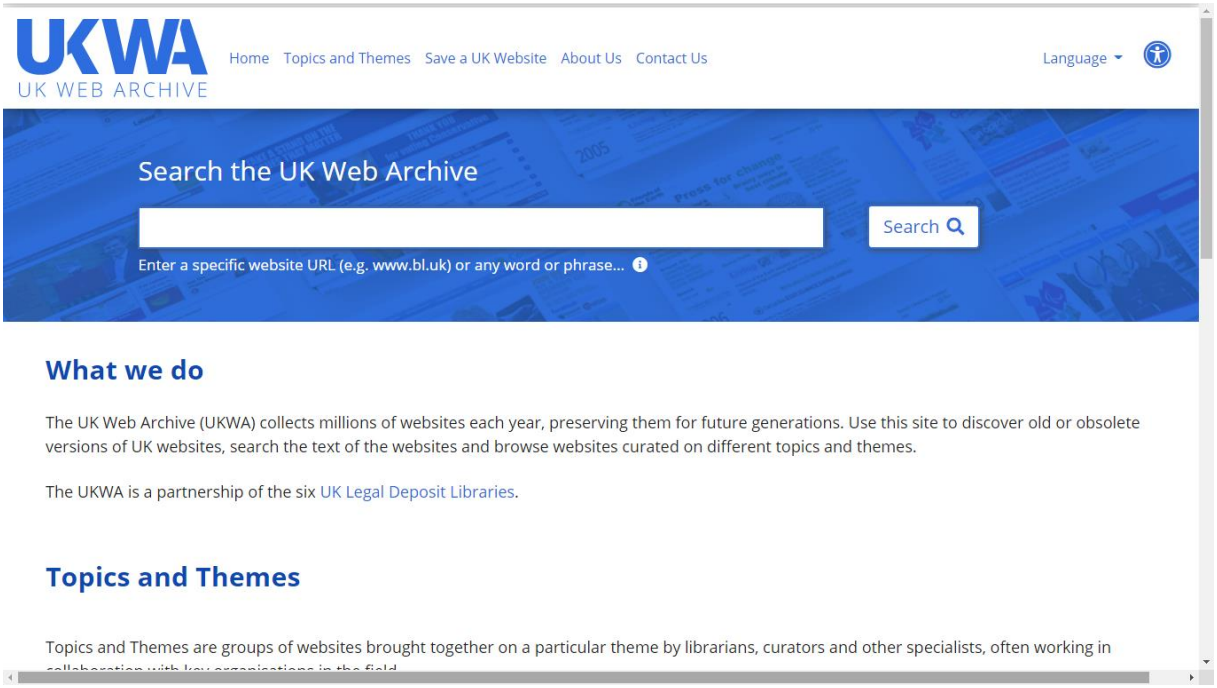
TNA od 2003. bilježi informacije koje proizvodi središnja vlada UK te ih pohranjuje u UKGWA. Prikupljeno je preko 5.000 web mjesta i većina materijala ima slobodan pristup i korištenje, osim ako ne dolaze s izvora čiji vlasnik mora dati dopuštenje za pristup i korištenje. Prikupljeni su i brojni materijali s društvenog weba, točnije s YouTubea i Twittera (ali samo oni sadržaji koje je stvorila britanska vlada ili se na nju odnose).



Slika 16. Web mjesto UKGWA. Izvor: UK Government Web Archive.

UKWA nastoji barem jednom godišnje prikupiti sve web sadržaje unutar UK. Uz automatsko pobiranje, provodi se i ručno u kojem sudjeluju kustosi i drugi stručnjaci. Prikuplja se samo materijal koji je slobodno dostupan na webu (dakle bez privatnih informacija kao što su sadržaji korisničkih računa i elektroničke pošte). Do 2017. je prikupljeno otprilike 500 TB podataka (60-70 TB godišnje). Većina sadržaja dolazi iz velikih godišnjih pobiranja glavnih domena (.uk, .scot, .wales, .cymru, .london) te s web mjesta čiji su poslužitelji ili vlasnici locirani unutar područja UK. Ponekad se mogu zatražiti i dopuštenja od vlasnika stranih web mjesta kako bi se izgradile dosljedne tematske zbirke. Odabrana web mjesta arhiviraju se prema različitoj učestalosti ažuriranja (portali s vijestima pobiru se svakodnevno). Arhivirani materijal može se pregledavati samo unutar prostora šest knjižnica koje sudjeluju u održavanju arhiva weba, osim ako vlasnici web mjesta ne daju dopuštenje za širi pristup. Zbirka UKWA pohranjena je na četiri lokacije istovremeno u gradovima St. Pancras (Engleska), Boston Spa (Engleska), Aberystwyth (Wales) i Edinburgh (Škotska) od kojih svaki posjeduje potpunu kopiju svih sadržaja unutar sustava. Kopije zbirke su u neprestanoj komunikaciji jedna s drugom i svaka može automatski rekonstruirati izgubljene ili oštećene datoteke koristeći podatke s bilo koje druge lokacije.

UKWA navodi problem nepotpunosti nekih svojih sadržaja iz više razloga. Neka web mjesta mogu se potpuno propustiti iako se provode i automatska i ručna pobiranja, a neka postoje samo unutar prostora čitaonice i ne može im se pristupiti bez dopuštenja vlasnika. UKWA trenutno ne može zabilježiti medije koji omogućuju internetski prijenos (engl. *streaming media*), skriveni web, sadržaje iza korisničkih računa i sl. Također, različiti web preglednici mogu drugačije prikazivati arhivirane web stranice iako UKWA nastoji biti kompatibilan sa što više različitih verzija. Korisnici mogu kontaktirati arhiv s pritužbom ako žele da se određeni materijali uklone radi zabrinutosti oko privatnosti ili intelektualnog vlasništva. UKWA upozorava korisnike da je potreban oprez pri stavljanju osobnih podataka na javno dostupan web jer se samo taj sadržaj prikuplja u arhiv.



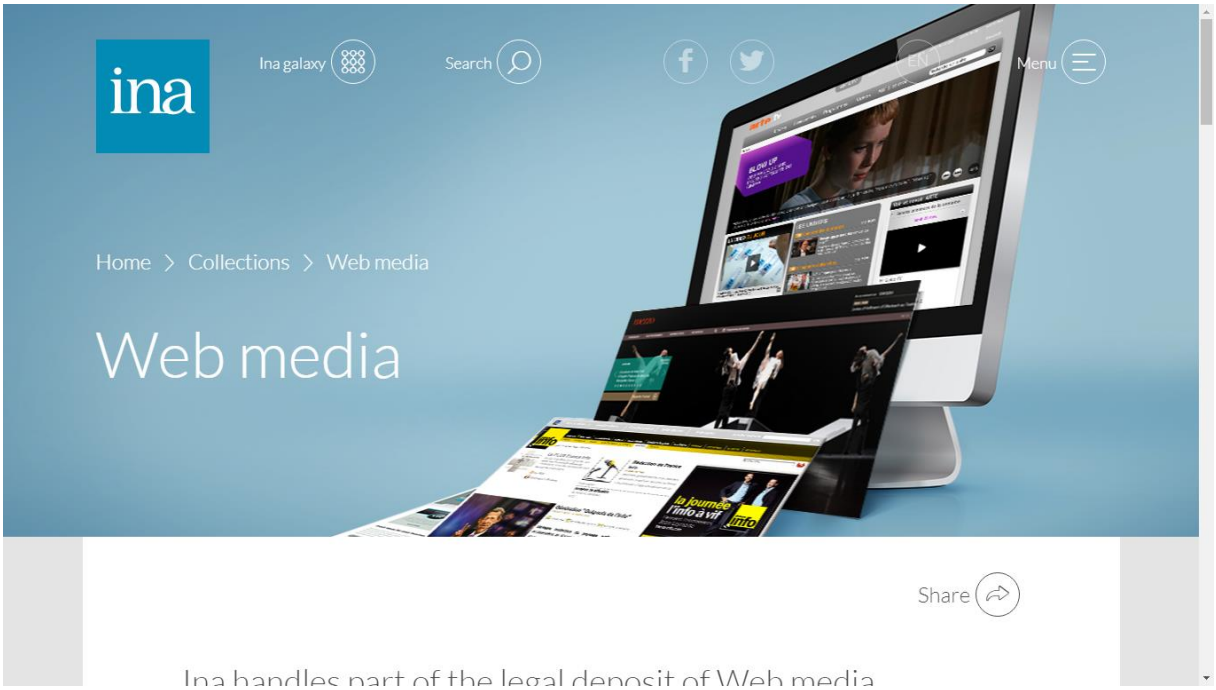
Slika 17. Web mjesto UKWA. Izvor: UK Web Archive.

9.3. Francuska

Nacionalna knjižnica Francuske (Bibliothèque nationale de France, BnF) i francuski Nacionalni audiovizualni institut (National Audiovisual Institute, INA) od 2006. surađuju u prikupljanju francuskih web materijala (slika 18). Pobiranje se provodi po domeni, temi i događajima. Pretraživanje se može vršiti po URL-u, ključnim riječima, punom tekstu i tematskim zbirkama. Domensko pobiranje obuhvaća sve sadržaje na .fr domeni te materijale

koji su stvoreni unutar Francuske ili čiji je izdavač iz Francuske. Posebne zbirke prikupljaju sadržaje o državnim, lokalnim i europskim političkim izborima, a tematske zbirke uključuju sadržaje koji dokumentiraju društvenu povijest francuskog weba (dnevnici, blogovi, aktivizam). BnF u stvaranju tematskih zbirki surađuje s različitim ustanovama poput regionalnih knjižnica i laboratorija. BnF arhivu weba se iz pravnih razloga može pristupiti samo unutar prostora knjižnice u Parizu. Knjižnica čuva dvije kopije arhiva na vrpci na dvije udaljene lokacije i jednu kopiju na disku. Do 2014. se koristio ARC format datoteka, a nakon njega WARC.

INA se bavi prikupljanjem audiovizualnih materijala francuskog weba s otprilike 34.000 različitih izvora. Danas je u INA-i očuvano preko 14.000 web mjesta u skladu sa zakonom o obveznim primjercima na webu. Sadržaji koje INA prikuplja uključuju radijske i televizijske programe emitirane putem weba te ostale sadržaje društvenih mreža i video platformi. Učestalost pobiranja i količina prikupljenih informacija određuje se prema ažuriranju i veličini svakog web mjesta. INA je kroz članstvo unutar IIPC-a i suradnju s IA uspjela dopuniti svoju zbirku sadržajima koji sežu sve do 1996., odnosno do samog početka arhiviranja weba u svijetu.



Slika 18. Web mjesto INA-e. Izvor: INA.

9.4. SAD

Arhiv weba Kongresne knjižnice (Library of Congress, LoC) u Washingtonu bavi se prikupljanjem web sadržaja od 2000. Ne provodi pobiranje vršne domene već stvara samo tematske i događajne zbirke čije sadržaje ručno odabiru knjižničari. Web mjesta koja se prikupljaju uključuju odabrana vladina tijela s područja zakonodavstva ili poslovanja, neka web mjesta stranih vlada, izvore posvećene političkim izborima u SAD-u i odabranim stranim državama, novinarstvo, vijesti i sl. Knjižnično osoblje odabire URL-ove koji će služiti kao početna točka pobiračima. Oni mogu biti puna domena, poddomena ili samo jedna web stranica ili dokument. Učestalost prikupljanja ovisi o prirodi samog web mjesta i odlukama osoblja (koje se ponekad mogu promijeniti). Korisnici mogu knjižnici poslati prijedloge za arhiviranje određenih web mjesta, ali ne računa se da će se oni uvažiti. LoC nastoji prikupiti što je moguće više sadržaja kako bi se web mjesta vjerno očuvala i prezentirala, što znači da se uključuju svi materijali potrebni za pružanje konteksta budućim istraživačima (HTML stranice, PDF datoteke, slike, audio i video sadržaji). Materijali su pohranjeni u formatima WARC i ARC (za starije datoteke) i unutar knjižnice postoji više kopija sadržaja arhiva. Knjižnica od 2009. godine koristi pobirač External koji provodi deduplikaciju sadržaja za smanjenje zauzetog prostora za pohranu. Deduplikacija se vrši provođenjem naknadnih pobiranja već posjećenih web mjesta: pobirač uklanja duplikate i pohranjuje samo nove i izmijenjene sadržaje. Za prikazivanje arhiviranog sadržaja koristi se verzija OpenWayback External. U arhivu je pohranjeno preko dva PB sadržaja i zbirka raste za otprilike 20-25 TB mjesečno.

Knjižnično osoblje odabire URL-ove koji će služiti kao početna točka pobiračima. Moguće je i da neki sadržaji neće biti potpuno opisani i katalogizirani zbog velikog opsega zbirke. Također se pojavljuju situacije u kojima su neka web mjesta dostupna kroz URL pretraživanje, ali još nisu potpuno obrađena za pristup od strane osoblja. Neka web mjesta nisu dostupna korisnicima zbog toga što se svi sadržaji najprije godinu dana zadržavaju u arhivu. Nakon isteka tog vremenskog roka sadržaju se pruža slobodan pristup, osim ako njegov vlasnik ne naloži drugačije. U takvim se slučajevima gradivu može pristupiti samo unutar prostora knjižnice. Ostali izazovi s kojima se LoC suočava su nemogućnost arhiviranja medija koji omogućuju internetski prijenos, skrivenog weba i sadržaja koji zahtijevaju korisničku prijavu, pretplatu ili naknadu za pristup. Društvene mreže i neke platforme izdavaštva također mogu biti teške za očuvanje. Također je teško rukovati sadržajima koji

imaju umetnute sadržaje s trećih strana npr. račune s društvenih mreža. Takvi URL-ovi imaju manji prioritet pri pobiranju pa se može dogoditi da se neka web mjesta nepotpuno arhiviraju tj. unutar arhiva se pojavljuju praznine.

Slika 19. Web mjesto LoC. Izvor: Library of Congress.

10. ARHIVIRANJE DRUŠTVENOG WEBA

Arhiviranje društvenog weba predstavlja poseban izazov zbog njegove ekstremne dinamičnosti i opsega. Također, mnoga takva web mjesta zahtijevaju stvaranje korisničkog računa prije ikakvog pristupanja sadržaju. Važne javne organizacije i ličnosti koriste barem neki oblik društvenih mreža za komunikaciju sa svijetom, zbog čega prikupljanje takvih sadržaja može biti važno za buduće generacije. Arhivi weba uglavnom ne prikupljaju sadržaje iza korisničkih računa već samo sadržaje koji se javno objavljuju, što znači da korisnici trebaju biti svjesni mogućnosti da će neke ili sve njihove objave s društvenih mreža možda biti trajno arhivirane za buduće generacije.

10.1. SAD

10.1.1. ArchiveSocial

ArchiveSocial (AS, slika 20) organizacija djeluje od 2011. sa sjedištem u Durhamu (Sjeverna Karolina) i razvila je program istog imena za arhiviranje društvenog weba. AS navodi da poslovne agencije imaju obvezu slijediti američke zakone: društvene mreže se u svih 50 saveznih država smatraju javnim spisima, zbog čega je potrebno osigurati pouzdan način za očuvanje tih spisa. AS se povezuje izravno s društvenim mrežama kako bi zabilježio i očuvao sadržaje tih stranica u originalnom formatu s potpunim tehničkim metapodacima. Svakom se zapisu dodjeljuje provjerena vremenska oznaka i digitalni potpis kako bi se osigurala autentičnost dokumenta tj. dokazalo njegovo postojanje i zajamčilo da se nije izmijenio ni na koji način. Tehnologija također može detektirati i zabilježiti preuređene, izbrisane ili skrivene sadržaje. AS nudi alat za upravljanje rizikom koji može obavijestiti vlasnika računa o neprimjerenim aktivnostima (zahtijevanje osobnih informacija, nedozvoljen vokabular) koje mogu dovesti do narušavanja javne slike agencije ili pravnih problema. AS navodi da bi se svi sadržaji trebali očuvati u izvornom formatu, što uključuje i bogate medije, npr. slika se mora sačuvati u svojoj punoj rezoluciji, a ne kao poveznica ili umanjena verzija. Pretraživanje je moguće kroz ključne riječi, datum, mrežu, korisnička imena, tip sadržaja ili tagove. AS je baziran na oblaku i ne zahtijeva instaliranje softvera niti informatičku stručnost.

Slika 20. Web mjesto AS. Izvor: ArchiveSocial.

10.1.2. Kongresna knjižnica (Twitter)

Kongresna knjižnica (LoC) u SAD-u 2010. je godine najavila da počinje s prikupljanjem svih javnih objava na Twitteru, počevši sa sadržajima iz 2006. godine tj. od samog osnutka te društvene mreže. Projekt je trajao do kraja 2017. kada je knjižnica objavila da se početkom iduće godine prebacuje na selektivno pobiranje tekstualnih sadržaja s Twittera. Razlozi za ovu promjenu su sljedeći: nagao i brz rast društvenih mreža, promjena prirode Twittera (sadržaji su postali sve više vizualni, a prvotni cilj knjižnice bio je prikupiti samo tekstualne objave) i proširenje opsega samih objava. LoC napominje da uglavnom ne provodi sveobuhvatna prikupljanja, ali takav se projekt započeo u slučaju Twittera zbog toga što je u to vrijeme smjer razvoja društvenog weba bio nepoznat. Budući da su društvene mreže danas razvijene, knjižnica će početi provoditi prikupljanja koja su više u skladu s njezinom politikom nabave. Sadržaji zbirke Twittera trenutno su izvan dohvata javnosti dok se ne riješe problemi pristupa.

10.2. Kina

Prema Deng (2019.), Narodna knjižnica Kine namjerava provesti neprofitan projekt arhiviranja preko 200 milijardi javnih objava s društvene mreže Weibo (funkcionira nalik

Twitteru) kako bi se očuvala „digitalna baština najveće internetske populacije“. Portal s vijestima Sina odabran je kao prvi partner u projektu (druge se tvrtke također pozivaju na sudjelovanje) zbog svoje goleme količine podataka koji bilježe značajne društvene događaje i reakcije javnosti. Podaci će se pohranjivati na poslužitelje Sine gdje će ih ona zajedno s knjižnicom analizirati za akademske i pravne svrhe.

Slika 21. Web mjesto Weibo-a. Izvor: Weibo.

11. ZAKLJUČAK

Arhiviranje weba zahtjevna je i kompleksna djelatnost koja postaje sve zastupljenija diljem svijeta. Neophodna je ne samo za arhive i knjižnice već i za svaku organizaciju ili pojedinca čije je poslovanje prisutno na webu. Također je korisno imati znanje o arhiviranju weba iz privatnih razloga kao što su stvaranje osobnih arhiva, pretraživanje materijala koji više nisu na živom webu i razumijevanje rizika od objavljivanja sadržaja na javno dostupnim web mjestima. Iako se ova djelatnost suočava s brojnim tehničkim i društvenim izazovima (što je i razumljivo s obzirom na njezinu relativno nedavnu pojavu), arhiviranje relevantnih dijelova weba trebalo bi postati redovita praksa u svim zemljama svijeta kako bi se očuvala digitalna kulturna baština.

Arhivi weba djeluju unutar različitih pravnih okvira i s različitim ograničenjima pristupa. Pravni stručnjaci moraju biti spremni za rješavanje sporova oko često nejasnih granica i pravila unutar djelatnosti arhiviranja weba. „Obični“ korisnici također se mogu uključiti u rad lokalnih arhiva i knjižnica te ih obavještavati o važnim web mjestima koja bi im mogla promaknuti. Stvaranje vlastitih zbirki weba u poslovne ili osobne svrhe također je moguće bez potrebe za stupanjem u kontakt s arhivima/knjižnicama ili posjedovanjem informatičkog predznanja. Dostupne su mnoge besplatne usluge za pohranu i pregledavanje arhiviranog web sadržaja.

Nužno je omogućiti pristup arhiviranim web sadržajima što većem broju korisnika jer bi baština (uključujući i digitalnu) trebala biti dostupna svima. Ovo je područje na mnogo načina još neistraženo i nedefinirano, ali sve veća svijest o nužnosti ove djelatnosti i konstantan razvoj novih usluga i inicijativa pruža obećavajući pogled na budućnost arhiviranja weba.

12. LITERATURA

1. Andrews, E. (2013.) Who Invented the Internet? History.com. URL: https://www.history.com/news/who-invented-the-internet (2.6.2020.)
2. Archive Access. URL: http://archive-access.sourceforge.net/projects/wera/articles/what-is-wera.html (14.9.2020.)
3. Archive Social. URL: https://archivesocial.com/ (8.6.2020.)
4. Arhiv. (2020.) Hrvatska enciklopedija, mrežno izdanje. Leksikografski zavod Miroslav Krleža. URL: http://www.enciklopedija.hr/Natuknica.aspx?ID=3764 (1.6.2020.)
5. Ball, A. (2010.) Web Archiving. University of Bath. URL: https://dcc.ac.uk/guidance/briefing-papers/technology-watch-papers/web-archiving (14.6.2020.)
6. Bouard, A. (2014.) Crawl Speed: How Many Pages/Second? 7 Points To Take Into Account. Botify. URL: https://www.botify.com/blog/crawler-impact-performance (2.6.2020.)
7. Celjak, D. i Milinović, M. (2012). Harvestiranje hrvatskoga weba: arhitektura programskoga sustava za harvestiranje i iskustva stečena njegovom upotrebom. U D. Hasenay, (Ur.), M. Krtalić, (Ur.), 15. seminar Arhivi, knjižnice, muzeji : Mogućnosti suradnje u okruženju globalne informacijske infrastrukture, 144-160. Zagreb: Hrvatsko knjižničarsko društvo. URL: https://urn.nsk.hr/urn:nbn:hr:102:141793 (23.6.2020.)
8. Cloudflare. URL: https://www.cloudflare.com/learning/bots/what-is-a-web-crawler/ (3.3.2020.)
9. Costa, M., Gomes, D., Silva, M.J. (2017.) The evolution of web archiving. Int J Digit Libr 18, 191–205. URL: https://link.springer.com/article/10.1007%2Fs00799-016-0171-9 (13.5.2020.)
10. Dans, E. (2019.) Immortality at last! Now there's a solution to link rot. Medium. URL: https://medium.com/enrique-dans/immortality-at-last-now-theres-a-solution-to-link-rot-cdc90ccce685 (7.6.2020.)
11. Deng, I. (2019.) China's national library to archive 200 billion Weibo posts in project to preserve country's digital heritage. South China Morning Post. URL: https://www.scmp.com/tech/apps-social/article/3007115/chinas-national-library-archive-200-billion-weibo-posts-project (8.6.2020.)

12. Everything you need to know about the "Right to be forgotten". GDPR.eu. URL: https://gdpr.eu/right-to-be-forgotten/ (14.9.2020.)
13. Heritrix. Internet Archive. GitHub. URL: https://github.com/internetarchive/heritrix3/wiki (24.6.2020.)
14. How much data is generated each day? World Economic Forum. URL: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/ (1.6.2020.)
15. Hrvatski arhiv weba. Nacionalna i sveučilišna knjižnica u Zagrebu. URL: https://haw.nsk.hr/ (13.4.2020.)
16. Ina. URL: https://institut.ina.fr/en/collections/web-media (17.6.2020.)
17. International Internet Preservation Consortium. URL: http://netpreserve.org/ (12.2.2020.)
18. Internet Archive. URL: https://archive.org/ (3.5.2020.)
19. ISO/TR 14873:2013. Information and documentation: Statistics and quality issues for web archiving. International Organization for Standardization. URL: https://www.iso.org/obp/ui/#iso:std:iso:tr:14873:ed-1:v1:en (15.6.2020.)
20. Larivière, J. (2000.) Guidelines for legal deposit legislation: A revised, enlarged and updated edition of the 1981 publication by Dr. Jean Lunn. United Nations Educational, Scientific and Cultural Organization. URL: https://www.ifla.org/publications/guidelines-for-legal-deposit-legislation (20.6.2020.)
21. Leetaru, K. (2016.) The Internet Archive Turns 20: A Behind The Scenes Look At Archiving The Web. Forbes. URL: https://www.forbes.com/sites/kalevleetaru/2016/01/18/the-internet-archive-turns-20-a-behind-the-scenes-look-at-archiving-the-web/#2b26162e82e0 (8.5.2020.)
22. Library of Congress. URL: https://www.loc.gov/ (5.3.2020.)
23. Link Rot: The Web is Decaying. (2019.) Decentralised Public Library. Medium. URL: https://medium.com/@arweave/link-rot-the-web-is-decaying-cc7d1c5ad48b (8.6.2020.)
24. Masanès, J. (2006.) Web Archiving. Berlin: Springer.
25. McCarthy, K. (2018.) Archive.org's Wayback Machine is legit legal evidence, US appeals court judges rule. The Register. URL: https://www.theregister.com/2018/09/04/wayback_machine_legit/ (4.6.2020.)
26. Memento Web. URL: http://mementoweb.org/about/ (18.6.2020.)
27. Netarkivet. URL: http://netarkivet.dk/in-english/ (17.6.2020.)

28. Newing, C. (2017.) How do you move a web archive? The National Archives. URL: https://blog.nationalarchives.gov.uk/move-web-archive/ (8.5.2020.)
29. Niu, J. (2012.) An Overview of Web Archiving. School of Information Faculty Publications. 308. URL: http://scholarcommons.usf.edu/si_facpub/308 (5.11.2019.)
30. Niu, J. (2012.) Functionalities of Web Archives. School of Information Faculty Publications. 309. URL: http://scholarcommons.usf.edu/si_facpub/309 (3.2.2020.)
31. Quintillion definition. Maths is fun. URL: https://www.mathsisfun.com/definitions/quintillion.html (9.6.2020.)
32. Sbforge. URL: https://sbforge.org/display/NASDOC60/Heritrix+Control+and+GUI-console+Access (14.9.2020.)
33. Taylor, N. (2012.) Designing Preservable Websites, Redux. The Signal, Library of Congress. URL: https://blogs.loc.gov/thesignal/2012/02/designing-preservable-websites-redux/ (19.2.2020.)
34. The Oakland Archive Policy. (2002.) Recommendations for Managing Removal Requests And Preserving Archival Integrity. School of Information Management and Systems, U.C. Berkeley. URL: https://groups.ischool.berkeley.edu/archive/aps/removal-policy (15.6.2020.)
35. The WARC Format 1.1, WARC Specifications. URL: https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/ (15.6.2020.)
36. UK Government Web Archive. The National Archives. URL: https://www.nationalarchives.gov.uk/webarchive/ (17.6.2020.)
37. UK Web Archive. URL: https://www.webarchive.org.uk/ukwa/ (17.6.2020.)
38. Weibo. URL: https://weibo.com/overseas (18.9.2020.)
39. Weigle, M. (2018.) On the Importance of Web Archiving. Social Science Research Council. URL: https://items.ssrc.org/parameters/on-the-importance-of-web-archiving/ (3.3.2020.)
40. Welcome to the Permaweb. (2019.) The Arweave Project. Medium. URL: https://medium.com/@arweave/welcome-to-the-permaweb-ce0e6c73ddfb (8.6.2020.)
41. Willer, M. (2017.) ARHIVIRANJE HRVATSKIH MREŽNIH PUBLIKACIJA: od projekta do programa arhiviranja u Nacionalnoj i sveučilišnoj knjižnici u
42. Zagrebu (NSK). Zagreb: Nacionalna i sveučilišna knjižnica. URL: http://bosniaca.nub.ba/index.php/bosniaca/article/view/159 (14.9.2020.)

43. Zakon o knjižnicama i knjižničnoj djelatnosti NN 17/19, 98/19, Zakon.hr URL: https://www.zakon.hr/z/2275/Zakon-o-knjižnicama-i-knjižničnoj-djelatnosti (10.6.2020.)

POPIS SLIKA

SAŽETAK

Diplomski rad bavi se pitanjima arhiviranja web sadržaja. Rad najprije donosi kratak pregled povijesti arhiviranja weba i objašnjenje osnovnih pojmova unutar ove djelatnosti (pobirači, vrste pobiranja, robots.txt datoteke itd.) Zatim se spominju pravni okviri arhiviranja weba koji uključuju smjernice UNESCO-a o očuvanju digitalne baštine te zakon o obveznom primjerku. Navode se najpoznatije svjetske organizacije, usluge i alati arhiviranja weba (Internet Archive, Wayback Machine, Heritrix, Archive-It, IIPC, ARC i WARC formati datoteka, WARCreate, Memento i sl.) Komentira se brz i širok rast inicijativa arhiviranja weba u razdoblju od samo četiri godine. Raspravlja se o važnosti arhiviranja weba za različite svrhe koje mogu uključivati i korištenje arhiviranih web sadržaja u sudskim procesima. Nabrojeni su neki od najčešćih tehničkih i društvenih izazova unutar kompleksne djelatnosti arhiviranja weba (propadanje poveznica, premještanje sadržaja, skriveni web, uklanjanje sadržaja itd.) Navode se primjeri arhiva weba u Hrvatskoj i svijetu (Danska, UK, Francuska i SAD), s naglaskom na Hrvatski arhiv weba (HAW). Uspoređuju se dva različita pobirača unutar HAW-a. Zaključno se ističu inicijative arhiviranja društvenog weba s primjerima iz SAD-a i Kine.

Ključne riječi: arhiviranje weba, pobirači, pobiranje, dugoročno očuvanje, arhiv

WEB ARCHIVING

SUMMARY

The thesis discusses the issues of archiving web content. Firstly, a brief overview of web archiving history is presented, as well as the explanations of common terms within this field (crawlers, types of harvesting, robots.txt files etc.) Legal frameworks of web archiving which include UNESCO's guidelines on preserving digital heritage and legal deposit laws are also elaborated. The most well-known world organizations, services and tools of web archiving are listed (Internet Archive, Wayback Machine, Heritrix, Archive-It, IIPC, ARC and WARC file formats, WARCreate, Memento etc.) The rapid and wide growth of web archiving initiatives during a span of only four years is discussed. The importance of web archiving for different purposes which can include usage of archived web content in court proceedings is pointed out. Some of the most common technical and social challenges within the complex profession of web archiving (link decay, transferring content, hidden web, removal of content etc.) are analysed. Examples of web archives in Croatia and worldwide (Denmark, the UK, France, and USA) are included, with emphasis on the Croatian web archive. Two different crawlers within the archive are compared. In conclusion, initiatives of archiving the social web in the USA and China are illustrated.

Key words: web archiving, crawlers, harvesting, long term preservation, archive