NAČRT ZA RAVNANJE Z RAZISKOVALNIMI PODATKI

OBRAZEC ARIS

0	**Splošne informacije**	
0.1	**Šifra projekta**	J2-70078
0.2	**Naziv projekta**	Od velikih podatkov do dobrih podatkov: pametna primerjalna analiza za zanesljivo umetno inteligenco (DATA-TRUST)
0.3	**Šifra vodje projekta**	50854
0.4	**Ime in priimek vodje projekta**	Tome Eftimov
0.5	**Ime in priimek osebe, ki je v RO zadolžena za podporo pri ravnanju z raziskovalnimi podatki**	/
0.6	**Interna pravila RO za ravnanje z raziskovalnimi podatki**	Načrt za ravnanje s podatki (https://zic.ijs.si/odprta-znanost/#NRRP)
0.7	**Verzija NRRP**	1.0
1	**Povzetek in opis raziskovalnih podatkov**	
1.1	**Ali boste pri projektu ponovno uporabili že obstoječe podatke predhodnih raziskav?**	☒ Da ☐ Ne Uporabili bomo javno dostopne benchmarking podatke za zbirke testnih problemov na področju zvezne enociljne optimizacije ter za klasifikacijo časovnih vrst. Poleg tega bomo uporabili tudi podatke za primerjalno vrednotenje (angl. benchmarking data)) velikih jezikovnih modelov (LLM), s katerimi bomo ocenjevali prenosljivost metodologij na druga področja strojnega učenja. Ti podatki bodo uporabljeni za preizkušanje metodologij za izbor reprezentativnih podatkov, z namenom zmanjšanja pristranskosti primerjav.
1.2	**Katere vrste podatkov boste ustvarili oz. ponovno uporabili in v katerih formatih bodo shranjeni?**	Za probleme enociljne optimizacije bomo uporabili primerjalne podatkovne zbirke (MA-BBOB, COCO, CEC, RANDOM), ki vsebujejo optimizacijske probleme. Za vsak problem bomo izračunali različne predstavitve (vektorje značilk), ki opisujejo lastnosti optimizacijske pokrajine, kot so ELA, DoE2Vec, DeepELA, TransOptAS, ter trajektorijske predstavitve, kot so DynamoRep, Probing, Trajectory Features in Iterative ELA. Vse primerjalne podatkovne zbirke bodo predstavljene v obliki tabelaričnih podatkov, shranjeno v formatih CSV, TXT ali JSON. Poleg tega bodo metapodatki o predstavitvah vključeni v ontologijo OPTION, ki je bila razvita v naši raziskovalni skupini za zagotavljanje interoperabilnosti podatkov. V primeru klasifikacije časovnih vrst bomo uporabili podatkovne zbirke iz repozitorija UCR. Primeri v posameznih podatkovnih zbirkah bodo predstavljeni s klasičnimi predstavitvami časovnih vrst, kot sta catch22 in tsfresh, ter z novejšimi metodami učenja predstavitev, ki omogočajo pridobivanje vektorskih predstavitev časovnih vrst. Vse primerjalne podatkovne zbirke bodo predstavljene s tabelaričnimi podatki, shranjeno v formatih CSV, TXT ali JSON. Ti formati so široko uporabljeni v raziskovalni skupnosti ter omogočajo enostavno ponovno uporabo podatkov in ponovljive analize. V primeru primerjalne analize za velike jezikovne modele (LLM) bodo pozivi (prompts) predstavljeni z različnimi besedilnimi vektorskimi

		predstavitvami (text embeddings), pridobljenimi z uporabo manjših in večjih jezikovnih modelov. Rezultati bodo shranjeni kot tabelarične podatkovne zbirke v formatih CSV, JSON ali TXT, kar omogoča enostavno ponovno uporabo podatkov in zagotavlja ponovljivost analiz. Za nove primerjalne podatkovne zbirke, ki bodo nastale v okviru projekta, bomo ponovno uporabili oziroma razvili podatkovne zbirke na podlagi metod učenja predstavitev, povezanih z domenami enociljne optimizacije (SOO), analize časovnih vrst (TSA) in velikih jezikovnih modelov (LLM). Vse primerjalne podatkovne zbirke bodo vsebovale tudi izvorne (angl. raw) podatke o uspešnosti različnih algoritmov, ovrednotenih na teh podatkih. Za podatke o uspešnosti bomo uporabljali javno dostopne lestvice rezultatov (leaderboards) in rezultate primerjalnih kampanj, ki so uveljavljene na posameznih raziskovalnih področjih.
1.3	**Kakšen je namen ustvarjanja, zbiranja oz. ponovne uporabe podatkov in njihova povezava s cilji projekta?**	Podatkovne zbirke, v katerih bo vsak primer predstavljen z ustreznimi metaznačilkami, bodo uporabljene za razvoj in preizkušanje metodologij iz delovnega sklopa DS2 za izbor reprezentativnih podmnožic podatkov z namenom zmanjšanja inherentne pristranskosti primerjav. Izbrane podmnožice bodo skupaj z izvornimi podatki o uspešnosti algoritmov uporabljene za vrednotenje in preverjanje metodologij, razvitih v okviru DS3. Glavni cilj projekta je izbrati reprezentativne podatkovne zbirke, ki bodo omogočile bolj robustne, zanesljive in manj pristranske rezultate primerjav.
1.4	**Kakšna je pričakovana velikost podatkov, ki jih nameravate ustvariti oz. ponovno uporabiti?**	☐ 0–10 GB ☒ 10–100 GB ☐ 100–1000 GB ☐ >1000 GB
2	**Shranjevanje in varnostno kopiranje podatkov**	
2.1	**Kje bodo podatki med izvajanjem projekta shranjeni in varnostno kopirani?**	Podatki bodo med izvajanjem projekta shranjeni na računalniški infrastrukturi Odseka za računaniške sisteme Instituta Jožef Stefan. Raziskovalna skupina ima dostop do visokozmogljivih računalniških virov, vključno s strežniki i (576 + 576 CPU jeder), 4 TB delovnega pomnilnika (RAM) ter grafičnimi procesnimi enotami (GPGPU), ki zagotavljajo do 51,4 TFLOPS vršne zmogljivosti pri dvojni natančnosti (FP64). Za shranjevanje podatkov je na voljo več kot 0,5 PB diskovnega prostora. Vsi podatki bodo hranjeni znotraj omrežja Instituta Jožef Stefan, kjer se izvajajo redni postopki varnostnega kopiranja in nadzora dostopa. Lokalna infrastruktura zadostuje za izvedbo osrednjega eksperimentalnega programa projekta. V primeru povečanih potreb po računalniških ali shranjevalnih zmogljivostih bodo uporabljeni tudi viri Slovenske iniciative za nacionalni grid (SLING), katere član konzorcija je Institut Jožef Stefan. SLING zagotavlja dodatne računske in podatkovne zmogljivosti ter s tem povečuje zanesljivost in razpoložljivost infrastrukture za izvajanje raziskav.
2.2	**Kako boste izbrali podatke za dolgoročno hrambo?**	Projekt je usmerjen v razvoj metodologij za izbor reprezentativnih podmnožic podatkov iz obsežnih primerjalnih podatkovnih zbirk. Za namene dolgoročne hrambe bomo zato ohranili vse analizirane primerjalne podatkovne zbirke skupaj z njihovimi predstavitvami v obliki tabelaričnih podatkov (CSV, JSON ali TXT), ki vključujejo izračunane metaznačilke, vektorske predstavitve in pripadajoče metapodatke. Poseben poudarek bo namenjen tudi hrambi reprezentativnih podmnožic, identificiranih z metodologijami, razvitimi v projektu, saj te predstavljajo ključni rezultat raziskave in omogočajo bolj robustno ter manj pristransko primerjalno analizo. Poleg tega bomo hranili tudi pripadajoče podatke o uspešnosti algoritmov, ki omogočajo ponovljivost analiz in nadaljnjo uporabo rezultatov v raziskovalni skupnosti.
2.3	**Ali bodo podatki shranjeni v zaupanja vrednem repozitoriju?**	☒ Da ☐ Ne Podatki bodo shranjeni v zaupanja vrednem odprtem repozitoriju Zenodo (https://zenodo.org/), ki ga razvija in vzdržuje organizacija CERN. Zenodo je splošno priznan raziskovalni repozitorij, ki ga široko uporablja mednarodna raziskovalna skupnost ter podpira načela FAIR.

		Repozitorij Zenodo omogoča dolgoročno hrambo raziskovalnih podatkov, dodelitev trajnih identifikatorjev DOI (Digital Object Identifier), različice zapisov ter izvajanje digitalnega skrbništva in dolgoročne dostopnosti podatkov. Podatki bodo objavljeni skupaj z ustreznimi metapodatki, kar bo omogočalo njihovo sledljivost, citiranje, ponovno uporabo in preverljivost raziskovalnih rezultatov. Izvorna koda za obdelavo in analizo podatkov (GitHub) bo javno dostopna ter povezana z ustreznimi podatkovnimi zbirkami v repozitoriju Zenodo, s čimer bomo zagotovili transparentnost, ponovljivost raziskav in ponovno uporabo rezultatov. Prednatisi vseh publikacij, ki bodo nastale v okviru projekta, bodo deponirani v slovenskem odprtodostopnem repozitoriju DIRROS in na platformi arXiv, s čimer bomo zagotovili odprt dostop, dolgoročno hrambo ter široko diseminacijo raziskovalnih rezultatov.
3.	**Zagotovitev podatkov na način FAIR**	
3.1	**Zagotavljanje najdljivosti podatkov (F)**	
3.1.1	**Ali bodo podatki označeni s trajnim identifikatorjem (PID)?**	☒ Da ☐ Ne Podatki bodo označeni s trajnim identifikatorjem DOI (Digital Object Identifier), ki bo dodeljen ob objavi podatkov v zaupanja vrednem repozitoriju Zenodo oziroma ob vključitvi podatkov v znanstvene objave, objavljene v znanstvenih revijah z odprtim dostopom. DOI bo omogočal trajno, enolično in citirno povezavo do raziskovalnih podatkov ter prispeval k njihovi sledljivosti, dostopnosti in ponovni uporabi.
3.1.2	**Kateri metapodatki bodo ustvarjeni in kateri metapodatkovni standardi bodo pri tem upoštevani?**	Za vsako analizirano podatkovno zbirko za vrednotenje (benchmark) bomo ustvarili metapodatke, ki bodo omogočali sledljivost, interpretacijo in ponovno uporabo podatkov. Metapodatki bodo med drugim vključevali identifikator podatkovne instance znotraj posamezne zbirke, ime in različico podatkovne zbirke, uporabljeno metodo za predstavitev podatkov (npr. metoda za učenje predstavitev oziroma ekstrakcijo značilk), imena in opise značilk ter informacije o postopku njihovega izračuna. Pri eksperimentalnih rezultatih bomo hranili metapodatke o uporabljenih algoritmih, njihovih konfiguracijah in uporabljenih metrikah uspešnosti. Za opis podatkov in eksperimentalnih rezultatov bomo uporabljali splošno sprejete standarde metapodatkov za raziskovalne podatke, kot so načela FAIR in sheme metapodatkov, ki jih podpirajo odprti repozitoriji. Če na posameznem področju ne obstajajo ustrezni standardi za opis specifičnih podatkov, bomo pripravili lastno shemo metapodatkov, dokumentirano v datotekah README in spremljajoči dokumentaciji. Ta bo vključevala jasen opis vseh spremenljivk, značilk, metrik, eksperimentalnih nastavitev in postopkov obdelave podatkov, kar bo omogočalo popolno reprodukcijo in ponovno uporabo rezultatov.
3.1.3	**Ali bodo metapodatki vsebovali ključne besede za izboljšanje najdljivosti in možnosti ponovne uporabe?**	☒ Da ☐ Ne Metapodatki bodo vsebovali skrbno izbrane ključne besede, ki bodo opisovale vsebino podatkovnih zbirk, raziskovalno področje, uporabljene metode in eksperimentalne postopke. Ključne besede bodo izbrane na podlagi uveljavljene terminologije v znanstveni literaturi ter izrazov, ki se pogosto uporabljajo v skupnosti avtomatiziranega strojnega učenja, optimizacije in podatkovnega rudarjenja. Za posamezne podatkovne zbirke in rezultate bodo med drugim uporabljene ključne besede, kot so: benchmark dataset, representation learning, meta-features, algorithm selection, algorithm configuration, automated machine learning (AutoML), black-box optimization, performance evaluation, machine learning benchmarking, explainable AI in reproducible research. Pri izbiri ključnih besed bomo upoštevali tudi terminologijo, uporabljeno v relevantnih repozitorijih podatkov in znanstvenih objavah, s čimer bomo izboljšali najdljivost podatkov v iskalnikih, repozitorijih raziskovalnih podatkov in drugih sistemih za odkrivanje raziskovalnih virov. To bo

		omogočilo lažje prepoznavanje vsebinskega konteksta ter učinkovitejšo ponovno uporabo podatkov s strani drugih raziskovalcev.
3.2	**Zagotavljanje dostopnosti podatkov (A)**	
3.2.1	**Ali bodo vsi podatki odprto dostopni?**	☒ Da ☐ Ne
3.2.2	**Kdaj bodo podatki odprto dostopni in za koliko časa?**	Podatki bodo odprtodostopni postopoma, skladno z načelom »čim bolj odprto, ter zaprto, kjer je to nujno«. Ključni raziskovalni podatki, ki podpirajo znanstvene publikacije, bodo odprto dostopni ob objavi posamezne znanstvene publikacije. Po zaključku projekta bodo odprto dostopni vsi ustrezno obdelani in anonimizirani podatki, skupaj z metapodatki in metodološko dokumentacijo, in bodo trajno shranjeni v zaupanja vrednem repozitoriju Zenodo.
3.2.3	**Na kakšen način bo v primeru omejitev pri uporabi omogočen dostop do podatkov med izvajanjem projekta in po njegovem zaključku?**	NA
3.2.4	**Ali bo za dostop do podatkov oz. njihovo branje potrebna dodatna dokumentacija oz. informacija o ustrezni programski opremi?**	☒ Da ☐ Ne Vsak podatkovni repozitorij bo opremljen z datoteko README, v kateri bodo podrobno opisani pogoji in navodila za ponovno uporabo podatkov, vključno z opisom podatkovnih zbirk, metapodatki, formati datotek, licenčnimi pogoji ter navodili za dostop in uporabo podatkov.
3.3	**Zagotavljanje interoperabilnosti podatkov (I)**	
3.3.1	**Katere geslovnike oz. šifrante boste uporabili pri pripravi podatkov in metapodatkov?**	Pri pripravi podatkov in metapodatkov bomo uporabljali uveljavljene ontologije in nadzorovane besednjake, ki se uporabljajo na področjih optimizacije, strojnega učenja in podatkovnega rudarjenja, saj ti omogočajo boljšo interoperabilnost, izmenjavo in ponovno uporabo podatkov. Za področje enociljne optimizacije (single-objective optimization) bomo uporabljali ontologijo OPTION (OPTimization algorithm benchmarking ONtology), ki je namenjena standardiziranemu opisu optimizacijskih problemov, algoritmov, eksperimentalnih nastavitev in rezultatov vrednotenja ter je že uveljavljena v raziskovalni skupnosti. Za področje strojnega učenja in podatkovnega rudarjenja bomo uporabljali ontologijo OntoDM (Ontology of Data Mining), ki omogoča formalni opis podatkovnih zbirk, nalog strojnega učenja, algoritmov, eksperimentalnih postopkov in rezultatov. OntoDM bomo uporabili predvsem za opis podatkovnih zbirk in eksperimentov na področju klasifikacije časovnih vrst ter drugih nalog strojnega učenja, kjer je to primerno.
3.3.2	**Ali boste primorani uporabiti manj poznane ali lastne geslovnike oz. šifrante?**	☐ Da ☒ Ne
3.4	**Zagotavljanje ponovne uporabe podatkov (R)**	
3.4.1	**Na kakšen način boste zagotovili dokumentacijo, potrebno za ponovno uporabo podatkov?**	Vsak podatkovni repozitorij bo opremljen z datoteko README, v kateri bodo podrobno opisani pogoji in navodila za ponovno uporabo podatkov, vključno z opisom podatkovnih zbirk, metapodatki, formati datotek, licenčnimi pogoji ter navodili za dostop in uporabo podatkov.
3.4.2	**Ali bodo vaši podatki javno dostopni in licencirani z odprtima licencama CC BY**	☒ Da ☐ Ne

	oz. CC BY-SA, da bo s tem omogočena čim širša ponovna uporaba?	
3.4.3	**Kakšne postopke zagotavljanja kakovosti podatkov boste uporabili?**	Zagotavljanje kakovosti podatkov predstavlja osrednji raziskovalni cilj projekta. Projekt bo razvijal in vrednotil nove metodologije za zmanjševanje pristranskosti (bias) v podatkovnih zbirkah ter za izbiro reprezentativnih podmnožic podatkov, ki ohranjajo ključne lastnosti celotnih zbirk. S tem bomo izboljšali kakovost podatkov za vrednotenje algoritmov in zmanjšali tveganje za pristranske ali nereprezentativne raziskovalne ugotovitve. Kakovost podatkov bomo zagotavljali z uporabo kvantitativnih metod za analizo reprezentativnosti, raznolikosti in pokritosti podatkovnih prostorov. Razvite metode bodo temeljile na predstavitvah podatkovnih instanc, grafovskih modelih, metodah strojnega učenja ter statističnih kazalnikih za ocenjevanje podobnosti med izbranimi podmnožicami in izvirnimi podatkovnimi zbirkami. Kakovost in učinkovitost predlaganih metod bomo merili z uporabo uveljavljenih statističnih in eksperimentalnih kazalnikov, kot so ohranjanje porazdelitve značilk, pokritost prostora problemov, stabilnost rezultatov, stopnja zmanjšanja pristranskosti ter vpliv izbranih podmnožic na rezultate vrednotenja algoritmov. Poleg tega bomo izvajali neodvisne validacijske eksperimente na različnih primerjalnih podatkovnih zbirkah, da preverimo posplošljivost in robustnost razvitih pristopov. Vsi postopki obdelave podatkov, izračuna značilk in eksperimentalnega vrednotenja bodo dokumentirani ter podprti z avtomatiziranimi postopki za preverjanje konsistentnosti in reprodukcije rezultatov.
4.	**Etični in pravni vidiki**	
4.1	**Ali obstajajo etična ali pravna vprašanja, ki bi lahko vplivala na deljenje podatkov?**	☐ Da ☒ Ne
4.2	**Ali boste med izvajanjem projekta obdelovali oz. hranili osebne podatke?**	☐ Da ☒ Ne
4.3	**Ali bodo med projektom ustvarjene oz. ponovno uporabljene posebne vrste osebnih podatkov?**	☐ Da ☒ Ne
4.4	**Kako boste uredili lastništvo podatkov in morebitne avtorske pravice na podatkih, ki jih boste ustvarili ali ponovno uporabili?**	Podatki, ustvarjeni v okviru projekta, bodo v lasti nosilne institucije projekta, tj. Instituta »Jožef Stefan«. Vsi raziskovalci, vključeni v projekt, bodo imeli pravico do uporabe podatkov za izvajanje raziskovalnih aktivnosti, pripravo znanstvenih objav in diseminacijo rezultatov v skladu z internimi pravili institucije in načeli odprte znanosti. Projekt bo pretežno uporabljal javno dostopne raziskovalne podatkovne zbirke in zbirke za vrednotenje algoritmov, ki so objavljene pod ustreznimi odprtimi licencami ali pogoji uporabe. Lastništvo teh podatkov ostaja pri njihovih izvirnih avtorjih oziroma ponudnikih. Pri njihovi ponovni uporabi bodo dosledno upoštevani licenčni pogoji, zahteve glede citiranja ter morebitne omejitve uporabe. Podatki, ustvarjeni v okviru projekta, kot so izračunane meta-značilke podatkovnih zbirk, predstavitve problemov, eksperimentalni rezultati, in spremljajoči metapodatki, bodo po zaključku projekta objavljeni v ustreznih odprtih repozitorijih, kadar to ne bo v nasprotju s pogoji uporabe izvornih podatkov. Za objavljene rezultate bodo uporabljene odprte licence, ki bodo omogočale njihovo nadaljnjo uporabo, preverjanje in reprodukcijo raziskav.

5.	**Drugi raziskovalni rezultati**	
5.1	**Ali boste poleg podatkov ustvarili ali ponovno uporabili tudi druge raziskovalne rezultate?**	☐ Da ☒ Ne
6.	**Finančna sredstva**	
6.1	**Kakšni bodo stroški ravnanja s podatki in drugimi rezultati projekta po načelih FAIR in kako bodo kriti?**	Neposredni stroški bodo obsegali pripravo in strukturiranje podatkov, pripravo metapodatkov ter dokumentacijo. Del stroškov bo povezan tudi s pripravo podatkov za arhiviranje in objavo v zaupanja vrednih repozitorijih. Posredni stroški bodo zajemali uporabo obstoječe IT infrastrukture (strežniki, varnostne kopije, shranjevanje podatkov). Stroški ravnanja s podatki in drugimi raziskovalnimi rezultati v projektu DATA-TRUST bodo vključeni v okvir rednega financiranja projekta.
6.2	**Kdo bo odgovorna oseba za ravnanje z raziskovalnimi podatki pri projektu?**	Odgovorna oseba za ravnanje z raziskovalnimi podatki pri projektu je vodja projekta. Na ravni raziskovalne organizacije (IJS) še ni vzpostavljene posebne službe za ravnanje z raziskovalnimi podatki, zato bo vodja projekta neposredno odgovoren za načrtovanje, izvajanje in nadzor nad vsemi aktivnostmi upravljanja raziskovalnih podatkov. To vključuje zagotavljanje skladnosti z načeli FAIR, pripravo podatkovnega načrta, nadzor nad shranjevanjem, varnostjo in deljenjem podatkov ter koordinacijo priprave podatkov za arhiviranje in objavo v zaupanja vrednih repozitorijih. Na ravni raziskovalne organizacije (IJS) je pripravljen načrt za ravnanje z raziskovalnimi podatki, kateremu bomo sledili. Na IJS nudi podporo pri ravnanju s podatki Znanstveno-informacijski center (ZIC) IJS.

Verzija dokumenta: 1.0
Datum: 29. 5. 2026