

NAČRT ZA RAVNANJE Z RAZISKOVALNIMI PODATKI

OBRAZEC ARIS

0	Splošne informacije	
0.1	Šifra projekta	J2-70265
0.2	Naziv projekta	AI4Food: Developing and Validating a Comprehensive Food Composition Database and a Knowledge Base Aligned with FAIR Principles, Artificial Intelligence Methods, and Large Language Models
0.3	Šifra vodje projekta	10824
0.4	Ime in priimek vodje projekta	Barbara Koroušić Seljak
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 V delovnem sklopu DS1 izvajamo raziskovalno delo na podatkih o sestavi referenčnih in industrijskih živil. Poleg nacionalne baze podatkov o sestavi živil, uporabljamo tudi tuje podatkovne baze, ki so prosto dostopne in usklajene s standardom CEN/TC 387. Omenimo ameriško bazo USDA FoodData Central, nemško bazo BLS (Bundeslebensmittelschlüssel) in francosko bazo Ciqua. V delovnem sklopu DS2 uporabljamo podatke iz javno dostopnih anotiranih korpusov s področja prehrane, kot sta npr. Cafeteria-FCD in Cafeteria-SA. Korpusa smo izbrali, saj sta anotirana z različnimi ontologijami in nadzorovanimi besednjaki, kot so SNOMED CT, FoodOn in Hansard. Uporabili bomo tudi javno dostopno podatkovno zbirko Recipe1M+, ki vsebuje več kot milijon receptov z opisi sestavin, količin, navodil za pripravo, pripadajočimi podatki o makrohranilih ter oznakami prehranske varnosti in kakovosti živil na podlagi semaforkega sistema označevanja živil. V DS3 bomo razvite metode uporabili za podporo pri zapolnjevanju, obogatitvi in posodabljanju slovenske baze podatkov o sestavi živil.
1.2	Katere vrste podatkov boste ustvarili oz. ponovno uporabili in v katerih formatih bodo shranjeni?	Uporabljali in hranili bomo podatke iz baz o sestavi živil, ki so strukturirane v tabelarni obliki in skladne z uveljavljenimi standardi za opis in izmenjavo podatkov o sestavi živil. Za posamezno živilo baze vsebujejo podatke o vsebnosti hranil in drugih sestavin, skupaj z ustreznimi merskimi enotami ter spremljajočimi metapodatki. Podatki so na voljo v standardnih formatih za izmenjavo in shranjevanje podatkov, kot so CSV, TXT, JSON, XLSX in SQL. Pri raziskovalnem delu bomo uporabili tudi podatke, pridobljene z avtomatskim luščenjem informacij iz besedil. Ti podatki opisujejo živila in njihovo hranilno vrednost ter so anotirani v skladu z ustreznimi prehranskimi ontologijami. Za njihovo predstavitev bomo uporabili

		standard BioC, namenjen interoperabilni izmenjavi biomedicinskih besedilnih anotacij, podatke pa bomo shranjevali v formatih XML ali JSON. Poleg tega bomo uporabljali javno dostopno zbirko Recipe1M+, ki zajema tekstovne podatke o receptih (informacije o sestavinah, količinah ter navodilih za pripravo) v tabelarni obliki. Ti podatki so povezani s hranilnimi vrednostmi (numerični podatki) ter oznakami prehranske kakovosti po sistemu prehranskega semaforja (kategorične ordinalne vrednosti). Podatki iz Recipe1M+ so dostopni v formatih JSON in CSV. Eden izmed rezultatov projekta bodo grafovne predstavitve relacij med živili, hranili, kemijskimi spojinami in drugimi relevantnimi koncepti, ki bodo avtomatsko izluščeni iz besedil z metodami obdelave naravnega jezika. Relacije bodo predstavljene v obliki trojic <subjekt, predikat, objekt> ter shranjene v standardnih formatih CSV, TXT ali RDF (Resource Description Framework). Zbirko Recipe1M+ in anotirane korpuse bodo nadgradili s podatkovnimi nizi vprašanj in odgovorov, ki so potrebni za učenje in vrednotenje velikih jezikovnih modelov (LLM). Pri tem bodo iz obstoječih strukturiranih in besedilnih podatkov generirani pari vprašanj in odgovorov, ki bodo nadalje povezani s prehranskimi, hranilnimi in varnostnimi informacijami. Nastale podatkovne zbirke bomo shranili v formatih JSON in TXT. Nenazadnje bomo v DS3 nadgradili slovensko bazo podatkov o sestavi živil z manjkajočimi in novimi podatki. V ta namen bomo uporabili metode, ki jih razvijamo v DS1 in DS2. Podatki bodo shranjeni v formatih CSV, TXT, JSON ter SQL, s čimer bo zagotovljena njihova interoperabilnost, ponovna uporaba in dolgoročna dostopnost.
1.3	Kakšen je namen ustvarjanja, zbiranja oz. ponovne uporabe podatkov in njihova povezava s cilji projekta?	Podatkovne zbirke o sestavi živil bomo uporabili za razvoj, vrednotenje in testiranje metod za imputacijo manjkajočih vrednosti (raziskovalni cilj 1). Besedilno anotirane korpuse bomo uporabili za razvoj semantičnih virov in orodij za povezovanje različnih podatkovnih zbirk o sestavi živil ter vzpostavitev njihove interoperabilnosti (raziskovalni cilj 2). Podatkovne zbirke vprašanj in odgovorov, pripravljenih na osnovi anotiranih korpusov in podatkovne zbirke Recipe1M+, bomo uporabili za prilagajanje (angl. fine-tuning) velikih jezikovnih modelov za področje prehrane in živilstva (raziskovalni cilj 2). Slovensko bazo podatkov o sestavi živil bomo uporabili za validacijo metod, razvitih v okviru prvih dveh raziskovalnih ciljev, ter za njeno nadaljnjo nadgradnjo in izboljšanje kakovosti (raziskovalni cilj 3). Validacijo bomo izvedli tudi na anonimiziranih in ustrezno kodiranih podatkih o prehranskih navadah ter industrijskih živilih. V ta namen bomo med drugim izvedli kemijsko analizo izbranih industrijskih živil (raziskovalni cilj 4).
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 računalniške sisteme Instituta »Jožef Stefan« (IJS). Raziskovalna skupina ima dostop do visokozmogljivih računalniških virov, vključno s strežniki (576 + 576 CPU jeder), 4TB delovnega pomnilnika (RAM) ter grafičnimi procesnimi enotami (GPGPU), ki zagotavljajo do 51,4TFLOPS zmogljivosti pri dvojni natančnosti (FP64). Za shranjevanje podatkov je na voljo več kot 0,5 PB diskovnega prostora. Analitski podatki za izbrana industrijska živila bodo shranjeni na računalniški infrastrukturi Odseka za okoljske znanosti IJS. Vsi podatki bodo hranjeni znotraj omrežja IJS, 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 podatkovnih zmogljivostih bo uporabljena nacionalna infrastruktura SLING, katere član konzorcija je IJS. Nacionalna baza podatkov o sestavi živil je trenutno dostopna preko spletne aplikacije NutriBase, ki omogoča upravljanje s podatki. NutriBase z vsemi podpornimi tehnologijami gostuje v oblaku AWS, kar mu zagotavlja visoko razpoložljivost, varnost in zanesljivo delovanje. Fizična lokacija strežnikov se nahaja v Amazonovem podatkovnem centru v Frankfurtu, v Nemčiji.
2.2	Kako boste izbrali podatke za dolgoročno hrambo?	Za namene dolgoročne hrambe bomo ohranili vse podatke, uporabljene in analizirane pri razvoju metod v DS1 in DS2. To vključuje podatkovne baze o sestavi živil v formatih CSV, TXT in JSON, anotirane podatke v formatu BioC ter podatkovne množice z vprašanji in odgovori v formatih TXT in JSON. Poleg tega bomo dolgoročno hranili tudi vse rezultate vrednotenja in testiranja modelov, razvitih v DS1 in DS2, v formatih CSV in TXT, kar bo omogočilo ponovljivost raziskav ter nadaljnjo uporabo rezultatov. Izboljšana in obogatena različica slovenske podatkovne baze o sestavi živil, ki bo praktični rezultat projekta, bo na voljo v obliki relacijske podatkovne baze (SQL) ter po potrebi tudi v interoperabilnih izvoznih formatih za nadaljnjo uporabo.
2.3	Ali bodo podatki shranjeni v zaupanja vrednem repozitoriju?	☒ Da ☐ Ne Podatki, pridobljeni z raziskovalnim delom v DS1 in DS2, bodo shranjeni v odprto dostopnem repozitoriju Zenodo (https://zenodo.org/), ki ga razvija in vzdržuje organizacija CERN. Zenodo je splošno priznan raziskovalni repozitorij, ki ga uporablja mednarodna raziskovalna skupnost ter podpira t.i. FAIR principe. Repozitorij omogoča dolgoročno hrambo raziskovalnih podatkov, dodelitev trajnih identifikatorjev DOI (Digital Object Identifier), različice zapisov ter izvajanje digitalnega skrbništva in dolgoročnega dostopa do podatkov. Podatki bodo objavljeni skupaj z ustreznimi metapodatki, kar bo omogočalo njihovo sledljivost, citiranje, ponovno uporabo in preverljivost raziskovalnih rezultatov. Izbrane podatke bomo shranili tudi v slovenski bazi podatkov o sestavi živil preko spletnega orodja NutriBase. 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. Prednatis 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 citabilno povezavo do raziskovalnih podatkov ter prispeval k njihovi sledljivosti, dostopnosti in ponovni uporabi.
3.1.2	Kateri metapodatki bodo ustvarjeni in kateri	Za vse podatkovne zbirke, uporabljene v projektu, bomo pripravili ustrezne metapodatke, ki bodo omogočali sledljivost, interpretacijo, interoperabilnost in ponovno uporabo podatkov. Metapodatki bodo med

	metapodatkovni standardi bodo pri tem upoštevani?	drugim vključevali identifikatorje živil, hranil, receptov ali besedilnih dokumentov, ime in različico uporabljene podatkovne zbirke, izvor podatkov, uporabljene ontologije in nadzorovane besednjake (npr. FoodOn, SNOMED CT), merske enote ter opise posameznih atributov. Pri anotiranih korpusih bomo hranili metapodatke o uporabljenih anotacijskih shemah, ontologijah, tipih entitet in relacij ter različicah uporabljenih virov. Pri podatkovnih množicah z vprašanji in odgovori bomo dokumentirali postopek generiranja primerov, uporabljene podatkovne vire in različice modelov oziroma pravil, uporabljenih pri pripravi podatkov. Za modele, razvite v DS1 in DS2, bomo hranili metapodatke o uporabljenih algoritmihi, njihovih konfiguracijah, učnih in validacijskih postopkih, uporabljenih podatkovnih množicah ter metrikah uspešnosti. Pri metodah za imputacijo manjkajočih vrednosti, luščenje znanja, gradnjo semantičnih virov in prilagajanje velikih jezikovnih modelov bomo dokumentirali tudi uporabljene eksperimentalne nastavitve in programsko opremo. Za opis podatkov, modelov in eksperimentalnih rezultatov bomo uporabljali splošno sprejete standarde za raziskovalne podatke ter t.i. principe FAIR (Findable, Accessible, Interoperable, Reusable), s čimer bomo zagotovili dolgoročno dostopnost, interoperabilnost in ponovno uporabo rezultatov projekta. Če se bo izkazalo, da še ne obstajajo 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 nastavitvev in postopkov obdelave podatkov, kar bo omogočilo 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 prehranske informatike, obdelave naravnega jezika, umetne inteligence in podatkovne znanosti. Pri izbiri ključnih besed bomo upoštevali tudi terminologijo, uporabljeno v relevantnih ontologijah, podatkovnih repozitorijih, standardih za prehranske podatke (EuroFIR, FoodEx2) ter znanstvenih objavah s področja prehranske znanosti, semantičnih tehnologij in umetne inteligence. S tem bomo izboljšali najdljivost podatkov v iskalnikih, repozitorijih raziskovalnih podatkov in drugih sistemih za odkrivanje raziskovalnih virov. Takšen pristop bo omogočil lažje razumevanje vsebinskega konteksta podatkov ter njihovo učinkovitejšo ponovno uporabo v nadaljnjih raziskavah in aplikacijah. Predvidevamo uporabo sledečih ključnih besed v angleškem jeziku: <i>food composition database (FCDB), food data integration, food ontology, FoodOn, SNOMED CT, nutrition informatics, food knowledge graph, natural language processing (NLP), named entity recognition, relation extraction, question answering (QA), large language models (LLM), food language models, missing value imputation, semantic interoperability, food recommendation systems, nutrient prediction, food safety, recipe analysis, machine learning, explainable AI, reproducible research in FAIR data</i>. Ustrezni prevodi v slovenščino: <i>baza podatkov o sestavi živil, integracija podatkov o živilih, ontologija živil, FoodOn, SNOMED CT, prehranska informatika, graf znanja o živilih, obdelava naravnega jezika, prepoznavanje poimenovanih entitet, luščenje relacij, iskanje odgovorov na vprašanja, veliki jezikovni modeli, jezikovni modeli za področje prehrane, imputacija manjkajočih vrednosti, semantična interoperabilnost, prehranski priporočilni sistemi, napovedovanje hranilne</i>

		<i>sestave, varnost hrane, analiza receptov, strojno učenje, razločljiva umetna inteligenca, ponovljive raziskave na podatkih FAIR.</i>
3.2	Zagotavljanje dostopnosti podatkov (A)	
3.2.1	Ali bodo vsi podatki odprto dostopni?	☐ Da ☒ Ne Večina podatkov, uporabljenih za razvoj in vrednotenje metod v okviru projekta, izvira iz javno dostopnih virov, kot so mednarodne podatkovne zbirke o sestavi živil, anotirani korpusi in podatkovne zbirke receptov. Ti podatki bodo ostali dostopni prek izvornih repozitorijev, pri čemer bomo zagotovili ustrezne povezave, metapodatke in dokumentacijo za njihovo ponovno uporabo. Podatki, ustvarjeni v okviru projekta, vključno z izpeljanimi podatkovnimi množicami, semantičnimi viri, podatkovnimi množicami z vprašanji in odgovori, eksperimentalnimi rezultati ter programske kodo, bodo praviloma javno dostopni prek ustreznih odprtih repozitorijev, skladno z načeli FAIR. Izjema bo izboljšana različica slovenske baze podatkov o sestavi živil. Ta vsebuje podatke, ki izvirajo iz različnih nacionalnih virov in institucij ter so lahko predmet avtorskih pravic, licenčnih omejitev ali drugih pogojev uporabe, ki ne dovoljujejo popolne javne objave celotne podatkovne zbirke. Zato bo javno dostopen le tisti del podatkov in metapodatkov, za katerega bodo pridobljena ustrezna dovoljenja oziroma za katerega ne bodo veljale omejitve glede nadaljnje distribucije. V primerih, kjer popolna objava ne bo mogoča, bomo zagotovili podrobne metapodatke, opis metodologije ter informacije o postopkih za pridobitev dostopa do podatkov pri njihovih izvornih upravljalcih. Prav tako obstajajo omejitve glede odprtega dostopa do podatkov pridobljenih s kemijsko analizo industrijskih živil. Ti podatki so predmet omejenega dostopa zaradi zagotavljanja znanstvene integritete ter varstva občutljivih informacij. Omejitve dostopa so povezane predvsem z naslednjimi vidiki: • zaščita še neobjavljenih raziskovalnih rezultatov in podatkov, ki so pomembni za nadaljnje znanstvene objave ali morebitno zaščito intelektualne lastnine, • začasni embargo na surove eksperimentalne podatke do objave ključnih rezultatov v znanstvenih revijah; • interni delovni podatki (npr. preliminarne analize in vmesni rezultati), ki niso namenjeni javni objavi ali bi brez ustreznega konteksta lahko vodili do napačne interpretacije.
3.2.2	Kdaj bodo podatki odprto dostopni in za koliko časa?	Podatki bodo odprto dostopni postopoma, skladno z načelom »čim bolj odprto ter zaprto, kjer je to nujno«. Ključni raziskovalni podatki, ki podpirajo znanstvene publikacije, vključno z izpeljanimi podatkovnimi množicami, eksperimentalnimi rezultati, semantičnimi viri, podatkovnimi množicami z vprašanji in odgovori ter programske kodo, bodo odprto dostopni ob objavi posamezne znanstvene publikacije. Po zaključku projekta bodo vsi podatki, za katere ne veljajo pravne, licenčne ali druge omejitve glede nadaljnje distribucije, skupaj z metapodatki in metodološko dokumentacijo trajno shranjeni v zanesljivem repozitoriju Zenodo. Javna dostopnost bo zagotovljena za vse podatke, ustvarjene v okviru projekta, kadar to ne bo v nasprotju s pogoji uporabe izvornih podatkovnih virov. Izboljšana različica slovenske baze podatkov o sestavi živil morda ne bo v celoti javno dostopna, saj lahko vsebuje podatke, za katere veljajo avtorske pravice, licenčne omejitve ali drugi pogoji uporabe izvornih ponudnikov podatkov. V takšnih primerih bodo javno dostopni metapodatki, opis podatkovne zbirke, metodologija obdelave ter tisti deli podatkov, za katere bo omogočena nadaljnja distribucija skladno z veljavnimi dovoljenji.

3.2.3	Na kakšen način bo v primeru omejitev pri uporabi omogočen dostop do podatkov med izvajanjem projekta in po njegovem zaključku?	V primeru omejitev pri uporabi podatkov bo dostop urejen skladno z licenčnimi pogoji, avtorskimi pravicami in drugimi omejitvami, ki jih določajo upravljavci posameznih podatkovnih virov. Med izvajanjem projekta bodo imeli dostop do podatkov le člani projektne skupine in pooblašeni sodelavci, ki podatke potrebujejo za izvajanje raziskovalnih aktivnosti. Za podatke, ki ne bodo mogli biti javno objavljeni, bodo javno dostopni metapodatki, dokumentacija in opis postopkov obdelave, s čimer bo zagotovljena transparentnost raziskave in ponovljivost metod. Kadar bo to mogoče, bodo objavljene tudi izpeljane ali agregirane različice podatkov, ki ne bodo kršile licenčnih ali drugih pravnih omejitev. Po zaključku projekta bodo uporabniki lahko dostop do omejenih podatkov pridobili neposredno pri njihovih izvornih upravljavcih oziroma lastnikih, skladno z njihovimi pogoji uporabe. V repozitoriju bodo objavljene informacije o izvoru podatkov, morebitnih omejitvah uporabe ter postopkih za pridobitev dostopa. Za podatke, ustvarjene v okviru projekta, za katere ne bodo veljale omejitve glede distribucije, bo zagotovljen odprt dostop prek repozitorija Zenodo.
3.2.4	Ali bo za dostop do podatkov oz. njihovo branje potrebna dodatna dokumentacija oz. informacija o ustrezni programski opremitvi?	☒ 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. šifrate boste uporabili pri pripravi podatkov in metapodatkov?	Pri pripravi podatkov in metapodatkov bomo uporabljali uveljavljene ontologije, nadzorovane besednjake in standarde, ki se uporabljajo na področjih prehranske znanosti, biomedicine, semantičnega spleta in umetne inteligence. Njihova uporaba bo omogočala boljšo interoperabilnost, povezovanje heterogenih podatkovnih virov ter ponovno uporabo podatkov in rezultatov projekta. Za opis živil, sestavin in prehranskih konceptov bomo uporabljali ontologijo FoodOn, ki predstavlja enega najpomembnejših standardov za semantični opis živil, prehranskih izdelkov in procesov povezanih s prehrano. Za opis biomedicinskih konceptov bomo uporabljali terminološke vire, kot je SNOMED CT, ki omogočajo standardizirano predstavitev zdravstvenih, prehranskih in biomedicinskih pojmov. Pri razvoju semantičnih virov in grafov znanja bomo uporabljali standarde RDF (Resource Description Framework) in OWL (Web Ontology Language), ki omogočajo formalno predstavitev entitet in relacij ter njihovo vključevanje v ekosistem povezanih podatkov (Linked Data). Za predstavitev in izmenjavo relacij, izluščenih iz besedil, bomo uporabljali model <subjekt, predikat, objekt>. Pri opisovanju raziskovalnih podatkov, eksperimentov in modelov umetne inteligence bomo upoštevali uveljavljene standarde za raziskovalne metapodatke ter načela FAIR. Kjer bo primerno, bomo uporabljali tudi obstoječe besednjake in klasifikacije s področja prehrane, hranil, receptov in sestavin, ki jih uporabljajo mednarodne podatkovne zbirke o sestavi živil ter sorodni raziskovalni projekti.
3.3.2	Ali boste primorani uporabiti manj poznane ali lastne geslovnike oz. šifrate?	☐ 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 oz. CC BY-SA, da bo s tem omogočena čim širša ponovna uporaba?	☐ Da ☒ Ne Večina podatkov, ustvarjenih v okviru projekta, vključno z izpeljanimi podatkovnimi množicami, semantičnimi viri, podatkovnimi množicami z vprašanji in odgovori, eksperimentalnimi rezultati in programsko kodo, bo javno dostopna in objavljena pod odprtimi licencami, kadar bo to mogoče. Vendar vseh podatkov ne bo mogoče licencirati z odprtima licencama CC BY ali CC BY-SA, saj projekt uporablja tudi podatke iz zunanjih virov, za katere veljajo licenčne, avtorske ali druge omejitve glede nadaljnje distribucije. To velja predvsem za določene dele slovenske baze s podatki o sestavi živil in druge podatkovne vire tretjih oseb. Za takšne podatke bodo javno dostopni metapodatki, dokumentacija in informacije o postopkih za pridobitev dostopa pri njihovih izvornih upravljavcih. Kadar omejitve ne bodo obstajale, bodo podatki in programska oprema objavljeni pod ustreznimi odprtimi licencami, ki bodo omogočale njihovo ponovno uporabo skladno s principi FAIR.
3.4.3	Kakšne postopke zagotavljanja kakovosti podatkov boste uporabili?	Kakovost podatkov predstavlja pomemben vidik projekta, saj bodo razvite metode temeljile na podatkovnih zbirkah o sestavi živil, anotiranih besedilnih korpusih, podatkovnih zbirkah receptov ter semantičnih virih. Kakovost podatkov bomo zagotavljali z uporabo avtomatiziranih in ročnih postopkov preverjanja konsistentnosti, popolnosti, točnosti in interoperabilnosti podatkov. Pri podatkovnih zbirkah o sestavi živil bomo izvajali preverjanje manjkajočih vrednosti, konsistentnosti merskih enot, podvajanja zapisov ter skladnosti s prehranskimi standardi in ontologijami. Razvite metode za imputacijo manjkajočih vrednosti bomo vrednotili z uporabo standardnih statističnih kazalnikov, kot so MAE, RMSE in druge ustrezne metrike napovedne uspešnosti. Pri anotiranih korpusih bomo preverjali kakovost anotacij z uporabo uveljavljenih mer kakovosti, kot so medoznačevalsko ujemanje (inter-annotator agreement), natančnost, priklic in mera F1. Dodatno bomo preverjali skladnost anotacij z uporabljenimi ontologijami in nadzorovanimi besednjaki (npr. FoodOn in SNOMED CT). Pri gradnji semantičnih virov in grafov znanja bomo izvajali preverjanje pravilnosti relacij, podvajanja entitet, ontološke skladnosti ter povezljivosti med koncepti. Kakovost ekstrakcije znanja iz besedil bomo ocenjevali z uporabo standardnih metrik za prepoznavanje entitet in ekstrakcijo relacij. Podatkovne množice z vprašanji in odgovori, namenjene prilagajanju velikih jezikovnih modelov, bodo podvržene preverjanju kakovosti generiranih vprašanj in odgovorov, odstranjevanju podvojenih primerov ter preverjanju dejanske skladnosti odgovorov z izvornimi podatki. Učinkovitost modelov bomo vrednotili na ločenih validacijskih in testnih množicah. Vsi postopki obdelave podatkov, anotiranja, integracije podatkovnih virov, razvoja modelov in eksperimentalnega vrednotenja bodo dokumentirani ter podprti z avtomatiziranimi postopki za preverjanje konsistentnosti, sledljivosti 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?	Lastništvo nad raziskovalnimi podatki, ustvarjenimi v okviru projekta, bo v celoti pripadalo nosilni raziskovalni organizaciji, Institutu »Jožef Stefan« (IJS), ki zagotavlja raziskovalno infrastrukturo in izvaja projekt. Upravljanje podatkov bo potekalo skladno z internimi pravili IJS ter pogodbenimi obveznostmi financerja projekta (ARIS). V okviru projekta bo potekalo sodelovanje z domačimi in mednarodnimi raziskovalnimi partnerji, zato bodo pravice do uporabe podatkov, objave rezultatov in deljenja raziskovalnih izsledkov usklajene med sodelujočimi raziskovalci in institucijami. Znanstvene objave bodo pripravljene skladno z načeli odgovornega raziskovanja in soavtorstva, pri čemer bodo upoštevani dejanski prispevki posameznih raziskovalcev. Projekt bo uporabljal predvsem javno dostopne podatkovne zbirke o sestavi živil, anotirane korpuse, podatkovne zbirke receptov in druge raziskovalne vire. Lastništvo teh podatkov ostaja pri njihovih izvirnih avtorjih oziroma upravljavcih, pri njihovi ponovni uporabi pa bodo dosledno upoštevani licenčni pogoji, zahteve glede citiranja in morebitne omejitve nadaljnje distribucije. Posebno pozornost bomo namenili slovenski bazi podatkov o sestavi živil, saj lahko vključuje podatke različnih ponudnikov in institucij, za katere veljajo posebni pogoji uporabe. Lastništvo teh podatkov ostaja pri njihovih imetnikih pravic, zato bo uporaba in morebitna objava izboljšane različice podatkovne zbirke potekala skladno z veljavnimi licenčnimi pogoji, avtorskimi pravicami in drugimi pravnimi omejitvami. Pred morebitno javno objavo podatkov ali rezultatov, ki vključujejo prispevke zunanjih partnerjev ali podatke z omejitvami uporabe, bodo pridobljena ustrezna soglasja vseh relevantnih deležnikov. Po objavi bodo podatki, ustvarjeni v okviru projekta, kot so semantični viri, grafi znanja, podatkovne množice z vprašanji in odgovori, eksperimentalni rezultati, programska koda in spremljajoči metapodatki, kjer bo to mogoče, deljeni pod odprtimi licencami (npr. CC BY) v ustreznih repozitorijih. Pri tem bo zagotovljeno ustrezno navajanje Instituta »Jožef Stefan« kot nosilca projekta ter vseh sodelujočih raziskovalcev in partnerjev.
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 se nanašajo na zbiranje in urejanje podatkov, anotiranje podatkov z metapodatki ter pripravo dokumentacije. Del stroškov je povezan s pripravo podatkov za arhiviranje in njihovo objavo v zaupanja vrednih repozitorijih. Posredni stroški zajemajo uporabo obstoječe IT infrastrukture (strežniki, varnostne kopije, shranjevanje podatkov). Stroški ravnanja s podatki in ostalimi raziskovalnimi rezultati projekta AI4Food so 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, ki je 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