Organizacija znanja, 30 (1–2), 2025, številka prispevka, https://doi.org/10.3359/oz2530002
Strokovni članek / Professional article

Kako nastajajo spletni arhivi: tehnični vidiki zajemanja spletnih vsebin

How are web archives created? Technical aspects of web content capture

Janko Klasinc[1]

IZVLEČEK: Spletni arhivi so zbirke podatkov, ki so rezultat prizadevanj knjižnic in drugih sorodnih ustanov za trajno ohranjanje spletne dediščine. Pogosto vsebujejo velike količine gradiva, ki je s spleta shranjeno z uporabo spletnih robotov, s stališča uporabe pa so spletni arhivi pogosto nepredvidljivi, netransparentni in nekonsistentni viri podatkov, ki vsebujejo številne vsebinske vrzeli. Poleg različnih družbenih, zakonodajnih in institucionalnih okoliščin, v okviru katerih nastajajo, njihove specifične lastnosti v veliki meri opredeljuje tudi heterogena, efemerna in fluidna narava svetovnega spleta. Ker uporabnikom spletni arhivi predstavljajo številne izzive, je pomembno, da se ti zavedajo okoliščin, ki vplivajo na naravo spletnih arhivov ter posledično na priložnosti in pasti uporabe arhiviranih podatkov. Da bi osvetlili ozadje teh relativno slabo poznanih virov podatkov, v prispevku na osnovi pregleda temeljne in druge relevantne literature opisujemo predvsem tehnične vidike nastajanja spletnih arhivov. Pri tem se osredotočamo na temeljne značilnosti svetovnega spleta v kontekstu ohranjanja, na različne pristope k zajemanju spletnih vsebin, njihove omejitve in vpliv teh okoliščin na naravo spletnih arhivov, ki se kot viri podatkov v marsičem razlikujejo od bolj tradicionalnih in uveljavljenih zbirk.

KLJUČNE BESEDE**:** spletni arhivi, arhiviranje spleta, spletna dediščina

ABSTRACT: Web archives are collections produced by libraries and other heritage institutions to permanently preserve online heritage. They often contain large amounts of material stored from the web through the use of web crawlers. From a usage perspective, they are often unpredictable, non-transparent and inconsistent data sources that contain numerous content gaps. In addition to the various social, legislative and institutional circumstances under which they are created, their specific characteristics are largely defined by the heterogeneous, ephemeral and fluid nature of the world wide web. Because they present numerous challenges to their users, it is important for them to be aware of the circumstances that influence the nature of web archives and, consequently, the opportunities and pitfalls of using archived data. To shed light on the background of these relatively poorly understood data sources, this paper, through a review of foundational and other relevant literature, describes primarily the technical aspects of web archives creation. It focuses on the fundamental characteristics of the world wide web in the context of preservation, different approaches to capturing web content, their limitations and the impact of these circumstances on the nature of web archives, which differ in many ways from more traditional and established data sources.

KEYWORDS: web archives, web archiving, web heritage

1 Uvod

Trajno ohranjanje kulturne, znanstvene in druge pisne dediščine je eno od temeljnih poslanstev knjižnic, arhivov in drugih sorodnih ustanov. Razvoj novih tehnologij, predvsem

[1] Janko Klasinc, univ. dipl. bibl., Narodna in univerzitetna knjižnica, Ljubljana, Slovenija, janko.klasinc@nuk.uni-lj.si.

tistih s področja informacijske tehnologije, botruje neprestanemu razvoju postopkov pridobivanja, ohranjanja in dajanja v uporabo raznovrstnih vsebin, ki v različnih kontekstih veljajo za kulturno dediščino. Tako so v preteklosti dediščinske ustanove poleg skrbi za gradivo na papirju in sorodnih nosilcih morale razviti tudi ustrezne postopke za zbiranje in varovanje gradiva na magnetnih in optičnih nosilcih ter gradiva v različnih digitalnih oblikah.

Čeprav smo gradivo v digitalni obliki poznali že pred svetovnim spletom, sta s pojavom slednjega zapisovanje in razširjanje informacij ter znanja dosegla nove dimenzije. Od sredine devetdesetih let prejšnjega stoletja svetovni splet igra vedno večjo vlogo v številnih družbah in predstavlja enega od glavnih temeljev današnje komunikacijske infrastrukture (Brüger in Laursen, 2019). Čeprav so se v preteklosti pojavljali pomisleki o smiselnosti sistematičnega ohranjanja vsebin na spletu, ki so bili utemeljeni z vprašljivo kakovostjo spletnih objav, prepričanjem, da je splet medij, ki se lahko arhivira sam (angl. *self-preserving medium*), ali z neizvedljivostjo naloge zaradi obsežnosti spleta in potencialnih avtorskopravnih omejitev (Masanès, 2006), je danes arhiviranje spleta prepoznano kot ena od pomembnih, čeprav še vedno relativno obrobnih dejavnosti za ohranjanje pisne dediščine.

V prispevku osvetljujemo bistvene značilnosti svetovnega spleta kot medija, ki pogojujejo pristope k njegovemu ohranjanju. Preučujemo tudi različne metode sistematičnega shranjevanja vsebin s spleta, pri katerih se osredotočamo na najbolj razširjen pristop zajemanja z uporabo spletnih robotov. Opisali smo tudi nekatere bistvene značilnosti spletnih arhivov[2], po katerih se razlikujejo od drugih, bolj uveljavljenih zbirk podatkov.

2 Svetovni splet v kontekstu ohranjanja

Bibliotekarski terminološki slovar (Kanič et al., 2009) definira svetovni splet kot »porazdeljen internetni informacijski sistem, v katerem so spletne strani, spletni dokumenti povezani s hiperpovezavami«. Splet je unikaten informacijski vir, ki vsebuje več milijonov spletnih mest in povezuje skupnosti ter posameznike z vsega sveta. Od svojega začetka, ko je predstavljal relativno omejeno storitev, ki so jo uporabljali predvsem raziskovalci, se je v le nekaj desetletjih razvil v globalni informacijski medij. Danes ne predstavlja več le sredstva za komuniciranje, pač pa unikaten vir informacij o življenju v 21. stoletju. Hitrost njegovega razvoja hkrati predstavlja grožnjo naši kulturni, družbeni in tehnični dediščini v digitalni obliki (Pennock, 2013). Človeštvo namreč še nikoli v zgodovini ni proizvajalo toliko informacij kot danes in te še nikoli niso bile tako široko dostopne, hkrati pa še nikoli do zdaj nismo imeli opravka s tako veliko količino izgubljenih informacij. Živimo v obdobju preobilja informacij in pomanjkanja spomina (Gomes et al., 2021).

Osnovni razlog za trajno ohranjanje informacij v kakršni koli obliki je spoznanje, da imajo različni informacijski objekti ne glede na namen, za katerega so bili ustvarjeni, tudi določeno trajno vrednost. Poleg tega je eden od glavnih razlogov za prizadevanja za ohranjanje spletnih virov, predvsem v okviru dediščinskih ustanov, relativno hitro izginjanje tovrstnih vsebin – splet je vseprisoten, vendar so posamezne spletne strani minljive (Pennock, 2013). Za

[2] Izraz spletni arhiv je prevod angleškega izraza web archive in je v Sloveniji najbolj uveljavljen izraz za tovrstne zbirke. Kljub temu je ta izraz zaradi slabe prepoznavnosti področja arhiviranja spleta pogosto narobe razumljen, tako strokovna kot širša javnost pa ga zamenjujeta z digitaliziranimi arhivi, digitalnimi knjižnicami in drugimi digitalnimi zbirkami. Kljub temu da bi bil morda ustreznejši izraz arhiv spleta, v besedilu uporabljamo bolj uveljavljeni spletni arhiv. Po mnenju Terminološke svetovalnice pri ZRC SAZU z dne 5. 7. 2024 sta sicer ustrezna oba izraza, za najboljšo rešitev pa bi bila gotovo smiselna širša razprava strokovnjakov z različnih področij (Atelšek et al., 2024).

izginjanje celotnih spletnih mest obstaja več razlogov. Eden od najpogostejših je namerno ali nenamerno zanemarjanje, zaradi katerega poteče veljavnost spletne domene, hkrati pa ni poskrbljeno za varnostne kopije vseh relevantnih datotek. Zanemarjanje je pogosto povezano s pomanjkanjem finančnih sredstev za vzdrževanje spletnega mesta ali z upadom motivacije lastnika za dopolnjevanje in vzdrževanje vsebine spletnega mesta. Razlogi so lahko tudi tehnične narave in vključujejo okvare strojne opreme, viruse in zlonamerno programsko opremo ter nenamerno brisanje datotek. Podobno kot velja za fizično gradivo, tudi računalniško opremo ogrožajo naravne nesreče, kot so požari in poplave ter burne družbeno-politične situacije, v katerih je lahko gradivo umaknjeno s spleta zaradi ideoloških ali političnih razlogov. Zaradi tržnih bojev med spletnimi podjetji se lahko pojavijo prevzemi spletnih storitev, ki so včasih potem ukinjene ali priključene večjim platformam (Barone, 2015; Milligan, 2019). Major (2021) navaja tudi opustitev spletnih mest, kadar lastniki komercialnih storitev ugotovijo, da te ne prinašajo (več) pričakovanega dobička, nekatera spletna mesta pa izginejo, ker ne izpolnjujejo več svojega osnovnega namena, na primer spletna mesta političnih kampanj, ki so aktualna samo v določenem časovnem obdobju.

Poleg izginevanja celotnih spletnih mest sta zelo pogosta tudi spreminjanje lokacije posameznih vsebin, ki vodi v odmiranje hiperpovezav (angl. *link rot*[3]), in izginevanje ali spreminjanje zgolj posameznih delov spletnih mest. V tem kontekstu Masanès (2006) kot eno od bistvenih lastnosti spleta, po katerem se razlikuje od gradiva na fizičnih nosilcih, navaja kardinalnost[4]. Arhivi in muzeji večinoma hranijo gradivo, ki obstaja v enem primerku (unikatu), medtem ko imajo knjižnice, predvsem od izuma tiska dalje, opravka z gradivom, pri katerem obstaja več kopij istega dela (npr. vsi izvodi določene izdaje publikacije). Ta lastnost zvišuje verjetnost za ohranitev vsebin, saj lahko knjižnice pridobijo izvod določene publikacije tudi dolgo po njenem izidu, večje število kopij pa zmanjšuje možnost, da bi določeno delo povsem izginilo. Po drugi strani je možno kopije spletnega gradiva ustvarjati v praktično neomejenem obsegu, preprosto in z nizkimi stroški, vendar se lahko te bolj ali manj razlikujejo med seboj in od originala. Prikaz določene vsebine je namreč lahko prirejen določenemu spletnemu brskalniku ali pogojen s časom dostopa, geografsko lokacijo in drugimi značilnostmi uporabnika. S stališča ohranjanja to pomeni, da je pogosto možno zajeti in shraniti le eno ali omejeno število vseh pojavnosti (kopij) določene vsebine.

Praktično delovanje spleta v največji meri opredeljujejo standardi URI[5], HTTP[6] in HTML[7], zaradi katerih je možno prek vsakega računalnika, ki je povezan z internetom, objavljati vsebine na spletu. Zaradi široke dostopnosti in enostavnosti objavljanja ter brisanja raznovrstnih vsebin na spletu se je izredno povečalo število ljudi, ki ustvarjajo in razširjajo informacije, in posledično tudi količina teh informacij. Če ob tem upoštevamo tudi hipertekstovno naravo spleta, s katero so heterogene vsebine na globalni ravni med seboj povezane na različne načine, lahko ugotovimo, da je ohranjanje spletnih vsebin možno le v omejenem obsegu.

[3] Angleški izraz *link rot* pomeni odmiranje hiperpovezav, ki zaradi prenehanja obstajanja ali premika strani, na katere kažejo, ne delujejo več (Kanič, I. et al., 2020).

[4] V kontekstu spletnega gradiva se pojem nanaša na številčnost instanc oziroma možnih pojavnosti iste vsebine.

[5] URI (Uniform Resource Identifier) je enotni identifikator vira, ki določa ime ali lokacijo fizičnega ali logičnega vira, ki ga uporabljajo spletne tehnologije (Kanič, I. et al., 2020).

[6] HTTP (HyperText Transfer Protocol) je protokol za izmenjavo hiperteksta ter grafičnih, zvočnih in drugih večpredstavnostnih vsebin na spletu (Kanič, I. et al., 2020).

[7] HTML (HyperText Markup Language): označevalni jezik za oblikovanje večpredstavnostnih dokumentov, ki omogoča povezave znotraj dokumenta ali med dokumenti (Kanič, I. et al., 2020).

Arhiviranje spleta neizogibno vsebuje tako ali drugačno obliko selekcije, spletni arhivi pa bodo vedno vsebovali le izsek živega spleta v določenem času (Masanès, 2006).

Izkušnja uporabe svetovnega spleta izhaja iz interakcije dveh njegovih temeljnih komponent – spletnega strežnika in odjemalca, kot je na primer spletni brskalnik. Na strežniku je vsebina shranjena v obliki datotek, kot so HTML-dokumenti in slike. Strežnik na podlagi zahtevkov brskalnika slednjemu dostavi ustrezne datoteke, ta pa prejete datoteke prikaže uporabniku. Interakcija med strežnikom in brskalnikom je enako pomembna kot komponente. Praviloma za komunikacijo med strežnikom in brskalnikom skrbi protokol HTTP (Brown, 2006).

V kontekstu shranjevanja spletnih vsebin je pomembno upoštevati, da so posamezne spletne strani ali njihovi deli lahko statični ali dinamični, saj to pomembno vpliva na možnost njihove shranitve. Statično spletno mesto sestavljajo različne spletne strani, na vsako od katerih vodi povezava vsaj z ene druge spletne strani. Vsaka spletna stran je običajno sestavljena iz enega ali več posameznih elementov. Struktura spletne strani in njena besedilna vsebina je običajno vsebovana v HTML-datoteki, ki vsebuje povezave do drugih elementov (npr. slik) in drugih spletnih strani. Vsi elementi spletne strani so na strežniku shranjeni v hierarhični strukturi map in podmap, URL posameznega elementa pa opisuje lokacijo tega elementa v strukturi (Brown, 2006). Tovrstna spletna mesta zaradi enostavne izvedbe običajno ne predstavljajo večjih težav za shranjevanje.

Na dinamičnih spletnih mestih se posamezne spletne strani generirajo sproti iz manjših vsebinskih elementov. Postopek se lahko izvede na strežniku, ki ob prejemu zahtevka posamezne elemente sestavi v spletno stran in jo pošlje odjemalcu. Če se postopek izvede na strani odjemalca (brskalnika), ta s strežnika pridobi skript[8], ga požene in tako sestavi spletno stran iz preostalih pridobljenih datotek. Dinamično generirane vsebine so pogosto vsebovane v bazah podatkov, iz katerih je možno vsebino pridobiti le z iskalnimi poizvedbami, ki sprožijo postopek pridobitve vsebine iz baze in njen prikaz. Druge pogoste oblike dinamično generiranih vsebin so na primer vsebine, dostopne prek različnih multimedijskih predvajalnikov, spletna mesta, ki uporabnikom omogočajo objavljanje in urejanje vsebin v različnih vdelanih aplikacijah, vsebine, ki so pridobljene iz drugih virov in so vdelane v spletno stran (npr. zemljevidi Google maps), personalizirane spletne strani itd. Dinamično generirane vsebine so izvedene na različne načine in so vedno pogostejše, vsem pa je skupno, da predstavljajo težavo za shranjevanje. Če se skripti, ki poskrbijo za generiranje vsebin, ne prenesejo k odjemalcu, jih ni možno shraniti, niti ni možno z njimi pridobiti vsebine spletnega mesta. Pri vsebinah, ki se generirajo na strani odjemalca (brskalnika), je s spletnimi roboti sicer možno shraniti potrebne skripte, vendar jih roboti ne morejo poganjati, saj delujejo popolnoma drugače kot spletni brskalniki, ki so namensko izdelani za odpiranje in prikazovanje spletnih vsebin.

Dinamično generirane vsebine predstavljajo del globokega spleta (angl. *deep web*), ki se od površinskega (angl. *surface web*) razlikuje predvsem po tem, da ni dostopen spletnim robotom. Izraz globoki splet je prvič uporabil Michael K. Bergman, ko je ugotovil, da spletni iskalniki spregledajo veliko količino dinamično generiranih informacij. Tradicionalni spletni iskalniki ustvarjajo indekse spletnih vsebin s sledenjem povezavam in lahko zato odkrijejo le tiste spletne strani, ki so statične in povezane z drugimi spletnimi stranmi. Poleg dinamično

[8] Skript (angl. *script*) je program, napisan v skriptnem jeziku (Kanič, I. et al., 2020).

generiranih vsebin v globoki splet spadajo tudi vsebine, ki so plačljive ali dostopne samo z uporabniškim imenom in geslom. Bergman je leta 2001 ocenil, da je globoki splet od 400- do 550-krat večji od spleta, kot ga pojmujemo običajno, in vsebuje 40-krat večjo količino podatkov kot površinski splet (Bergman, 2001). Hatta (2020) ocenjuje, da se je globoki splet zaradi individualnih dogovorov, ki jih je Google sklenil z lastniki različnih baz podatkov za potrebe indeksacije vsebin v kontekstu, kot ga obravnava Bergman, v zadnjih 20 letih precej skrčil. To seveda ne vpliva bistveno na omejitve spletnih robotov in drugih orodij, ki so v uporabi za zajem in shranjevanje teh vsebin.

3 Pristopi k arhiviranju spleta

Ko govorimo o arhiviranju spleta v kontekstu ohranjanja dediščine, za kar skrbijo ustanove, kot so knjižnice in arhivi, imamo največkrat v mislih zajemanje vsebin z uporabo spletnih robotov. Gre za najbolj razširjen način arhiviranja spleta, s katerim je shranjenih največ podatkov in se tudi največkrat pojavlja v področni literaturi. Poleg tega obstajajo tudi drugi pristopi in metode, ki so bili razviti za različne potrebe shranjevanja spletnih vsebin. Ti pristopi in metode se med seboj razlikujejo glede na predvideni tip izvajalca (posameznik, dediščinska ustanova, lastnik spletnega mesta) in glede na to, ali se zajem izvaja prek strežnika ali odjemalca.

Za najbolj osnovne pristope arhiviranja, ki jih lahko uporabi vsak posameznik, nista potrebna posebno tehnično znanje in oprema. Ena od najbolj enostavnih metod je **izdelovanje statičnih posnetkov zaslona**. Rezultat je slika dela ali celotne spletne strani, ki je shranjena v slikovni ali PDF-datoteki[9]. Podobna metoda je **izdelovanje gibljivih posnetkov zaslona** oziroma ustvarjanje videoposnetkov dogajanja na zaslonu, kar lahko zajema snemanje premikanja po spletnem mestu, igranja spletnih iger ali predvajanega spletnega pretočnega videoposnetka. Za obe metodi je značilno, da posnetki verodostojno ponazarjajo videz spletne strani v določenem trenutku, vendar ne vsebujejo aktivnih hiperpovezav. Preprosta metoda je tudi **shranjevanje posameznih datotek s spleta**, kar lahko pomeni shranjevanje HTML-datotek, vključno s celotno pripadajočo kodo, ali shranjevanje drugih posameznih datotek, ki sestavljajo spletno mesto, kot so npr. slike, videoposnetek, zvok (Brügger, 2018). Tovrstne metode so večinoma uporabljene pri dokumentiranju za potrebe raziskovanja ali pri arhiviranju za osebne potrebe, manj pa so primerne za shranjevanje večjih količin podatkov.

Lastniki spletnih mest, ki imajo neposreden dostop do strežnikov, lahko uporabljajo tudi drugačne metode. Ena od teh je **transakcijsko arhiviranje**, s katerim je možno dokumentirati dejansko uporabo spletnega mesta, ne pa tudi shraniti njegove celotne vsebine. Metoda je bila razvita za primere, ko mora lastnik spletnega mesta dokazovati, da je bila na določen datum na njegovem spletnem mestu dostopna določena vsebina ali tudi, da je bila na določen datum obiskana. V nekaterih državah so namreč podjetja in organizacije zakonsko odgovorni za vsebino, ki jo objavljajo na spletu in morajo biti zmožni dokazati, kakšne so bile stare verzije njihovih spletnih mest (Masanès, 2006). Glavna omejitev tega pristopa s stališča trajnega ohranjanja spletne dediščine predstavlja shranjevanje zgolj tistih vsebin, ki so bile dejansko obiskane (Brown, 2006). Enostaven, vendar dokaj omejen pristop predstavlja tudi **neposredno kopiranje spletnega mesta s strežnika** brez odjemalca in uporabe protokola HTTP. Enako kot

[9] Primer zajema spletne strani z zaslonskim posnetkom: https://web.archive.org/web/20130202005324/http:/web.archive.org/screenshot/http:/blog.dshr.org/

pri transakcijskem arhiviranju je tudi ta pristop za potrebe trajnega ohranjanja spletnih vsebin, za kar skrbijo dediščinske ustanove, večinoma neustrezen, saj zahteva neposreden dostop do spletnega strežnika in posledično sodelovanje lastnika spletnega mesta (Brown, 2006).

V primerih nekaterih spletnih mest je možno večje količine gradiva z njih pridobiti in shraniti z uporabo **aplikacijskih programskih vmesnikov**[10] (angl. *application programming interface, API*). To metodo pogosto uporabljajo spletne platforme (predvsem družbena omrežja), ki neprestano ustvarjajo podatke, ti pa so lahko uporabni za različne analize. Z uporabo programskega vmesnika ni možno pridobiti vsebin, kot se prikažejo v spletnem brskalniku, pač pa zgolj posamezne elemente spletnih mest, kot so podatki iz uporabniških profilov, slike, všečki, podatki o geolokaciji itd. (Brügger, 2018). Če lastnik spletnega mesta razpolaga s programskimi vmesniki in zunanjim uporabnikom omogoči tovrstno pridobivanje podatkov iz njegove baze, je možno s to metodo zajeti in shraniti njeno vsebino ali vsaj del nje. Uporabnost tega postopka za trajno ohranjanje je omejena, saj z njim ni možno pridobiti izvornega uporabniškega vmesnika in njegovih funkcionalnosti, pač pa zgolj surove podatke. Še večja pomanjkljivost je, da je izvajalec arhiviranja odvisen od tega, ali lastnik spletnega mesta sploh omogoča tovrsten dostop do podatkov in če ga, kako dolgoročno zanesljiv je in kakšne funkcionalnosti vključuje (Laska, 2019).

Vsi opisani pristopi k arhiviranju spletnih vsebin so lahko uporabni le v specifičnih okoliščinah in za določene posameznike ali organizacije. Nekaterim je skupno, da terjajo neposreden dostop do strežnika, ki gosti spletno mesto, večini pa, da omogočajo shranjevanje relativno omejene količine podatkov ob precejšnjem časovnem vložku. Te omejitve pomenijo, da nobeden od pristopov ni ustrezen za sistematično arhiviranje večje količine spletnih vsebin za namene ohranitve nacionalne ali drugače opredeljene spletne dediščine. V nadaljevanju podrobneje opisujemo še zadnji pristop, ki sicer ne odpravlja vseh do zdaj opisanih pomanjkljivosti, vendar je glede na rezultate, ki jih omogoča, in načine, kako pridemo do njih, daleč najučinkovitejši in najprimernejši za arhiviranje spleta, ki ga izvajajo dediščinske ustanove. To je arhiviranje z uporabo **spletnih robotov**.

4 Arhiviranje z uporabo spletnih robotov

4.1 Kratka zgodovina arhiviranja spleta z uporabo robotov

Zajemanje z uporabo robotov je najbolj kompleksna in avtomatizirana ter hkrati najprimernejša metoda arhiviranja spleta za shranjevanje večjih količin podatkov. Namenski roboti[11] za tovrstno zajemanje delujejo enako kot programi, ki jih spletni iskalniki uporabljajo za indeksacijo spleta (Brügger, 2018). Začetek arhiviranja z uporabo spletnih robotov sega v leto 1996, ko se je po svetu zagnalo več tovrstnih projektov. Tega leta sta v ZDA dva izmed razvijalcev sistema za indeksacijo spleta WAIS[12], in sicer Brewster Kahle in Bruce Gilliat, ustanovila podjetje Alexa Internet. Podjetje je za namene beleženja spletnega prometa in

[10] Vmesnik, ki zagotavlja, da ima računalniški program na razpolago funkcije operacijskega sistema ali drugega računalniškega programa (Kanič, I. et al., 2020).

[11] Tovrstni programi spadajo v družino programske opreme, na splošno poimenovano roboti ali pajki. Roboti so programi, ki so primarno namenjeni zbiranju podatkov in do spletnih vsebin dostopajo na podoben način kot človeški uporabniki. Med najpogostejšimi roboti so orodja, ki jih spletni iskalniki, kot je Google, uporabljajo za zbiranje in indeksacijo spletnih strani (Brown, 2006).

[12] WAIS (Wide Area Information Server) je sistem za iskanje po besedilu, ki za iskanje po bazah podatkov na oddaljenih računalnikih uporablja ANSI-standard Z38:50 (The history of domains, 2020).

rangiranja spletnih mest slednja tudi shranjevalo (Milligan, 2019). Istega leta je Kahle z namenom vzpostavitve digitalne knjižnice, ki bi omogočala trajen dostop do historičnih zbirk v digitalni obliki, ustanovil tudi neprofitno organizacijo Internet Archive (Brown, 2006). Ta je v naslednjih letih na osnovi zajemov, ki jih je kot donacije prejemala od podjetja Alexa Internet, vzpostavila največji spletni arhiv na svetu (Mohr et al., 2004), ki deluje še danes.

Istega leta se je v okviru projekta Kulturaw3 z arhiviranjem nacionalne dediščine na spletu začela ukvarjati tudi Švedska nacionalna knjižnica (Kungliga biblioteket, KB); le-ta je v svoje zajeme uvrstila spletna mesta na vrhnjih domenah .se in .nu ter tista, ki so bila registrirana s švedskimi naslovi ali telefonskimi številkami (Milligan, 2019). Enega od zgodnejših projektov predstavlja tudi Pandora (Preserving and Accessing Networked Documentary Resources of Australia); ta projekt je leta 1996 začela izvajati Nacionalna knjižnica Avstralije. Pristop je bil zelo selektiven, za zajem so v poštev prišla spletna mesta avstralskih avtorjev in tista, ki so se vsebinsko nanašala na Avstralijo (Brown, 2006). V Sloveniji se je zajemanje spleta eksperimentalno začelo v letih od 2002 do 2004 v okviru projekta Metodologija zbiranja in arhiviranja slovenskih elektronskih publikacij na medmrežju, pri katerem sta sodelovala Narodna in univerzitetna knjižnica (NUK) in Institut Jožef Štefan (IJS) (NUK, 2025). Med projektom je NUK razvila teoretični model arhiviranja in trajnega ohranjanja spletnih publikacij, IJS pa je razvil spletni robot WebBird, s katerim je opravil poskusni zajem spletne domene .si (Kavčič-Čolić, 2004). NUK je zakonsko podlago za arhiviranje spleta dobila v letih 2006 in 2007 z novim *Zakonom o obveznem izvodu publikacij* (2009) in *Pravilnikom o vrstah in izboru elektronskih publikacij za obvezni izvod* (2007). Po vzpostavitvi ustrezne tehnične infrastrukture je s sistematičnim zajemanjem spletnih mest začela leta 2008.

Po letu 2000 se je krog pobud za arhiviranje spleta postopno razširil, predvsem v okviru nacionalnih knjižnic in nekaterih univerz. Razvite so bile različne aplikacije za zajem spletnih mest, ki so jih razvili arhivske ustanove za lastne potrebe ali drugi zainteresirani posamezniki. V tem obdobju se je pojavila potreba po zmogljivem in enostavno nastavljivem orodju, ki bo lahko izvajalo zajeme večjega obsega. Pobuda za razvoj tovrstnega orodja je prišla iz organizacije Internet Archive, ki se je pri tem želela povezati z drugimi ustanovami. Po opravljeni analizi tedanjih odprtokodnih orodij so ugotovili, da nobeno ni dovolj fleksibilno in hkrati zmožno izvajati večje zajeme. Tako so v sodelovanju z razvijalci nekaterih skandinavskih nacionalnih knjižnic, združenih v projekt Nordic Web Archive, v letih 2003 in 2004 razvili odprtokodno orodje Heritrix[13] (Mohr, 2004). Kmalu po tem so številne ustanove s prejšnjih rešitev prešle na Heritrix, ki je bil od svojih začetkov večkrat posodobljen in je še danes najbolj razširjen robot za zajem spletnih mest.

4.2 Delovanje robotov

Infrastruktura, ki je potrebna za delovanje spletnega robota, je lahko zelo osnovna. Aplikacijo je treba namestiti na računalnik z internetno povezavo in zadostnimi pomnilniškimi kapacitetami za shranjevanje zajetih podatkov. Pri večini večjih projektov arhiviranja spleta so aplikacije za zajem nameščene na mrežnih strežnikih s pripadajočimi diskovnimi polji (Brown, 2006). Pri interakciji s strežnikom robot posnema uporabo spleta običajnih uporabnikov. Strežnikom pošilja HTTP-zahtevke in shranjuje vsebino, ki jo tako pridobi. Delovanje robota

[13] Celotna aplikacija z vso pripadajočo dokumentacijo je dostopna na naslovu https://github.com/internetarchive/heritrix3

usmerja seznam izhodiščnih URL-naslovov (semen[14]), ki naj bi jih obiskal. Robot obišče prvi URL-naslov na seznamu, shrani spletno stran, ki je na tem naslovu, znotraj nje identificira hiperpovezave in jih doda na seznam URL-jev za nadaljnje zajemanje. Potem postopek ponavlja s pridobljenimi dokumenti, dokler mu ne zmanjka povezav, ki jih lahko obišče in pridobi glede na nastavitve obsega zajema. Tako lahko robot, ki zajem začne na domači strani določenega spletnega mesta, obišče in zajame vse spletne strani znotraj spletnega mesta. S tem postopkom je možno shraniti vsebino na posameznih URL-naslovih, še bolj pogosto pa se ga uporablja za zajem vsebine celotnih spletnih mest ali spletnih domen (Brown, 2006; Masanès, 2006; Brügger, 2018).

Delovanje robotov je možno nastaviti glede na številne parametre. Pomemben skupek nastavitev predstavlja določitev obsega in globine zajema. Te nastavitve so pomembne, ker robotu preprečijo preširoko zajemanje, ki se lahko pojavi zaradi medsebojne povezanosti spletnih mest. Obseg zajema je v osnovi določen s številom povezav, ki jim robot sledi od izhodiščnega URL-naslova, in smerjo, v kateri naj se premika (Brown, 2006). Glede na namen zajemanja in razpoložljiva sredstva so lahko zajemi globlji ali plitkejši po vertikalni dimenziji ter širši ali ožji po horizontalni dimenziji. V tem kontekstu nekateri avtorji navajajo ekstenzivno in intenzivno ali široko in fokusirano zajemanje (glej Mohr, 2004 in Masanès, 2006). Čim večja globina je zaželena, kadar je namen zajeti spletna mesta v celoti in čim bolj popolno. Tovrstni zajemi običajno shranijo veliko količino podatkov, kar je lahko problematično, če ima arhivska ustanova na voljo omejena sredstva in čas. Plitkejši zajemi so pogostejši pri tematskih zajemih, kjer cilj ni zajeti celotnih spletnih mest, pač pa le posamezne vsebine, ki se nanašajo na določeno tematiko. V takih primerih so pogosto zajete le posamezne spletne strani. Manj globoki zajemi pridejo v poštev tudi v primerih zajemov celotnih nacionalnih domen ali podobno obsežnih delov spleta. Ker je pri tem cilj zajeti veliko količino spletnih domen, ima širina zajema prednost pred globino. Na splošno so ožji zajemi pogosto globlji, širši pa plitkejši.

Čeprav je delovanje robota omejeno že z določitvijo izhodiščnih URL-naslovov in nastavitvijo globine zajema, je obseg zajema običajno omejen tudi glede maksimalne količine zajetega gradiva (v bajtih), števila zajetih datotek ali trajanja. S tem arhivska ustanova poskrbi za racionalno porabo sredstev, ki jih ima na voljo, hkrati pa prepreči tudi negativne posledice morebitnih pasti, v katere se lahko zaplete robot. Pasti so segmenti spletnih mest (npr. koledarji), ki lahko generirajo neomejeno število povezav, katerim mora robot slediti in lahko povzročijo, da se zajem nikoli ne zaključi. Tovrstnim nevarnostim se je možno vsaj do neke mere izogniti tudi z bolj specifičnimi nastavitvami, s pomočjo katerih lahko izvajalec iz zajema izloči posamezne direktorije spletnih mest, določene tipe datotek ali URL naslove, ki vsebujejo določen niz znakov (Brown, 2006). Po drugi strani je možno v zajem vključiti tudi vsebine, ki se ne nahajajo na domeni izhodiščnega URL naslova. Kadar celotna vsebina spletnega mesta ni dostopna na eni domeni, pač pa se nekateri elementi prenesejo z drugih domen, je možno robotu določiti, da sledi določenemu številu povezav, ki z zajetega spletnega mesta vodijo na druga spletna mesta. S tem se zagotovi zajem tudi tovrstnih zunanjih vsebin, če predstavljajo del vsebinske celote spletnega mesta.

[14] V angleščini je za tovrstne URL-naslove v uporabi izraz *seeds* ali *seed URLs*, kar v slovenščino lahko prevedemo kot *izhodiščni URL-naslovi* ali *semenski URL-naslovi*. S pomočjo teh URL-jev robotu določimo, kje naj začne zajem in do neke mere tudi območje, znotraj katerega naj zajame vsebino (npr. posamezna spletna domena, poddomena, direktorij ali zgolj posamezna spletna stran).

Spletna mesta so lahko zajeta samo enkrat, še pogosteje pa jih arhivske ustanove zajemajo kontinuirano. Robotu je možno določiti urnik zajemanja in datum, ko naj začne zajemati gradivo z določene lokacije. Frekvenca zajema je lahko dnevna, tedenska, mesečna, letna ali kakršna koli druga, odvisna pa je predvsem od pogostosti spreminjanja vsebin spletnih mest. Po zaključku vsakega zajema je zajeto gradivo lahko shranjeno v enaki drevesni strukturi kot izvorno spletno mesto, kjer so v posameznih mapah in podmapah različne datoteke, še pogosteje pa se ga shranjuje v standardiziranem arhivskem formatu WARC ali sodobnejšem formatu WACZ, ki omogoča shranjevanje različnih virov (digitalnih objektov), sestavljenih iz podatkov in metapodatkov, v eno datoteko (ISO 28500, 2017).

4.3 Omejitve zajemanja z roboti

Zaradi tehnične kompleksnosti, prepletenosti in obsežnosti spleta noben pristop k zajemanju ne predstavlja popolne rešitve, ki bi omogočala učinkovito shranitev vseh oblik spletnih vsebin, vendar je zajemanje z roboti glede na druge načine najbolj enostavno, fleksibilno in učinkovito. Postopke je možno vzpostaviti hitro in z relativno enostavno infrastrukturo, pristop pa omogoča zajem velikega števila spletnih mest v kratkem času. Metoda je dodobra preizkušena, na voljo pa so tudi številna orodja, ki jih lahko ob ustreznem uvajanju uporabljajo tudi tehnično manj podkovani uporabniki. Ker pri zajemih ni potrebna participacija lastnikov spletnih mest, ima ustanova, ki izvaja arhiviranje, popoln nadzor nad vsemi postopki (Brown, 2006).

Vsak pristop k arhiviranju spleta vsebuje omejitve in zajemanje z roboti pri tem ni izjema. Omejitve so predvsem tehnične narave in so pogojene s številnimi možnostmi, ki so na voljo razvijalcem in oblikovalcem sodobnih spletnih mest. V začetnem obdobju razvoja svetovnega spleta so bila spletna mesta in interakcije uporabnikov z njimi relativno enostavna. Temu ustrezno je bil enostaven tudi klasičen model zajemanja z roboti, ki je bil osnovan na standardih HTTP in HTML ter je primarno vključeval razčlenjevanje (angl. *parsing*) besedila in identifikacijo ter sledenje povezavam. Takrat je bilo možno z relativno preprostimi metodami v doglednem času in z zmernimi stroški shraniti verodostojne reprezentacije spletnih mest. Nekatere omejitve zajemanja z roboti so obstajale že v tistem obdobju, še več pa se jih je pojavilo z razvojem spletnih tehnologij v zadnjih dveh desetletjih. Te tehnologije so pripeljale do družbenih omrežij in spletnih tehnologij, ki uporabnikom omogočajo ustvarjanje in objavljanje lastnih vsebin ter spletnih mest kot interaktivnih, prepletenih storitev, ki niso več zgolj zbirke datotek v drevesni strukturi, pač pa bolj spominjajo na aplikacije (IIPC, 2012).

Vsebine, pri katerih se kažejo najpomembnejše omejitve klasičnih robotov, lahko združimo v naslednje kategorije:

- vsebine, ki so v bazah podatkov in jih je možno pridobiti le z iskalnimi poizvedbami;
- multimedijske vsebine, dostopne v predvajalnikih in prikazovalnikih;
- vsebine, zaščitene z gesli, še posebej, če se URL za prijavo generira dinamično;
- vsebine, ki so dostopne prek menijev ali drugih elementov, izdelanih s programskim jezikom JavaScript;
- dinamično generirani, spreminjajoči se URL-naslovi, ki postrežejo isto vsebino;
- mehanizmi za postopno nalaganje spletnih strani, ki ne zahtevajo hkratnega prenosa vseh virov, uporabljenih za prikaz ali funkcioniranje spletne strani na strani odjemalca;

- mehanizmi za filtriranje prikazanih vsebin, ki temeljijo na izpolnjevanju obrazcev;
- skripti, ki se ne prenesejo k odjemalcu, pač pa se izvajajo na strani strežnika (Kavčič-Čolić, 2011; International Internet Preservation Consortium, 2012; Pennock, 2013).

Navedene težave lahko povzročijo, da se določene vsebine ne zajamejo ali pa se zajamejo, vendar se ne prikažejo na ustrezen način ali se sploh ne prikažejo, ko uporabnik odpre arhivirano verzijo spletnega mesta. Zaradi tovrstnih izzivov se v zadnjih letih pojavljajo rešitve, s katerimi je možno nadgraditi delovanje robotov in vsaj do neke mere izboljšati kakovost zajemov (Sigurðsson, 2016b; Besser, 2017). Ker spletna mesta primarno niso ustvarjena z namenom, da bi jih zajemali roboti, pač pa zato, da bi jih uporabniki odpirali in uporabljali v spletnih brskalnikih, je ena od smiselnih rešitev vključevanje brskalnikov v postopek zajema.

Primer orodja, ki za zajem uporablja spletni brskalnik, je Brozzler[15], razvit pri organizaciji Internet Archive. Brozzler za pridobitev spletnih strani in iskanje povezav znotraj njih uporablja spletni brskalnik brez uporabniškega vmesnika (angl. *headless browser*) Chromium. Namesto običajnega sledenja povezavam in shranjevanja njihove vsebine, kot je to značilno za klasične robote, Brozzler pred shranitvijo naloži vse vsebine spletnega mesta v spletnem brskalniku in tako bolje kot robot posnema interakcijo uporabnika s spletnim mestom, npr. drsenje, klikanje, uporaba kontrolnih gumbov (Besser 2017; Rollason-Cass, 2022). Podobno orodje je Browsertrix, del paketa orodij, razvitih v okviru projekta Webrecorder, pri katerem sodelujejo različne knjižnice in druge ustanove. Browsertrix poleg orodja za zajem vsebuje tudi uporabniški vmesnik za določanje urnikov zajemov in pregledovanje njihove kakovosti ter s tem predstavlja celostno rešitev za arhiviranje spletnih mest[16].

Uporaba brskalnikov predstavlja bistven napredek pri tehnikah zajemanja sodobnega spleta, verjetno pa ta način ne bo nikoli povsem zamenjal zajemanja z roboti, saj so ti še vedno najprimernejši za izvajanje obsežnejših zajemov. V prihodnosti bomo verjetno priča uporabi različnih kombinacij obeh pristopov (in morda še kakšnega novejšega). Hkrati je treba poudariti, da tudi te naprednejše tehnike ne odpravljajo vseh težav, ki jih predstavlja arhiviranje sodobnega spleta. Ker se spletne tehnologije neprestano razvijajo, bodo metode arhiviranja spleta verjetno vedno vsaj kakšen korak za časom in z njimi nikoli ne bo možno povsem ustrezno zajeti vseh oblik spletnih vsebin. Nekatere pomanjkljivosti so zaradi fluidne in kompleksne narave svetovnega spleta skupne prav vsem pristopom ter predstavljajo neizogibne lastnosti vsakega spletnega arhiva.

5 Splošne omejitve arhiviranja spleta

5.1 Časovna neskladja

Časovna skladnost v kontekstu arhiviranja spleta je lastnost določenega skupka arhiviranih spletnih strani, ki nakazuje, da so bile ob določenem času vse prisotne na živem spletu (Ball, 2010). Časovna neskladnost je posledica hipertekstovne narave svetovnega spleta in se lahko pojavi pri vsem gradivu, ki vsebuje hiperpovezave. Časovni soobstoj vira in cilja hiperpovezave se prekine, če je cilj zajet ure ali celo dneve kasneje kot vir, kar se dogaja relativno pogosto (Brügger, 2018). Časovno neskladje se pojavi, ko se v času, ki ga robot porabi za zajem

[15] Ime orodja Brozzler je kombinacija angleških besed *browser* (brskalnik) in *crawler* (spletni robot) ter s tem nakazuje, da gre za pristop, ki zajemanje z roboti dopolnjuje z uporabo spletnega brskalnika.
Več o orodju Brozzler: https://github.com/internetarchive/brozzler
[16] Več o projektu Webrecorder in orodju Browsertrix: https://webrecorder.net/

spletnega mesta, določeni deli spletnega mesta spremenijo, vsebina na vrhnjih nivojih izhodiščnega URL-ja (npr. na domači strani) pa se zato časovno več ne ujema z vsebino na globljih nivojih. Tovrstna težava se lahko pojavi že pri manjših spletnih mestih, še bolj pereča pa je pri obsežnejših zajemih (npr. vrhnje nacionalne domene), ki lahko glede na obseg trajajo več tednov ali mesecev. Arhivirane vsebine ne moremo smatrati kot reprezentativno kopijo živega spleta (ali spletnega mesta) v določenem trenutku, pač pa le v določenem časovnem razponu (Pennock, 2013). Rezultat zajema spletnega mesta je lahko njegova kopija, ki morda v taki različici nikoli ni bila dostopna na živem spletu in je povsem unikatna. Glede na vstopno točko (običajno domača stran) lahko zajeta različica vsebuje posamezne spletne strani, ki ob času zajema vstopne točke še niso obstajale in so bile objavljene med zajemom, hkrati pa ne vsebuje spletnih strani, ki so ob zajemu vstopne točke obstajale in so bile med zajemom umaknjene s spletnega mesta, še preden jih je dosegel robot. Brügger (2018) navaja, da so časovna neskladja možna pri vseh spletnih arhivih, hkrati pa je zelo težko oceniti, kje in kako obsežna so.

5.2 Omejitve, ki jih ustvarjajo ustvarjalci spletnih vsebin

Globoki splet je svoje ime dobil ravno zaradi svoje nedosegljivosti spletnim robotom. Posamezne vsebine globokega spleta je sicer možno shraniti, vendar so postopki kompleksni, časovno potratni in zahtevajo veliko ročnega dela. Kakršna koli avtomatizirana rešitev je lahko razvita zgolj za vsak posamezen primer, univerzalne rešitve niso možne. Uporabnik zato lahko pričakuje, da v spletnih arhivih ne bo številnih vsebin, ki so na spletu shranjene v bazah podatkov, so dosegljive samo prek iskalnikov ali dostopne le ob registraciji.

Tovrstne omejitve niso nastale z namenom omejevanja delovanja robotov, pač pa so tehnične in praktične narave. Ustvarjalci spletnih mest pa lahko delovanje robotov tudi namerno omejujejo. Pogost in uveljavljen način je uporaba datoteke robots.txt, ki se nahaja v vrhnjem direktoriju vsakega spletnega mesta in robotom na osnovi standardnega protokola določa, katere dele spletnega mesta lahko obiščejo in katerih ne, ali celo, da spletnega mesta ne smejo obiskati. Upoštevanje pravil v robots.txt je del pravil dobrega vedenja na spletu, vendar ni obvezno (Sigurðsson, 2016a). Glede na zakonsko podlago, namen zajemanja in politiko arhivske ustanove lahko robot pravila upošteva ali ignorira. Upoštevanje robots.txt je lahko dvorezen meč. Robot ne bo zajel vsega relevantnega gradiva ali celo ne bo zajel nič gradiva, če se je lastnik spletnega mesta tako odločil. Po drugi strani je z upoštevanjem navodil poskrbljeno za vljudnost zajema, robots.txt pa lahko vsebuje tudi koristne izključitve, ki robotu preprečijo nepotrebno zajemanje nevsebinskih delov spletnega mesta.

Drug razširjen način omejevanja delovanja robotov so različni testi za ugotavljanje, da je obiskovalec spletnega mesta res človek (npr. Captcha). Ti se uporabljajo predvsem za preprečevanje zlorab spletnih mest s strani zlonamernih robotov in vključujejo različne vrste preizkusov (besedilnih, zvočnih, slikovnih, itd.), ki jih lahko opravi le človek (Guerar et al., 2022). Čeprav se neprestano razvija tudi zlonamerna programska oprema, ki lahko zaobide enostavnejše načine preverjanja, ti mehanizmi učinkovito preprečujejo avtomatizirano shranjevanje spletnih vsebin z roboti.

5.3 Omejitve, ki jih ustvarjajo izvajalci zajemov

Vsak zajem je praviloma namenoma zamejen glede na količino zajetega gradiva, število zajetih datotek ali čas zajema; kateri omejevalni parameter je izbran, pa je odvisno od namena zajema in politike arhivske ustanove. Tovrstne zamejitve izvirajo predvsem iz racionalizacije izrabe razpoložljivih sredstev, saj nobena ustanova nima na voljo neomejenega števila zaposlenih, časa in prostorskih kapacitet. Zato so potrebni kompromisi, s katerimi je doseženo optimalno razmerje med razpoložljivimi sredstvi arhivske ustanove in težnji k ohranitvi čim več relevantnih vsebin. To pomeni, da čeprav je namen zajeti določeno spletno mesto v celoti, lahko nekatere vsebine izpadejo iz zajema, ker je robot na določeni točki zaradi nastavljenih omejitev zaključil svoje delo.

Kot smo že navedli, je omejitev zajema izražena tudi z globino, ki je odvisna od namena zajemanja v povezavi s sredstvi, ki so na voljo. Pri tematskih zajemih so na primer pogosto zajete le posamezne spletne strani ali segmenti spletišč, ki obravnavajo določeno tematiko. Uporabnik ima tako na voljo le manjše koščke spletnih mest, ki so jih za zajem izbrali kuratorji zbirke. Ker je izbira vsebin za zajem vedno vsaj do neke mere subjektivna in ker je splet tako obsežen, je v zajem nemogoče vključiti čisto vse vsebine, ki so relevantne za določeno temo. Zato so takšne zbirke same po sebi nepopolne. Uporabnikovo zavedanje, kaj zbirka vsebuje in česa ne, pa je odvisno od tega, kako podrobno je arhivska ustanova dokumentirala kriterije in različne druge odločitve, ki so oblikovale vsebino zajema. Tovrstna dokumentacija, če sploh obstaja, uporabniku pogosto ni na voljo.

Vsako spletno mesto na živem spletu se lahko uporabniku prikaže na različne načine. Prikaz je lahko odvisen od spletnega brskalnika, geografske lokacije uporabnika ali drugih nastavitev, ki se v obliki piškotkov shranijo na uporabnikov računalnik. Različne verzije istih vsebin ustvarjajo tudi arhivisti. Različne tehnike zajemanja lahko povzročijo, da se kopiji spletnega mesta, ki sta ju ustvarili dve različni arhivski ustanovi, med seboj razlikujeta, nobena od njiju pa ni identična spletnemu mestu, kot je obstajalo v času zajema. Zbirka arhiviranega spleta je v bistvu zbirka različnih verzij, od katerih vsaka predstavlja unikatno rekonstrukcijo izvirnega spletnega mesta, ki verjetno ne obstaja več. Vsaka od verzij je lahko le ena od mnogih in zelo težko je ugotoviti, ali je katera od njih identična izvornemu spletnemu mestu. Prav tako je na podlagi vseh zajetih različic težko ugotoviti, kakšno je bilo izvorno spletno mesto. Arhivirana spletna mesta je zato treba obravnavati kot unikatne različice, ne kot kopije živih spletnih mest. Unikatnost zajetih spletnih mest je drugačna kot na primer pri digitaliziranem gradivu, kjer so posamezni digitalizirani objekti s stališča vsebinske popolnosti veliko bližje ali celo popolnoma identični izvirniku (Brügger, 2018).

6 Nepopolnosti, specifične za spletne arhive

Nepopolnost je sestavni del vsake zbirke, vključno z zbirkami digitaliziranega in izvorno digitalnega gradiva, vendar je ta pri spletnih arhivih drugačna. V primeru zbirk digitaliziranega gradiva se nepopolnost lahko pojavi, ker je bila že izvorna fizična zbirka ali primerek nepopoln, pomanjkljivost pa lahko ugotovimo še pred digitalizacijo. Določene sistemske in predvidljive nepopolnosti se lahko pojavijo tudi s postopkom digitalizacije. Ena od bistvenih razlik med arhiviranim spletnim gradivom in digitaliziranim gradivom je, da zaradi neobstojnosti spletnih vsebin uporabnik pri uporabi arhiviranega spletnega gradiva pogosto nima možnosti, da bi preveril, kakšen je bil original. Slednji se je morda spremenil ali pa je povsem izginil s spleta. Po drugi strani pri digitaliziranem gradivu zelo pogosto še vedno obstaja fizični izvirnik, s katerim po potrebi primerjamo digitalizirano verzijo. K netransparentnosti prispeva tudi

kompleksnost postopka zajema, ki je lahko pogojen s številnimi odločitvami in specifičnimi nastavitvami uporabljenih orodij. Te modalitete so redko podrobno dokumentirane, zato je težko ugotoviti, do kakšnih nepopolnosti so pripeljale (Brügger, 2018).

Nepopolnost pogosto predstavlja nerešljivo težavo tudi za izvajalca arhiviranja. Bolj je postopek kompleksen, večja je možnost za različne človeške ali tehnične napake, ki jih je težko identificirati. V idealni situaciji bi vsak zajem po zaključku podrobno pregledal operater, ki bi po potrebi zajem ponovil z drugačnimi nastavitvami in tako dosegel boljšo kakovost zajetega gradiva. V praksi je to zaradi velike količine gradiva nemogoče. Čeprav je postopke ocenjevanja kakovosti do neke mere mogoče avtomatizirati, je neizogibno, da se bodo v zajetem gradivu pojavljale tudi napake in pomanjkljivosti, ki bi se jih dalo odpraviti, če bi bile zaznane. Hitrost spreminjanja spletnih vsebin in njihovega izginjanja še dodatno poslabšuje možnosti za ponovni, izboljšan zajem prvotno neuspešno pridobljenih vsebin.

Najbolj bistvena razlika med spletnimi arhivi in drugimi vrstami zbirk izhaja iz transformativne narave shranjenega gradiva. Kot ugotavlja Masanès (2006), spletno okolje omogoča neprestano spreminjanje, posodabljanje in brisanje vsebin, zaradi česar splet ni stanoviten informacijski prostor, pač pa dinamičen preplet različnih informacijskih sistemov in vsebin. Spletni arhivisti morajo zato gradivo, ki ga zbirajo, izločiti iz tega neprestano spreminjajočega se okolja in poskrbeti za njegovo odpornost proti spremembam, značilnim za svetovni splet. Posledica ločitve gradiva od njegovega izvora (strežnika) je lahko izguba nekaterih funkcionalnosti, ki jih zagotavlja izvorno okolje. Ker so spletne vsebine med postopki zajema in shranitve podvržene različnim spremembam, je ohranitev vseh značilnosti in funkcionalnosti izvornih digitalnih objektov pogosto nemogoča. Brügger (2018) vsebino spletnih arhivov označuje kot digitalno transformirano oziroma preobraženo gradivo in s tem uvaja novo kategorijo digitalnega gradiva, ki se po svoji naravi razlikuje tako od digitaliziranega (angl. *digitized*) kot od izvorno digitalnega gradiva (angl. *born-digital*). Uporabnik spletnih arhivov lahko pričakuje odsotnost posameznih elementov, kot so slike, videoposnetki in različne oblike interaktivnosti, ali celo manjkajoče spletne strani in celotna spletna mesta.

K temu je treba prišteti tudi selektivno naravo spletnih arhivov, katerih vsebina je periodično zajeta zgolj z izbranih spletnih lokacij. Pri izboru sodelujejo posamezniki z različnimi stopnjami poznavanja spleta, izkušnjami in profili znanj, postopek pa usmerjajo različni zakonski okviri in institucionalne politike. Arhiviranje spleta zato pogosto bolj spominja na vzorčenje kot pa na celosten pristop k ohranjanju dediščine, saj je slednji zaradi obsežnosti, minljivosti in tehničnih značilnosti spleta v praksi neizvedljiv, spletni arhivi pa, kot navaja Hofheinz (2010), pogosto vsebujejo več vrzeli kot vsebine. Kot ugotavlja Brügger (2018), nepopolnost sama po sebi ni posebnost spletnih arhivov, pač pa se ti od drugih zbirk razlikujejo predvsem po tem, da je zelo težko ugotoviti, do kakšne mere so nepopolni, kateri deli vsebine manjkajo in zakaj.

7 Zaključek

Ker se je svetovni splet od svojih začetkov razvil v nepogrešljivo globalno komunikacijsko orodje, ki vsebuje veliko količino raznovrstnega znanja in informacij, so prizadevanja za ohranjanje teh vsebin prav tako nujna kot pri drugih digitalnih in fizičnih virih pisne dediščine človeštva. Ena od bistvenih značilnosti svetovnega spleta v primerjavi z drugimi viri informacij je neobstojnost njegove vsebine, kar pomeni, da bi odlašanje s prizadevanji za njeno ohranitev pripeljalo do velikih izgub naše skupne dediščine. Arhiviranje spleta se je na srečo začelo kmalu

po njegovem nastanku, vendar se dediščinske ustanove že od začetka soočajo s številnimi izzivi pri ohranitvi vsebine tega dinamičnega, tehnično kompleksnega in izjemno obsežnega prostora.

Pristopi k shranjevanju spletnih vsebin se med seboj razlikujejo po kompleksnosti in tehnični dovršenosti, nobeden od njih pa ne zagotavlja celostne shranitve vseh spletnih vsebin. Zajemanje z uporabo spletnih robotov, ki predstavlja najbolj razširjen pristop, je praktično in uporabno, vendar omejeno glede tega, kako daleč v tkivo spleta lahko prodrejo roboti in katere vrste vsebin lahko dosežejo. Kljub nenehnemu razvoju tehničnih rešitev za zajemanje bo velik del vsebine svetovnega spleta tudi v prihodnje ostal neshranjen, spletni arhivi pa bodo verjetno vedno nepopolni. Kot ugotavlja Brügger (2018), je zbirka arhiviranega spletnega gradiva zbirka koščkov preteklega živega spleta, vendar je težko z gotovostjo ugotoviti, koliko relevantnih koščkov je prisotnih v zbirki, ali so vsi iz istega časa in ali so tisti, ki bi morali biti, med seboj povezani. Zaradi pogoste odsotnosti originala in unikatnosti vsake arhivirane različice so spletni arhivi konstitutivno nezanesljivi v smislu zrcaljenja živega, izvorno digitalnega spleta. K temu poleg tehničnih okoliščin prispevajo tudi različni človeški in družbeni dejavniki, ki tako na strani ustvarjalcev spletnih mest kot na strani arhivskih ustanov ustvarjajo omejitve, ki se jim ni mogoče izogniti. Kot navaja Ben-David (2021), je treba pri vrednotenju vsebinske (ne)celovitosti spletnih arhivov poleg tehničnih okoliščin, ki jih oblikujejo, upoštevati tudi družbena okolja, v katerih arhivi nastajajo, ter vrednotne in ideološke pristranskosti, ki jih vsebujejo.

Tovrstni razmisleki se dotikajo epistemoloških vprašanj o tem, kakšno znanje vsebujejo spletni arhivi in kako vplivajo na možne načine poznavanja, raziskovanja in razumevanja preteklega spleta. Ker so spletni arhivi nepredvidljivi in netransparentni viri podatkov, je pomembno, da poleg spletnih arhivistov njihove značilnosti dobro poznajo tudi njihovi uporabniki. To še posebej velja za tiste, ki arhivirane podatke uporabljajo za raziskovalno delo, saj morajo pri oceni verodostojnosti rezultatov takšnih raziskav razpolagati s čim boljšimi informacijami o različnih nepopolnostih uporabljenih podatkov. Pri zagotavljanju tovrstnih informacij zagotovo najbolj ključno vlogo igrajo arhivske ustanove z različnimi oblikami izobraževanj in usposabljanj ter z drugimi prizadevanji za ozaveščanje različnih skupin uporabnikov in širše javnosti o pomenu trajnega ohranjanja spletne dediščine ter o številnih izzivih in priložnostih, ki jih predstavlja arhivirani splet.

Reference

Atelšek, S. et al., 2024. Spletni arhiv. *Mnenje Terminološke svetovalnice pri ZRC SAZU.* Dostopno na: https://isjfr.zrc-sazu.si/sl/terminologisce/svetovanje/spletni-arhiv [25. 1. 2025].

Ball, A., 2010. *DCC State of the art report: web archiving*. Edinburgh: University of Edinburgh; UKOLN, University of Bath; HATII, University of Glasgow; Science and Technology Facilities Council.

Barone, F., Zeitlyn, D. in Mayer-Schönberger, V., 2015. Learning from failure: the case of the disappearing web site. *First Monday*, *20*(5–4). Dostopno na: https://doi.org/10.5210/fm.v20i5.5852 [25. 1. 2025].

Ben-David, A., 2021. Critical web archive research. V D. Gomes, Demidova E., Winters J. in Risse T. (ur.), *The Past Web* (str. 181–188). Springer Cham.

Bergman, M. K., 2001. White paper: The deep web: surfacing hidden value. *Journal of Electronic Publishing*, 7(1). DOI: doi:10.3998/3336451.0007.104

Besser, H., 2017. Archiving websites containing streaming media. V *Archiving 2017 – Final Program and Proceedings*, *2017–May*, str. 11–13. Dostopno na: https://doi.org/10.2352/issn.2168-3204.2017.1.0.11 [25. 1. 2025]

Brown, A., 2006. *Archiving websites: a practical guide for information management professionals*. London: Facet.

Brügger, N. in Laursen, D., 2019. *The historical web and digital humanities : the case of national web domains*. London [i. e.] Abingdon; New York: Routledge, Taylor & Francis Group.

Brügger, N., 2018. *The archived web*. London: The MIT Press.

Gomes D., 2021. Part 1. The era of information abundance and memory scarcity. V D. Gomes, Demidova E., Winters J. in Risse T. (ur.), *The past web* (str. 1–3). Springer.

Guerar, M., Verderame, L., Migliardi, M., Palmieri, F. in Merlo, A., 2022. Gota CAPTCHA 'em all: a survey of 20 years of the human-or-computer dilemma. *ACM Computing Surveys*, *54*(9), str. 1–33. Dostopno na: https://doi.org/10.1145/3477142 [25. 1. 2025]

Hatta, M., 2020. Deep web, dark web, dark net: A taxonomy of "hidden" internet. *Annals of Business Administrative Science*, 19 (2020), str. 277–292.

Hofheinz, A., 2010. A History of Allah.com. V N. Brügger (ur.) *Web history* (str. 105–135). New York: Peter Lang.

International Internet Preservation Consortium, 2012. *IIPC future of the web workshop – introduction* & overview. International Internet Preservation Consortium. Dostopno na: https://digital.library.unt.edu/ark:/67531/metadc1638392/ [25. 1. 2025]

International Organization for Standardization, 2017. Information and documentation — WARC file format (ISO Standard No. 28500:2017). Dostopno na: https://www.iso.org/standard/68004.html [25. 1. 2025]

Kanič, I. (ur.) et al., 2020. *Islovar*. Ljubljana: Slovensko društvo Informatika. Dostopno na: http://islovar.org/ [25. 1. 2025].

Kanič, I., Leder, Z., Ujčič, M., Vilar, P. in Vodeb, G., 2011. *Bibliotekarski terminološki slovar*. Ljubljana: Amebis.

Kavčič-Čolić, A. in Grobelnik, M., 2004. Archiving the slovenian web: recent experiences. V *4th International Web Archiving Workshop*. Dostopno na: https://www.researchgate.net/publication/228413950_Archiving_the_Slovenian_Web_Recent_Experiences [25. 1. 2025]

Kavčič-Čolić, A. in Klasinc, J., 2011. Arhiviranje spletnih strani v Narodni in univerzitetni knjižnici. *Knjižnica*, *55*(1), str. 209–232.

Laska, K., 2019. *The pros and cons of using APIs for web archiving* (6. 5. 2019). Hanzo. Dostopno na: https://www.jdsupra.com/legalnews/the-pros-and-cons-of-using-apis-for-web-70814/ [25. 1. 2025]

Masanès, J., 2006. *Web archiving*. Berlin; Heidelberg; New York: Springer.

Major, D., 2021. The problem of web ephemera. V D. Gomes, Demidova E., Winters J. in Risse T. (ur.), *The past web* (str. 5–10). Springer.

Milligan, I., 2019. *History in the age of abundance*? Montreal & Kingston; London; Chicago: McGill-Queen's University Press.

Mohr, G., Stack M., Ranitovic I., Avery D. in Kimpton, M., 2004. An introduction to heritrix : an open source archival quality web crawler. V *4th International Web Archiving Workshop*. Dostopno na: http://crawler.archive.org/Mohr-et-al-2004.pdf [25. 1. 2025]

NUK, 2025. *Zaključeni raziskovalni projekti*. Dostopno na: https://www.nuk.uni-lj.si/nuk/raziskovalna-dejavnost-zakljuceni [25. 1. 2025].

Pennock, M., 2013. *Web-archiving: DPC technology watch report 13-01 March 2013*. Digital Preservation Coalition. Dostopno na: https://www.dpconline.org/docs/dpc-technology-watch-publications/technology-watch-reports-1/865-dpctw13-01-pdf/file [25. 1. 2025].

Pravilnik o vrstah in izboru elektronskih publikacij za obvezni izvod, 2007. Uradni list RS, št. 90/07. Dostopno na: http://www.pisrs.si/Pis.web/pregledPredpisa?id=PRAV8482 [25. 1. 2025].

Rollason-Cass, S., 2022. *What is Brozzler?* Dostopno na: https://support.archive-it.org/hc/en-us/articles/360000343186-What-is-Brozzler- [25. 1. 2025].

Sigurðsson, K., 2016a. *3 things I shouldn't have to tell you about running a "good" crawler*. (24. 2. 2016). Dostopno na: https://kris-sigur.blogspot.com/2016/02/3-things-i-shouldnt-have-to-tell-you.html [25. 1. 2025].

Sigurðsson, K., 2016b. *3 crawlers: 1 writer*. (26. 9. 2016). Dostopno na: https://kris-sigur.blogspot.com/2016/09/3-crawlers-1-writer.html [25. 1. 2025].

The history of domains, 2020. Dostopno na: https://www.historyofdomains.com/wais/ [25. 1. 2025].

Weiss, R., 2003. On the web, research work proves ephemeral. *Washington Post,* 24. 11. 2003. Dostopno na: https://www.washingtonpost.com/archive/politics/2003/11/24/on-the-web-research-work-proves-ephemeral/959c882f-9ad0-4b36-88cd-fb7411db118d/ [25. 1. 2025].

Zakon o obveznem izvodu publikacij, 2009. Uradni list RS, št. 69/06 in 86/09. Dostopno na: http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO3606 [25. 1. 2025].