



Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities: LLM4DH

Izroček projekta/ Project Deliverable

Vizualizacija grafov imenskih entitet / Visualization of extracted named entity graphs

Št. Izročka / Deliverable No.	4.1.2
Datum oddaje / Delivery date	26. 9. 2026 (Verzija 1.0)
Vrsta / Nature ¹ :	R
Nivo diseminacije / Dissemination level ¹ :	PU
Delovni Sklop / Work package	WP4 Advanced Technologies for Digital Humanities T4.1 Named entity graphs
Avtor, Vodilni partner / Author, Lead partner	Matija Marolt, FRI
Sodelujoči partnerji / Contributing partners	Ciril Bohak, Uroš Šmajdek, Matija Marolt UL FRI Darja Fišer, Tina Munda, Filip Dobranič, Vojko Gorjanc, Oliver Pejić INZ

¹Nature:

R = Report, P = Prototype, D = Demonstrator, O = Other (specify)

¹Dissemination level

PU = Public

CO = Confidential, only for members of the consortium (including ARIS)

Projekt financira ARIS – Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije, št projekta GC-0002

Informacijska tabela projekta / Project information card	
Akronim projekta / Project acronym	LLM4DH
Naslov projekta / Project title	Veliki jezikovni modeli za digitalno humanistiko / Large language models for digital humanities
Spletna stran / Website	https://www.cjvt.si/llm4dh/
Št projekta / Project No.	GC-0002
Razpis / Call	Javni razpis za (so)financiranje Gravitacije, št: 5100-70/2024-6 z dne 7.5.2024 in Sprememb javnega razpisa za (so)financiranje Gravitacije, št. 5100-70/2024-11 z dne 29.5.2024.
Koodrinator projekta / Project coordinator	Marko Robnik-Šikonja
Trajanje projekta / Project duration	1.10.2025 – 30.9.2027, 36 mesecev / 36 months
Povzetek projekta / Abstract of the project	Jezikovne tehnologije igrajo vse pomembnejšo vlogo na številnih področjih znanosti. Z bliskovitim napredkom velikih jezikovnih modelov se ti pozicionirajo kot osrednja tehnologija umetne inteligence z daljnosežnim vplivom ne le na znanstveno raziskovanje ampak tudi na širši družbeni razvoj in celotno družbo. Veliki jezikovni modeli, kljub svojim velikim zmožnostim, izkazujejo vrsto pomanjkljivosti, kot so nekonsistentni in napačni odgovori, velike računske zahteve, šibko poznavanje jezikov z manj viri, neprilagojenost za nekatere pomembne naloge in domene ter slabosti v razumevanju družbe, etike in človekovih potreb. Projekt bo tehnologijo velikih jezikovnih modelov izboljšal z različnimi znanji in podatki, jim zmanjšal računsko zahtevnost pri govornih tehnologijah in jim omogočil dostop do zanesljivih zunanjih informacij. Izboljšave v temeljnih zmožnostih velikih jezikovnih modelov bodo vodile do prebojnih dosežkov na več področjih digitalne humanistike: v jezikoslovju, leksikografiji, izobraževanju, analizi družbenih pojavov in sodobne zgodovine, političnih vedah, folkloristiki in pravu. Language technologies are increasingly important in many areas of science. With the rapid advances in large-scale language models, they are positioning themselves as a core AI technology with far-reaching impact on scientific research, wider societal development, and society. Despite their great capabilities, large language models exhibit several shortcomings, such as inconsistent and incorrect answers, high computational demands, weak knowledge of less-resourced languages, poor performance in some important tasks and domains, and weaknesses in understanding society, ethics, and human needs. The project will improve the technology of large language models by enriching them with diverse knowledge and data, using them to improve the performance of speech technologies, and giving them access to reliable external information. Improvements in the core capabilities of large language models will lead to breakthroughs in several areas of the digital humanities: linguistics, lexicography, education,

	analysis of social phenomena and contemporary history, political sciences, folkloristics, and law.
Vodilni partner prprojekta/ Lead project partner	Univerza v Ljubljani, Fakulteta za računalništvo in informatiko (UL FRI):
Sodelujoči partnerji / Participating project partners	Institut "Jožef Stefan" (IJS) Inštitut za novejšo zgodovino (INZ) Inštitut IRRIS za raziskave, razvoj in strategije družbe, kulture in okolja (IRRIS) Inštitut za kriminologijo pri Pravni Fakulteti v Ljubljani (IK) Univerza v Ljubljani, Filozofska fakulteta (UL FF) Univerza v Ljubljani, Fakulteta za elektrotehniko (UL FE) Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko (UM FERi)
Opis aktivnosti 4.1.2 / Description of the Action 4.1.2	Digitalna humanistika se vse bolj osredotoča na raziskovanje omrežij oseb, organizacij, krajev in drugih entitet, ki se pojavljajo v različnih besedilnih korpusih. Obstoječe metode za prepoznavanje imenske entitet (NER) in povezovanje imenskih entitet (NEL) (Boros idr., 2020) niso ustrezno prilagojene jezikom z manj razpoložljivimi viri. To omejuje možnosti za izgradnjo celovitih grafov imenskih entitet, s katerimi bi lahko analizirali pomembne entitete in njihove povezave skozi različna časovna obdobja. Čeprav lahko analiza kompleksnih omrežij pomaga pri razvrščanju in vizualizaciji teh entitet, napredne metode vizualizacije omogočajo ohranjanje pristopa človek v zanki ter s tem zmanjšujejo težave, povezane s kakovostjo in dvoumnostjo podatkov. Z izgradnjo in vizualizacijo grafov imenskih entitet želimo izboljšati razumevanje povezav, sentimenta in časovnega dogajanja ter zagotoviti pomemben vir za slovenščino in druge jezike z manj razpoložljivimi viri. There is a notable need to investigate networks of people, organizations, places, and other entities across various corpora. Current Named Entity Recognition (NER) and Linking (NEL) methods (Boros et al., 2020) are not adequately adapted to less-resourced languages, leaving these languages underrepresented in multilingual historical corpora (Schweter et al., 2022). This lack of adaptation limits the ability to build comprehensive interaction graphs for analyzing important entities and their connections across different time periods. While complex network analysis can help rank and visualize these entities, advanced visualization methods allow us to retain a human-in-the-loop approach, alleviating issues with data quality and ambiguity. By constructing and visualizing comprehensive interaction graphs, we aim to significantly improve our understanding of relationships, sentiment, and narrative structures within these texts, providing a much-needed resource for Slovenian and other less-resourced languages.

Zgodovina dokumenta / Document history			
Datum / Date	Verzija	Avtor / Author	Komentar / Comment
26.9.2026	1.0	Uroš Šmajdek, Ciril Bohak	Izdelana različica za končni pregled
30.9.2026	1.1	Uroš Šmajdek	Dodan DOI za tehnično poročilo

Kazalo vsebine / Table of content

1. CILJI	5
2. REZULTATI	5
3. PUBLIKACIJE / PUBLICATIONS	6

1. Cilji

Cilj 4.1.2 je izgraditi celovite grafe odnosov, ki omogočajo analizo povezav med entitetami skozi različna časovna obdobja, ter razviti napredne metode vizualizacije za raziskovanje in interpretacijo kompleksnih omrežij ob učinkovitem vključevanju raziskovalca v proces analize. Posebno pozornost smo namenili izboljšanju kakovosti in zanesljivosti samodejno pridobljenih podatkov z interaktivnim preverjanjem in dopolnjevanjem rezultatov. Končni cilj je omogočiti analizo odnosov med entitetami, čustvene naravnosti in pripovednih struktur v zgodovinskih in sodobnih časopisih, parlamentarnih prepisih ter literarnih besedilih ter ustvariti raziskovalni vir za nadaljnje raziskave slovenščine in drugih jezikov z manj razpoložljivimi viri. V okviru cilja smo primerjalno ovrednotili tudi metode za prepoznavanje imenskih entitet in z njimi povezane podporne tehnologije na zgodovinskih slovenskih besedilih.

2. Rezultati

Pripravili smo zbirko ročno označenih imenskih entitet v zgodovinskih parlamentarnih dokumentih zbirke Kranjska 1.0. Objavljena bo v okviru zbirke Kranjska 2.0, ki bo prav tako vsebovala samodejno označene imenske entitete s pomočjo jezikovnega modela Gemma 4.

Na podlagi označenih podatkov smo razvili interaktivno vizualizacijo grafov imenskih entitet, ki prikazuje omrežja ljudi, organizacij, krajev in drugih entitet ter njihove medsebojne povezave. To vizualizacijo smo, prek povratnih informacij s strani domenskih raziskovalcev, razširili z vizualizacijo negotovosti označb, tako na nivoju entitete in njenih atributov kot povezav. Razširjen sistem prav tako omogoča večkratne povezave kot simulacijo entitet z več interpretacijami, urejanja podatkov, dodajanja novih vnosov in naprednejše funkcije filtriranja.

V okviru novega sistema za vizualizacijo pripravljamo članek z naslovom *EntityWeaver: Visual Exploration and Curation of Named-Entity Relationships in Document Collections*, ki potrebuje še poglobljeno evalvacijo.

Prav tako smo izvedli poglobljeno evalvacijo preverili učinkovitost obstoječih velikih jezikovnih modelov za prepoznavanje imenskih entitet, sentimenta na nivoju kolektivnih identitet na izbranih zgodovinskih korpusih in zaznavanja dogodkov v slovenskih zbirkah. Pokazalo se je, da se ti modeli na zgodovinskih podatkih obnesejo precej slabše kot na novejših, še posebej pri zaznavanju entitet, ki označujejo organizacije.

Prav tako smo v sodelovanju z Inštitutom za kulturno zgodovino pri ZRC SAZU pripravili predlog prenove njihovega digitalnega biografskega leksikona PZSBL¹, tako da bo ta omogočal vizualizacijo in poglobljeno analizo omrežij ljudi in krajev, prav tako pa tudi sledenje provenience informacij in analize virov. V nadaljnjem načrtujemo implementacijo novega sistema in podpornega vizualizacijskega

¹ Podatkovna zbirka Slovenskega biografskega leksikona / Database of the Slovenian Biographical Lexicon

sistema, ki bo omogočal interaktiven pregled, vnos, analizo in popravljanje informacij o entitetah in virih. Trenutni napredek smo predstavili na simpoziju *Prozopografija – pristopi, besedila in tehnologije*.

3. Publikacije / Publications

Konferenčni prispevki / Conference publications:

Šmajdek, U., & Bohak, C. (2025). NERVIS : an interactive system for graph-based exploration and editing of named entities. HCI SI Koper 2025 : Proceedings of the 10th Human-Computer Interaction Slovenia Conference : University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Koper, Slovenia, 13 October 2025, 147–156. <https://doi.org/10.26493/978-961-293-559-7.14>

Uroš Šmajdek, Tina Munda & Matija Marolt. (2026). Low-Resource Historical Named Entity Recognition in the Era of Large Language Models. JT-DH 2026

Dobranić, F., Šmajdek, U., Pejić, O., Bohak, C., Gorjanc, V., Munda, T., & Fišer, D. (2026). Thematic landscapes of the past : analysing Slovene historical periodicals with topic modelling. LREC 2026 : First Workshop on Creating Interoperable Corpora of Historical Newspapers : Workshop Proceedings, 92–99. <https://doi.org/10.63317/263a2ep7iiyv>

Munda, T., Dobranić, F., Šmajdek, U., Pejić, O., Bohak, C., Gorjanc, V., & Fišer, D. (2026). Benchmarking LLMs for aspect-based sentiment classification in Slovene historical periodicals. LREC 2026 : Shaping Multilingual, Multimodal AI for the Social Sciences and Humanities (LLMs4SSH) @ LREC 2026 : Workshop Proceedings, 103–113. <https://doi.org/10.63317/22hvcbc23rts>

Vojko Gorjanc, Filip Dobranić, Uroš Šmajdek, Oliver Pejić, Tina Munda & Darja Fišer. (2026). Combining Graph Visualisation and Critical Discourse Analysis in the Study of Slovene Historical Newspapers. JT-DH 2026.

Tina Munda, Oliver Pejić, Vojko Gorjanc, Uroš Šmajdek, Filip Dobranić & Darja Fišer. (2026). Evaluative Framing of Collective Identities in Leading Slovene Newspapers from the Late Habsburg Empire. JT-DH 2026.

Ahac Meden, Uroš Šmajdek. (2026). Defining Data Uncertainty in the PZSBL. *Prozopografija – pristopi, besedila in tehnologije*.

Tehnična poročila:

Šmajdek, U., & Bohak, C. (2026). EntityWeaver: Visual Exploration and Curation of Named-Entity Relationships in Document Collections. <https://arxiv.org/abs/2609.37329>