Dolgotrajno ohranjanje raziskovalnih podatkov v manjših raziskovalnih infrastrukturah
Uporaba odprtokodne aplikacije Archivematica

Andrej Pančur,* Bogomir Rožman†

* Inštitut za novejšo zgodovino
Kongresni trg 1, 1000 Ljubljana
andrej.pancur@inz.si
† UniPort – DR, računalniški inženiring, d.o.o.
Fosterjeva 40, 1000 Ljubljana
bogomir.rozman@uniport.si

1 Uvod

Evropska raziskovalna politika je v zadnjih letih začela intenzivno uvajati načela odprtega dostopa do raziskovalnih podatkov. Program Obzorje 2020 tako uvaja pilot za odprte raziskovalne podatke, v okviru katerega morajo udeleženci pripraviti načrt za ravnanje s podatki. Ta načrt mora zajemati življenjski cikel vseh raziskovalnih podatkov, pridobljenih ali ustvarjenih v okviru projekta. Naposled morajo udeleženci izbrati primeren raziskovalni podatkovni repozitorij, ki bo trajno hranil njihove podatke in metapodatke (European Commission, 2016). Države članice Evropske unije naj bi v skladu s priporočili Evropske komisije enaka določila kot za Obzorje 2020 uveljavile tudi za nacionalno financiranje raziskovalnih dejavnosti. V skladu s temi priporočili je Vlada Republike Slovenije septembra 2015 sprejela Nacionalno strategijo odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015-2020, v kateri je med drugim tudi predvideno, da bodo morali udeleženci pilotnega programa odprtega dostopa do raziskovalnih podatkov le-te predati področnim, institucionalnim ali splošnim repozitorijem raziskovalnih podatkov (Vlada Republike Slovenije, 2015, 20).

Pričakujemo lahko, da bodo v prihodnosti v Sloveniji poleg dveh obstoječih repozitorijev raziskovalnih podatkov (Arhiv družboslovnih podatkov[1] in CLARIN.SI[2]) začela nastajati še nova področna podatkovna središča, katera bodo zadovoljevala potrebe posameznih disciplinarnih področij v vseh fazah življenjskega cikla podatkov (Štebe et al., 2013, 7-8). Pri tem se v humanistiki glede na zelo različne vire in raziskovalna vprašanja življenjski cikli podatkov med seboj precej razlikujejo (Puhl et al., 2015). Zato nadaljnji razvoj podatkovne infrastrukture za humanistiko zahteva fleksibilnost, »ki bo zagotovila specializirano obravnavano glede na vsebino in tip podatkov« (Štebe in Bezjak, 2014, 11). Posledično lahko pričakujemo, da bo v Sloveniji (in verjetno tudi v drugih manjših članicah Evropske unije) v naslednjih letih nastalo več manjših specializiranih podatkovih središč, katera bodo lahko optimalno pokrivala celoten življenjski cikel raziskovalnih podatkov posameznih raziskovalnih področij.

Ker bodo denarna sredstva za razvoj teh podatkovnih središč razmeroma omejena, je nujno, da se bodo pri razvoju svoje infrastrukture čim bolj naslonile na obstoječe odprtokodne programske rešitve.

2 Dolgotrajno ohranjanje raziskovalnih podatkov

V okviru življenjskega cikla raziskovalnih podatkov je dolgoročno arhiviranje ponavadi sicer umeščeno na konec raziskovalnega procesa, vendar priprave nanj potekajo že od začetka (izbor metapodatkov in formatov). V predstavitvi se bova osredotočila zgolj na končno arhiviranje. Pri tem bova predpostavljala, da mora vsak zaupanja vreden sistem dolgotrajnega ohranjanja raziskovalnih podatkov temeljiti na standardu odprtega arhivskega informacijskega modela (OAIS).[3]

2.1 Archivematica – odprtokodna aplikacija za dolgotrajno arhiviranje

V zadnjih letih se manjše ustanove, ki se ukvarjajo s hranjenem digitalne kulturne dediščine, prav tako kot raziskovalne ustanove srečujejo s številnimi ovirami (tehnična in intelektualna kompleksnost ter visoki stroški) pri implementaciji najnovejših standardov s področja dolgotrajnega ohranjanja njihovih zbirk. Kot

[1] Arhiv družboslovnih podatkov, http://www.adp.fdv.uni-lj.si/.

[2] CLARIN.SI Repository, https://www.clarin.si/repository/xmlui/.

[3] ISO 14721:2012, Space data and information transfer systems – Open archival information system (OAIS) – Reference model, http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284.

odgovor na te potrebe je leta 2009 nastala Archivematica (Garderen, 2010), katero aktivno razvijajo še danes.[4] Archivematica je integrirana zbirka odprtokodnih programskih orodij, ki uporabniku omogoča obdelavo digitalnih objektov v skladu s funkcionalnim modelom OAIS (zajem, vzdrževanje arhiva, dostop). Uporabnik preko nadzorne plošče upravlja in nadzira izbrane mikrostoritve. Te so razdrobljen sistem opravil, katere delujejo na konceptualnem nivoju OAIS informacijskih paketov: sprejemni informacijski paket (SIP), arhivski informacijski paket (AIP) in dostavni informacijski paket (DIP). Informacijski paketi vsebujejo datoteke, XML metapodatke, dokumentacijo, checksum, log podatke ipd. Poleg originalno razvitih programskih rešitev Archivematica pri tem uporablja še mnoga odprtokodna programska orodja (bulk_extractor, Clam AV, ElasticSearch, ExifTool, FITS, fido, JHOVE, MediaInfo, Tesseract, Imagemagick, md5deep itd.). Archivematica uporablja METS, PREMIS, Dublin Core, BagIt in druge priznane standarde, na podlagi katerih tvori zaupanja vredne, avtentične, zanesljive in sistemsko neodvisne arhivske informacijske pakete (AIP), namenjeni hrambi v poljubnih repozitorijih. Trenutno je Archivematica že integrirana v repozitorije Islandora, dSpace in DuraCloud, ki jih uporabljajo infrastrukture s področja digitalne humanistike. Hkrati je integrirana še v številne sisteme, ki jih uporabljajo v glavnem ustanove s področja kulturne dediščine (CONTENTdm, LOCKSS, AtoM, OpenStack, Archivists' Toolkit, Arkivum, ArchivesSpace).

2.2 Raziskovalna infrastruktura Slovenskega zgodovinopisja

Archivematico smo za dolgotrajno arhiviranje začeli uporabljati tudi v okviru Raziskovalne infrastrukture Slovenskega zgodovinopisja na Inštitutu za novejšo zgodovino (RI INZ). Na RI INZ tako upravljamo portal Zgodovina Slovenije – SIstory,[5] v okviru katerega je najbolj opazna digitalna knjižnica digitaliziranih in digitalnih publikacij ter posnetki konferenc in predavanj. Poleg le-te pa razpolagamo še z različnimi zbirkami raziskovalnih podatkov v relacijskih bazah in v XML zbirkah. Vsi ti raziskovalni podatki niso shranjeni v enem repozitoriju. Hkrati je lahko digitalna kulturna dediščina, ki je najprej dostopna v okviru digitalne knjižnice, kasneje tekom (ponovne) uporabe v življenjskih ciklih raziskovalnih podatkov kot povsem nova zbirka raziskovalnih podatkov dostopna v drugih specializiranih repozitorijih ali bazah podatkov.

3 Zaključek

Na konferenci bo podrobneje predstavljena implementacija Archivematice v RI INZ. Sprva so Archivematico uspešno testirali in začeli uporabljati v številnih repozitorijih s področja ohranjanja kulturne dediščine. Šele pred kratkim pa so bolj intenzivno preizkusili uporabo Archivematice pri hranjenju raziskovalnih podatkov (Mitcham, 2016). Smatramo, da bodo na konferenci predstavljene izkušnje in testne meritve konstruktivno prispevali pri vzpostavljanju dolgotrajne hrambe raziskovalnih podatkov v manjših specializiranih podatkovnih centrih v Sloveniji in drugje po Evropi.

Literatura

European Commission. 2016. *Guidelines on Data Management in Horizon 2020*, verzija 2.1. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Peter Van Garderen. 2010. Archivematica: Lowering the Barrier to Best Practice Digital Preservation. V: Archiving 2010 Final Program and Proceedings, str. 39-41, Society for Imaging Science and Technology.

Jenny Mitcham, Chris Awre, Julie Allinson, Richard Green in Simon Wilson. 2016. Filling the Digital Preservation Gap: A Jisc Research Data Spring project Phase Two report – February 2016. http://dx.doi.org/10.6084/m9.figshare.1481170.

Johanna Puhl, Peter Andorfer, Mareike Höckendorff, Stefan Schmunk, Juliane Stiller in Klaus Thoden. 2015. *Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften*. Göttingen: GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2015-4-4.

Janez Štebe in Sonja Bezjak. 2014. Nastavki odprtih podatkovnih zbirk kot podlaga za družboslovno in humanistično raziskovanje. *Glasnik (Slovensko etnološko društvo)*, 54(1/2): 8-16.

Janez Štebe, Sonja Bezjak in Sanja Lužar. 2013. *Odprti podatki: načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v Sloveniji.* Fakulteta za družbene vede, Založba FDV, Ljubljana. http://www.dlib.si/details/URN:NBN:SI:DOC-US3XRRB2.

Vlada Republike Slovenije. 2015. *Nacionalna strategija odprtega dostopa do znanstvenih objav in raziskovalnih podatkov v Sloveniji 2015-2020.* http://www.mizs.gov.si/fileadmin/mizs.gov.si/pageuploads/Znanost/doc/Zakonodaja/Strategije/Nacionalna_strategija_odprtega_dostopa.pdf.

[4] Archivematica, https://www.archivematica.org/en/.

[5] http://sistory.si/.