

Univerza v Ljubljani
Fakulteta *za družbene vede*



Napredne tehnologije za vzpostavitev raziskovalne podatkovne infrastrukture v Sloveniji

22. maj 2013, Poročilo z delavnice

Janez Štebe

Sonja Bezjak

Sanja Lužar

30. maj 2013, ADP, Univerza v Ljubljani



Janez Štebe, Sonja Bezjak, Sanja Lužar

Napredne tehnologije za vzpostavitev raziskovalne podatkovne infrastrukture v Sloveniji

Poročilo projekta Odprti podatki – priprava akcijskega načrta za vzpostavitev sistema odprtega dostopa do podatkov iz javno financiranih raziskav v Sloveniji

Način dostopa (URL): http://www.adp.fdv.uni-lj.si/o_arhivu/publikacije/odpp10_delavnica_maj2013/



To delo je ponujeno pod [Creative Commons Priznanje avtorstva 2.5 Slovenija](https://creativecommons.org/licenses/by/2.5/si/)

Projekt	ODPRTI PODATKI - Priprava akcijskega načrta za vzpostavitev sistema odprtega dostopa do podatkov iz javno financiranih raziskav v Sloveniji Raziskovalni projekt št. V5-1018 v okviru Ciljnega raziskovalnega programa »KONKURENČNOST SLOVENIJE 2006-2013«
Vodja projekta	Janez Štebe, predstojnik ADP
Vsebinski spremljevalki	Petra Tramte, Direktorat za visoko šolstvo in znanost, MIZŠ, Liljana Lučič, ARRS
Financer	MIZŠ, ARRS

KAZALO

1. KAZALO	3
2. Uvod	4
3. Dnevni red tretje delavnice	5
4. Udeleženci tretje delavnice	5
5. Napredne tehnologije za vzpostavitev raziskovalne podatkovne infrastrukture v Sloveniji	7
6. Predstavitve naprednih podatkovnih tehnologij	8
1. Predstavitve možnosti vključitve raziskovalnih podatkov v nacionalno infrastrukturo odprtega dostopa, Milan Ojsteršek	8
2. Seadatanet: Upravljanje z oceanografskimi in morskimi podatki v okviru pan-evropske infrastrukture, Branko Čermelj	9
3. Enotna prijava in napredni avtentikacijski sistemi, Rok Papež	9
4. Uporaba odprtokodnih rešitev pri zagotavljanju podatkovne infrastrukture, Tomaž Erjavec	9
5. Iskanje po multimedijskih podatkih: primer na arhivu EtnoMuza, Matija Marolt	9
6. Slovenska infrastruktura in omrežja evropskega raziskovalnega prostora, Jan Jona Javoršek	10
7. Sklep	11

Uvod

V sklopu projekta Odprti podatki, ki ga financirata Ministrstvo za izobraževanje, znanost in šport ter ARRS, smo v Arhivu družboslovnih podatkov izvedli tretjo delavnico, tokrat z naslovom Napredne tehnologije za vzpostavitev raziskovalne podatkovne infrastrukture v Sloveniji. Delavnica je logično nadaljevanje preteklih delavnic, v katerih smo najprej problematizirali vprašanje vzpostavitve podatkovnih storitev in potem še vprašanje politik ravnanja z raziskovalnimi podatki financiranimi z javnimi sredstvi.

Cilj navedenih aktivnosti je spoznati okoliščine in pogoje, v katerih deluje slovenska raziskovalna skupnost ter detektirati poznavanje problemov in rešitev na področju odpiranja raziskovalnih podatkov financiranih z javnimi sredstvi. Vse to s ciljem priprave učinkovitega akcijskega načrta za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov financiranih z javnimi sredstvi.

Na tokratni delavnici smo si zadali naslednji cilj: izpostaviti pomen tehnoloških rešitev kot vmesne opreme, ki lahko olajša izmenjavo podatkov med raziskovalci – ustvarjalci podatkov, ponudniki podatkovnih storitev in raziskovalci – uporabniki podatkov.

Ta dokument je poročilo tretje delavnice in je namenjen seznanjanju zainteresirane javnosti s problematiko odpiranja raziskovalnih podatkov financiranih z javnimi sredstvi.

Dnevni red tretje delavnice

9.00 - 9.20 **Uvodna beseda in predstavitev obeh delavnic** *Janez Štebe, Arhiv družboslovnih podatkov*

9.20 - 9.35 **Predstavitev možnosti vključitve raziskovalnih podatkov v nacionalno infrastrukturo odprtega dostopa** *Milan Ojsteršek, Laboratorij za heterogene računalniške sisteme, Fakulteta za elektrotehniko, računalništvo in informatiko, UM*

9.35 - 9.50 **Seadatanet: Upravljanje z oceanografskimi in morskimi podatki v okviru pan-evropske infrastrukture** *Branko Čermelj, Nacionalni inštitut za biologijo, Morska biološka postaja Piran*

9.50 - 10.05 **Enotna prijava in napredni avtentikacijski sistemi** *Rok Papež, ARNES*

10.05 - 10.20 **Uporaba odprtokodnih rešitev pri zagotavljanju podatkovne infrastrukture** *Tomaž Erjavec, Odsek za tehnologije znanja, Institut "Jožef Stefan"*

10.20 - 10.35 **Podatkovni viri v projektu XLike za medjezično luščenje znanja** *Blaž Fortuna (se ni izvajalo)*

10.35 - 10.50 **Iskanje po multimedijskih podatkih: primer na arhivu EtnoMuza** *Matija Marolt, Laboratorij za računalniško grafiko in multimedije, Fakulteta za računalništvo in informatiko, UL*

10.50 - 11.05 **Slovenska infrastruktura in omrežja evropskega raziskovalnega prostora** *Jan Jona Javoršek, Institut "Jožef Stefan" in SLING*

11.05 - 11.35 **Razprava**

11.45 - 12.15 **Odmor in pogostitev**

Udeleženci tretje delavnice

Prvi delavnici sta bili interaktivni v smislu, da smo v Arhivu pripravili vsebine in udeležence pozvali, da bodisi podajo svoje predloge, pripombe, izboljšave bodisi pomagajo pri iskanju rešitev. Tretjo delavnico smo zasnovali kot sklop predstavitev, ki so jih izvajali vabljeni govorniki. Med šestimi govorniki sta bila po dva iz fakultet za računalništvo in informatiko, po dva z Instituta Jožef Stefan in po en z ARNES – a in Morske biološke postaje Piran.

Poslušalstvo je bilo veliko bolj raznoliko zastopano. Saj so bili med udeleženci štirje predstavniki treh različnih ministrstev, trije z dveh fakultet za računalništvo in informatiko, sedem predstavnikov z nacionalne in različnih področnih knjižnic, po nekaj predstavnikov zasebnih in javnih podjetij ter raziskovalnih inštitutov in fakultet ter po en predstavnik SURS in ZRC SAZU.

Sledi poimenski seznam udeležencev delavnice Napredne tehnologije za vzpostavitev raziskovalne podatkovne infrastrukture v Sloveniji:

1. Tjaša Bezenšek, Centralna medicinska knjižnica
2. Sonja Bezjak, Arhiv družboslovnih podatkov
3. Branko Čermelj, Nacionalni inštitut za biologijo, Morska biološka postaja Piran
4. Barbara Černač, Nacionalni inštitut za biologijo, Biotehniška fakulteta, UL
5. Tomaž Erjavec, Institut "Jožef Stefan"
6. Blaž Fortuna, Institut "Jožef Stefan"
7. Ivan Fučak, RRC Računalniške storitve d.d.
8. Jana Intihar Ferjan, Moderna galerija
9. Manica Jakič Brezočnik, Ministrstvo za izobraževanje, znanost in šport
10. Jan Jona Javoršek, Institut "Jožef Stefan"
11. Borut Kovačič, Javno podjetje Prlekija d.o.o.
12. Mojca Kotar, Univerza v Ljubljani
13. Tanja Koželj, Filozofska fakulteta, UL
14. Igor Kuzma, Statistični urad RS
15. Dejan Lavbič, Fakulteta za računalništvo in informatiko, UL
16. Žiga Lipar, Gozdarski inštitut Slovenije
17. Sanja Lužar, Arhiv družboslovnih podatkov
18. Darko Majcenović, Centralna medicinska knjižnica
19. Matija Marolt, Fakulteta za računalništvo in informatiko, UL
20. Vlasta Miklavžin, Prometni inštitut
21. Uroš Mladenovič, Ministrstvo za infrastrukturo in prostor
22. Mojca Mlinar Strgar, Biblioteka SAZU
23. Milan Ojsteršek, Fakulteta za elektrotehniko, računalništvo in informatiko, UM
24. Rok Papež, ARNES
25. Miha Peče, ZRC SAZU
26. Danica Šaponja, Ministrstvo za notranje zadeve in javno upravo
27. Janez Štebe, Arhiv družboslovnih podatkov
28. Petra Tramte, Ministrstvo za znanost, izobraževanje in šport
29. Gorazd Vodeb, NUK
30. Samo Zorc, Ministrstvo za izobraževanje, znanost in šport
31. Aljaž Zrnc, Fakulteta za računalništvo in informatiko, UL
32. Slavko Žitnik, Optilab d.o.o. & Laboratorij za podatkovne tehnologije
33. Petra Vide Ogrin, Biblioteka SAZU
34. Dalibor Radovan, Geodetski inštitut Slovenije

Uvod v Napredne tehnologije za vzpostavitev raziskovalne podatkovne infrastrukture v Sloveniji

Mišljenje, da napredne tehnologije omogočajo učinkovitejšo izrabo raziskovalnih podatkov, je že bolj ali manj privzeto. Pa vendarle je na poti do uresničevanja tega cilja še precej ovir, tako na ravni financ, kulture, kadrov itd. Pri premoščanju le-teh so velikega pomena mednarodne pobude (npr. OECD in Evropska komisija), ki kažejo smernice in zlasti tiste države, ki so v zaostanku, spodbujajo in tudi zavezujejo, da se lotijo reševanja problemov. V Sloveniji je na področju odpiranja raziskovalnih podatkov eden pomembnejših korakov projekt Odprti podatki, ki ga financirata Ministrstvo za izobraževanje, znanost in šport in ARRS, izvaja pa ga Arhiv družboslovnih podatkov.

Z uvodno predstavitevjo smo udeležencem delavnice osvetlili okoliščine, v katerih nastajajo, se hranijo in izmenjujejo raziskovalni podatki v Sloveniji. Predstavljene ugotovitve temeljijo na preglednem poročilo z naslovom *Opis stanja na področju raziskovalnih podatkov v Sloveniji*, ki je izšlo oktobra 2012 in je utemeljeno na mednarodni primerjavi ter na poglobljenih intervjujih s slovenskimi raziskovalci. Dopolnili smo jih z izsledki delavnic *Problemi in rešitve na področju vzpostavljanja podatkovnih storitev v Sloveniji* (december 2012) ter *Politike ravnanja z raziskovalnimi podatki v Sloveniji* (marec 2013).

Slovenski raziskovalci vključeni v našo raziskavo so med ovirami, ki otežujejo izmenjavo raziskovalnih podatkov, navajali: odsotnost politik, ki bi s pravili in dogovorom raziskovalne skupnosti regulirale področje, pomanjkanje finančnih sredstev ter pomanjkanje kadrov, ki bi poskrbeli za ustrezno pripravo in hrambo podatkov in s tem za večjo možnost izmenjave. Kot pomembno oviro so navajali tudi tekmovalnost in konkurenčnost, ki bi jo bilo mogoče preseči z ustreznim ovrednotenjem napora vloženega v ustvarjanje in pripravo podatkov.

Slovenijo na poti k učinkovitejši izrabi raziskovalnih podatkov financiranih z javnimi sredstvi čaka izoblikovanje krovnih politik o ravnanju z raziskovalnimi podatki, ki bodo opredelile vloge in odgovornosti med deležniki ter vzpostavitev podatkovnih storitev in infrastrukture, ki bodo raziskovalcem nudile oporo pri pripravi, hrambi in izmenjavi raziskovalnih podatkov ter pri ovrednotenju njihovega napora.

S tem pa smo že prešli k napovedi tretje delavnice, na kateri so bile predstavljene napredne tehnološke rešitve s področja vzpostavljanja dostopa do podatkov in njihovega izkoriščanja, potenciali GRID-a, semantičnih orodij in drugo. Vse s ciljem, da bi izpostavili pomen tehnoloških rešitev kot vmesne opreme, ki lahko olajša izmenjavo podatkov med raziskovalci – ustvarjalci podatkov, ponudniki podatkovnih storitev in raziskovalci – uporabniki podatkov.

Predstavitve naprednih podatkovnih tehnologij

Pomembnejši del delavnice smo namenili vabljenim govorcem, ki so vsak z svojega področja predstavili podatkovne tehnologije, njihove možnosti in izrabo v slovenskem raziskovalnem okolju. Govorci so si sledili v naslednjem vrstnem redu

1. *Milan Ojsteršek, Laboratorij za heterogene računalniške sisteme, Fakulteta za elektrotehniko, računalništvo in informatiko, UM*
2. *Branko Čermelj, Nacionalni inštitut za biologijo, Morska biološka postaja Piran*
3. *Rok Papež, ARNES*
4. *Tomaž Erjavec, Odsek za tehnologije znanja, Institut "Jožef Stefan"*
5. *Matija Marolt, Laboratorij za računalniško grafiko in multimedije, Fakulteta za računalništvo in informatiko, UL*
6. *Jan Jona Javoršek, Institut "Jožef Stefan" in SLING*

1. Predstavitev možnosti vključitve raziskovalnih podatkov v nacionalno infrastrukturo odprtega dostopa, Milan Ojsteršek

V okviru projekta ODUN poteka izgradnja repozitorijev Univerz v Ljubljani, Mariboru, Novi Gorici in na Primorskem. Namen projekta je omogočiti odprt dostop do intelektualne produkcije slovenskih univerz. Univerzitetni repozitoriji so v prvi vrsti namenjeni zagotavljanju odprtega dostopa do objav, vanje pa bo mogoče vključiti tudi raziskovalne podatke.

V projektu ODUN so si zadali naslednje naloge:

- izdelati pilotni nacionalni portal in univerzitetne repozitorije, ki bodo omogočali dostop preko spletnih brskalnikov, prenosnih naprav in mobilnih telefonov v slovenskem in angleškem jeziku. Spletna različica bo prijazna do ljudi s posebnimi potrebami in bo poskušala zagotoviti glavne značilnosti spletnih aplikacij, ki ustrezajo specifikaciji WAI;
- vzpostaviti proces, ki bo raziskovalcem in študentom omogočal vnos gradiv in podatkov raziskav v repozitorij univerze, na kateri delajo ali študirajo;
- integrirati repozitorije z informacijskimi sistemi slovenskih univerz in sistemom eVŠ;
- povezati repozitorije z evropskim portalom znanstvenih objav (OpenAIRE), zaključnih del študija (DART-Europe) in registri odprtega dostopa (ROAR, OpenDOAR in BASE);
- vzpostaviti interoperabilnost repozitorijev z informacijskim sistemom COBISS.SI;
- izdelati sistem za priporočanje sorodnih vsebin;
- izdelati sistem za ugotavljanje podobnih vsebin. (iz predstavitve M. Ojstrška, 22.5.2013)

Projekt se konča 30. avgusta 2013.

2. Seadatanet: Upravljanje z oceanografskimi in morskimi podatki v okviru pan-evropske infrastrukture, Branko Čermelj

Morska biološka postaja Piran je vključena v SeaDataNet - pan evropsko infrastrukturo namenjeno upravljanju morskih in oceanografskih podatkov. Gre za vzpostavitev evropskih zbirk oziroma katalogov morskih projektov, podatkovnih zbirk, organizacij, opazovalnih sistemov v morskem okolju ter križarjenj.

V infrastrukturo je trenutno povezanih 83 centrov, ki razvijajo skupne standarde za "morske" podatke, imajo skupno strategijo, licenco in orodja, ki se med seboj "razumejo". Centri imajo enoten spletni portal. Eden izmed glavnih ciljev je zagotavljanje celovitega pregleda in dostop do podatkov, ki jih hranijo podatkovni centri. (iz predstavitve B. Čermelja, 22.5.2013)

3. Enotna prijava in napredni avtentikacijski sistemi, Rok Papež

Dostop do podatkov velja za eno pomembnejših storitev na področju vzpostavljanja sistema odprtega dostopa do raziskovalnih podatkov. V predstavitvi je bilo govora o različnih oblikah dostopa in tveganjih, ki izvirajo iz njih. Predstavljeni so bili problemi množice uporabniških imen in gesel ter napredni mehanizmi avtentikacije, kot sta dvojna prijava in uporaba certifikatov za prijavo. Kot rešitev je bila predstavljena ločitev avtentikacije in storitve.

4. Uporaba odprtokodnih rešitev pri zagotavljanju podatkovne infrastrukture, Tomaž Erjavc

Govorec je predstavil "infrastrukturo" za področje (slovenskih) jezikovnih virov. Na portalu ijs.nl, ki obstaja že od leta 1995, teče več projektov. Med njimi: projekt JOS, ki je namenjen spodbujanju razvoja tehnologij za slovenski jezik, korpusi, slovarji, leksikoni, registri, digitalne knjižnice ... Pri uresničevanju projektov uporabljajo odprtokodne rešitve, ki prinašajo več prednosti:

- odprta koda: brezplačna, možnost popraviljanja napak in dograjevanja,
- odprti standardi: jasen, dobro dokumentiran, od implementacije neodvisen in trajen način zapisa besedil,
- odprte licence: enostavna in predvsem čim širša nadaljnja uporaba razvitih (jezikovnih) virov, zgrajenih z državnim financiranjem. (iz predstavitve T. Erjavca, 22. maj 2013).

5. Iskanje po multimedijskih podatkih: primer na arhivu EtnoMuza, Matija Marolt

Velika količina multimedijskih vsebin z majhnim deležem metapodatkov predstavlja poseben izziv. Na primeru EtnoMuza - digitalni arhiv slovenske ljudske glasbe in plesa GNI ZRC SAZU - je bilo predstavljeno orodje za iskanje neposredno po gradivih. Med prvimi koraki je bila vzpostavitev arhiva, ki obsega orodje za vnos metapodatkov in spletni vmesnik za iskanje in

brskanje po metapodatkih. Naslednji korak je prinesel razvoj naprednih tehnologij, s katerimi je mogoče iskati po notografijah in zvočnih posnetkih, o katerih ni metapodatkov ali jih je zelo malo.

Predstavljeno je bilo orodje za avtomatsko prepoznavanje vsebine. Členitev gradiva na: govor, solo petje, večglasje, instrumental, pritrkavanja. Ter možnost členitve vsebine na:

a) petje in inštrumentali:

- iskanje po melodiji, parametrih melodije
- sestav, število glasov, spol

b) govor:

- število govorcev, spol

c) pritrkavanja:

- iskanje po vzorcih
- število zvonov. (iz predstavitve M. Marolta, 22. maj 2013)

6. Slovenska infrastruktura in omrežja evropskega raziskovalnega prostora, Jan Jona Javoršek

V zadnji predstavitvi je bilo predstavljeno mrežno računalništvo, predvsem pa potencial grida oz. kdaj, kako in zakaj ga uporabiti. Grid je vmesna programska oprema, ki jo je smiselno uporabiti, ko imamo opravka z veliko količino podatkov, ko izvrševanje nalog zahteva veliko procesorsko moč ali ko gre za veliko število nalog. Grid omogoča lažjo uporabo mrežnih računalniških kapacitet (enotni, standardizirani vmesniki, samodejna izbira, gruče, upravljanje z nalogami in podatki), uporabo novih tehnologij (high throughput computing, nameščanje programske opreme, virtualizacija na različnih stopnjah ipd.) ter lažje znanstveno-raziskovalno sodelovanje (virtual research communities).

Predstavljena je bila tudi Slovenska iniciativa za nacionalni grid (SLING). Ključna razloga zanjo sta vzpostavitev stalne nacionalne infrastrukture in vključitev v mednarodno infrastrukturo EGI.

Sklep

Ker so potrebe po naprednih podatkovnih tehnologijah velike, je bil odziv udeležencev na predstavljene rešitve pričakovan. Udeležencem so z vprašanji aktivno sodelovali in ustvarjali tok razprave, ki se je vlekel od prvih seznanitev s posameznimi rešitvami, do izmenjave dobrih izkušenj ter navezovanja prihodnjih stikov.

Zaključiti je mogoče, da v Sloveniji že imamo strokovnjake, ki se na svojih področjih poglobljeno ukvarjajo z razvojem podatkovnih tehnologij in lahko predstavljajo pomembna središča pri vzpostavljanju sistema odprtega dostopa do raziskovalnih podatkov. Ker prihajajo z različnih raziskovalnih področij (naravoslovje, humanistika, družboslovje) poznajo problematiko različnih tipologij podatkov in tehnoloških ovir ter prednosti tehnologije, s katero delajo.

Udeleženci so pozdravili srečanje in poudarili potrebo po nadaljevanju tovrstnih aktivnosti, s katerimi področne ali parcialne rešitve vstopijo v skupno raziskovalno okolje.