Plan upravljanja istraživačkim podacima - IntRegVar

Barešić, Anja

Data management plan / Plan upravljanja istraživačkim podacima

Plan upravljanja istraživačkim podacima

Opće informacije		
	Ime i prezime predlagatelja	Anja Barešić
	Matična organizacija	Institut Ruđer Bošković
	Naziv projekta	Istraživanje interakcija između regulatornih varijanti u kontekstu bolesti čovjeka, UIP-2020-02-1623
	Upravitelj podacima	Anja Barešić, Zavod za elektroniku, IRB, Bijenička cesta 54, 10000
1.	Prikupljanje podataka i dokumentacija	
	Koje ćete podatke prikupljati, obrađivati, stvarati ili se ponovno njima koristiti? (navedite format, vrstu i opseg podataka)	Tijekom projekta se prikupljaju podaci iz javnih izvora: 1. O1-O3 se temelje na podacima o genomima, različitim „-omics" značajkama te varijantama dobivenim tijekom prethodno izvedenih eksperimenata van ove grupe i projekta, objavljenim u obliku znanstvenih publikacija i priloženih setova rezultata te objavljeni za javnu upotrebu znanstvenoj zajednici preko javno-dostupnih repozitorija. U pravilu je riječ o .csv tabularnom prikazima, kao i genomskim vrpcama (engl. genomic tracks) u kojima se signal zapisuje za svaku poziciju u genomu: ordinalni u .bed, .bigBed, .vcf i sl. te kontunuirani u .wig, .bigWig i sličnim formatima (https://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#CustomTracks) 2. Ostali podaci koji će biti obrađeni u sklopu suradnji sa drugim grupama prikupljani su unutar tih projekata i u skladu sa njihovim PUIP-om. Članovi grupe uvijek u fazi planiranja suradnje jasno komuniciraju da su im za rad (mi analiziramo podatke koji su nam proslijeđeni nakon faze prikupljanja) potrebni anonimizirani podaci koji su prikupljeni u skladu sa etičkim i zakonskim normama te FAIR principima i potičemo ispravljanje eventualnih propusta u FAIR procedurama prilikom faze planiranja projekta, ako je to moguće.
	Kako će se podaci prikupljati, obrađivati ili stvarati? (ukratko navedite metodologiju i procese osiguranja kvalitete te načine organiziranja podataka)	Podaci **prikupljeni** iz javno dostupnih baza se prikupljaju i pohranjuju na diskovlju nabavljenom specifično za ovaj projekt u digitalnom obliku. Prikupljene datoteke u istoj mapi posjeduju tekstualni zapis web poveznice na izvor i vremensku oznaku kako bi se osigurala informacija o verziji seta podataka na kojima se radi (README file) te strukturu mape i opis datoteka. Gdje je potrebno (u slučaju kada se koriste verzije podataka iz istog izvora u različitom vremenu), ime same skinute datoteke ukazuje na vrijeme skidanja. Datoteke sa podacima nastale **u daljnjoj obradi** se pohranjuju na istom mjestu i prate princip imenovanja ulaznih podataka osiguravajući jednoznačno povezivanje sa ulaznim podacima te se informacije o istima također nalaze u README dokumentu. Ulazni podaci (varijante genoma čovjeka) prolaze rigoroznu provjeru (semantičke i logičke) **ispravnosti** formata podataka te se, s obzirom na velike količine podatak uočene pogreške ne pokušavaju ispraviti (zbog nemogućnosti kontaktiranja primarnog izvora), nego se nepotpuni i djelomično pogrešni podaci eliminiraju iz daljnje analize.
	Koju ćete dokumentaciju i metapodatke ustupiti osim podataka?	Većina računalnih analiza se izvode u programskom jeziku R kroz tzv. R notebook koji, uz upotrebu specificiranih ulaznih podataka, osigurava potpuno reproducibilni postupak dobivanja identičnih rezultata u svakom trenutku. Ova metodologija

	(navedite koje su informacije potrebne korisnicima kako bi mogli čitati i interpretirati podatke u budućnosti te koji će se standardi koristiti pri tumačenju podataka)	pohranjuje informaciju o autoru, datumu izvođenja samog računalnog eksperimenta, svim potrebnim paketima i verzijama istih koji su korišteni te lokaciji i imenu ulaznih podataka i proizvedenih rezultata. Markdown metodologija omogućava *free-text* komentiranje kojima se mogu dodati razni dodatni elementi gore navedenim podacima, i idealna je jer objedinjava metapodatke i računalni kod u istoj datoteci. Analogno, sve potrebne analize koje su obavljene u pythonu se izvode u Jupyter notebooku koji također svojim bilježenjem svih meta podataka osigurava potpunu reproducibilnost, kao i docker slike sa docker kontejnerom.
2.	Pravna i sigurnosna pitanja	
	Jeste li ograničeni sporazumom o povjerljivosti? Imate li potrebna dopuštenja za prikupljanje, obradu, čuvanje i dijeljenje podataka? Jesu li osobe čiji se podaci pohranjuju informirani o tome i jesu li dali privolu? Kojim ćete se metodama koristiti u svrhu zaštite osjetljivih podataka (GDPR - posebne kategorije osobnih podataka)?	Na projektu ne postoje ograničenja u vidu ugovora o povjerljivosti. Podaci koji su prikupljeni su povijesno uzeti od strane pacijenata u sklopu velikih studija izvedenih na inozemnim institucijama kontrolirane od strane nadležnih tijela te im mi pristupamo tek kada se u anonimiziranom obliku stavljeni na raspolaganje u javnu domenu (članci, javni repozitoriji). Stoga smatramo da se pri izvedbi ovog projekta neće kršiti etička načela.
	Kako će se regulirati pristup podacima i njihova sigurnost? Koji su potencijalni rizici koje treba uzeti u obzir? Kako ćete osigurati sigurnost pohrane osjetljivih podataka?	Sami podaci (među kojima nema osjetljivih podataka) korišteni unutar projekta neće ponovno biti dani na pristup jer su isti već javno dostupni, objavljena metodologija jasno ukazuje gdje ih se može pronaći (čime štedimo lokalne resurse ne duplicirajući pohranjene podatke), no bit će ih moguće dobiti od autora na zahtjev (osiguravana je lokalna pohrana). Rezultati analiza će biti objavljeni u skladu sa uvjetima objave u open access časopisima te pospremljeni u baze odgovarajuće zahtjevima određenog časopisa. U slučaju da poveznice na primarne izvore podataka budu uklonjene, dodat ćemo našim resursima i tim datotekama. Osigurana je pohrana **svih podataka** minimalno za period trajanja projekata, a minimalnog seta koji osigurava reproducibilnost rezultata još barem 3 godine.
	Kako ćete upravljati zaštitom autorskih prava i intelektualnog vlasništva? Tko će biti vlasnik podataka? Koje će se licencije primjenjivati na podatke? Koja će se ograničenja primjenjivati na ponovnu uporabu osobnih podataka?	Rezultati istraživanja će se, prema uvjetima Uspostavnih projekata HRZZ-a, staviti na raspolaganje javnosti kroz „open access" publikacije. U slučaju dodatnih rezultata koji imaju potencijal za patentiranje, primjenjivat će se pravila utvrđena internim „Pravilnikom o intelektualnom vlasništvu" IRB-a te u skladu sa „Pravilnikom o upravljanju rezultatima znanstvenih projekata koji su prikladni za zaštitu pravima intelektualnog vlasništva". IRB ima ustaljene procese za ovakve scenarije (Ruđer Inovacije).
3.	Pohrana i čuvanje podataka	

	Kako će podaci biti pohranjeni i kako će biti napravljena sigurnosna kopija podataka (*backup*) tijekom istraživanja? Koji su kapaciteti čuvanja podataka kojim raspolažete? Kojim se procedurama koristite za sigurnosnu kopiju (*backup*)?	Podaci su pohranjeni na diskovlju stroja Orthus koje je nabavljeno sredstvima sa ovog projekta (https://www.croris.hr/oprema/oprema/4215) te se fizički nalazi u prostorima Instituta. Podatkovni čvor sa 80TB prostora je više nego dovoljan za potrebe dvaju uspostavnih projekata iz čijih sredstava je nabavljen ovaj stroj. Svi važniji međuprodukti analiza, čija bi rekonstrukcija zahtjevala značajniju upotrebu računalnih resursa se nalaze u barem još jednoj kopiji na osobnim računalima 4 člana grupe. Svaki član grupe ima svoje prijenosno računalo, nabavljeno uz garanciju produljenog jamstva na 5 godina. Sav programski kod je pospremljen na Orthus, kao i na cloud platformi Github, gdje grupa ima svoj zajednički korisnički račun - vidljiv članovima, ali ne i javnosti. S obzirom da članovi svoja prijenosna računala redovito nose kući, smatramo da za sve podatke postoji off-site backup za vrijeme čitavog trajanja projekta. U zadnjoj godini projekta planirat će se strategija dodatnog backupa onog dijela podataka koji neće biti u javnoj domeni na u tom trenutku dostupna, novija računala.
	Koji je vaš plan čuvanja podataka? U kojim će se formatima čuvati?	Podaci će se čuvati deponiranjem u javne baze računalnog koda (Github – otvoren za javnost po objavljivanju članaka) i javne baze podataka za genomske podatke (npr. ENA, u sklopu ELIXIR-a https://www.ebi.ac.uk/ena/browser/home ili prikladnije). Ne predviđa se veliki volumen generiranih podataka koje je potrebno pospremiti na dulje vrijeme: svi potrebni izračuni na kojima se temelji targPred klasifikacija i modeliranje se svode na tablični zapis (.csv) reda veličine u megabajtima, a vizualizacije zauzimaju tek nekoliko gigabajta. Ostali prostor je potreban za fazu izračuna na ukkupnom setu podataka te se datoteke privremene upotrebljivosti brišu čim za njima više nema potrebe jer se reproducibilnim principom razvoja metodologije osigurava da se u svakom trenutku mogu rekreirati. Sve vizualizacije na targPred web alatu se mogu rekreirati iz tabličnih podataka i računalnog koda koji će biti dostupan na Githubu po objavi članka sa zanemarivim računalnim resursima za pojedinačne primjere izračuna te pohrana vizualizacija postaje suvišna kada se implementira „on-the-fly" opcija prikazivanja za sve varijante u targPred-u.
4.	Dijeljenje i ponovna uporaba podataka	
	Kako i gdje će se podaci dijeliti? Na kojem repozitoriju planirate dijeliti podatke? Kako će potencijalni korisnici doznati za podatke?	Svaki objavljeni članak će imati cjelokupni računalni kod i ulazne podatke jasno navedene i dostupne: podatke na originalnim web poveznicama sa kojih su dobiveni (npr. https://www.ebi.ac.uk/gwas/docs/file-downloads) i kod sa ukupnim, potpuno reproducibilnim tijekom računalnog eksperimenta na Github (https://github.com/) korisničkom računu grupe te Zenodo poveznicom. Sve diseminacijske aktivnosti navode web poveznice na relevantne izvore, a posebno na web stranicu sa rezultatima projekta http://targpred.irb.hr/. Svi članovi grupe imaju otvoreni korisnički račun u Dabar, ORCID, itd. te se svi objavljeni podaci u sklopu projekta redovito unose u ove baze.
	Ako postoje podaci koji se ne smiju dijeliti (prijavitelji vezani zakonskim, etičkim, autorskim pravila, povjerljivošću i sl.), pojasnite razloge ograničenja.	Ne postoje.
	Potvrdite da ćete se koristiti digitalnim repozitorijem koji je u skladu s načelima *FAIR-a*.	Findable – jasne poveznice na sve komponente uključene u izvještaje, publikacije i diseminacijske materijale. Accessible – open access publikacije, kod dostupan na githubu, javno dostupan web resursweb resurs Interoperable – podaci se čuvaju u standardnim formatima za multiplatformsko čitanje (.csv datoteke) te standardnim za uobičajene vizualizacijske alate (npr. R pakete).

		Reusable – sve komponente sustava su dizajnirane kako bi web alat bio upotrebljiv ne samo na ulaznim parametrima, nego i na bilo kojem podatku tog tipa te je dodavanje te funkcionalnosti planirano unutar projekta u drugoj fazi razvoja targPred alata. Pohranom podataka u repozitorije poput Dabar i Github+Zenodo osiguravamo da pohrana podataka bude u skladu sa FAIR načelima.
	Potvrdite da ćete se koristiti digitalnim repozitorijem koji održava neprofitna organizacija (ako ne, objasnite zašto ne možete dijeliti podatke na digitalnom repozitoriju koji nije komercijalan).	Koristit će se SVI oni repozitoriji koji dokazuju svoju sigurnost, ali i komercijalni koji su standard u domeni (poput Githuba za računalni kod) dok god ne podliježu dodatnim troškovima.

Ref:

[1] Celjak, D., Dorotić Malič, I., Matijević, M., Poljak, Lj., Posavec K. i Turk, I.: „Istraživački podaci - što s njima?" Istraživački podaci - što s njima? : priručnik o upravljanju istraživačkim podacima | Digitalni repozitorij Srca (unizg.hr)