Smernice Janes-NER za označevanje imenskih entitet v slovenskem jeziku
V1.1, 2017-09-18
Katja Zupan, Nikola Ljubešić, Tomaž Erjavec

[Spremembe glede na V1.0: dodano – v tabeli podkategorija Omembe oz. sklici na Twitterju, v razdelku 4 točka 1.4 in v 5.2 “ključniki”]

Ločimo 5 kategorij imenskih entitet:

* **oseba** (ang. person), oznaka PER
* **izpeljano iz osebe** (ang. person derivative), oznaka DERIV-PER
* **lokacija** (ang. location), oznaka LOC
* **organizacija** (ang. organization), oznaka ORG
* **drugo** (ang. miscellaneous), oznaka MISC

Gre za standardne tipe imenskih entitet, izjema je le DERIV-PER, ki smo jo dodali za označitev pridevnikov, izpeljanih iz osebnega lastnega imena, kar bo omogočilo boljšo anonimizacijo osebnih podatkov.

# KAJ OZNAČUJEMO KOT IMENSKO ENTITETO (IE)?

Osnovni princip: **samostalniki in samostalniške besedne zveze**, ki identificirajo neko osebo, lokacijo, organizacijo ali drug edinstven objekt v realnem prostoru in času. Poleg tega v širšem smislu označujemo tudi **svojilne pridevnike**, izpeljane iz osebnega lastnega imena, ki se nanašajo na dotično osebo, npr. DERIV-PER[Obamova] izvolitev.

Na ortografski ravni so pogosto izražene z **veliko začetnico** (Slovenska tiskovna agencija) ali **kratico** (STA), vendar pa velika začetnica in kratica ne označujeta samo imenskih entitet (npr. BDP). Kot IE obravnavamo tudi lastnoimenske besedne zveze, katerih zapis (z veliko oz. malo začetnico) variira, npr. ORG[ministrstvo za kulturo] = ORG[Ministrstvo za kulturo Republike Slovenije]. Občeimenskih poimenovanj, npr. ‘kulturno ministrstvo’, ne označujemo kot IE.

# 1 OSNOVNE USMERITVE

1. Posamezni IE pripišemo samo eno kategorijo.
2. Smiselno preferiramo daljše enote, npr. LOC[Jamova 39, 1000 Ljubljana] in ne LOC[Jamova 39], LOC[Ljubljana].
3. Posebej ne označujemo (ne gnezdimo) morebitnih samostojnih sestavin entitete, označimo samo hierarhično najvišjo enoto, npr. ORG[Ministrstvo za kulturo Republike Slovenije] in ne ORG[Ministrstvo za kulturo LOC[Republike Slovenije]].
4. Če je prvi del imenske entitete občno ime (kot npr. letališče, mejni prehod, občina, osnovna šola, zavod), ga označimo *samo* v primeru, da je rabljeno kot del uradnega poimenovanja (torej bi se po veljavnem pravopisu lahko zapisalo tudi z veliko začetnico) in da jasno ločuje med dvema IE, npr. lokacija prometne infrastrukture vs. lokacija kraja, naselja: Peljem se na LOC[letališče/Letališče Portorož] vs. Počitnikujem v LOC[Portorožu].

# 2 KATERO KATEGORIJO UPORABITI?

1. Glavna usmeritev je **Tipologija imenskih entitet**, podana v 3. razdelku. Vsebuje 5 kategorij s podkategorijami, ki podrobneje opredeljujejo, katere IE obsega posamezna kategorija.
2. Kjer je možna interpretacija v dve različni kategoriji, se običajno odločimo za privzeto oz. primarno kategorijo, tj. države so v vseh primerih kategorizirane kot lokacije, tudi če so omenjene kot organizacije, npr. Pristal je v LOC[Sloveniji] = LOC[Slovenija] je zaprosila za članstvo v Natu, institucije pa so vedno obravnavane kot organizacije, tudi če so omenjene kot lokacije, npr. Državno tekmovanje je potekalo na ORG[osnovni šoli Franceta Prešerna].
3. Če se je težko odločiti, kateri pomen je primarni, entiteti pripišemo pomen, ki ga izraža kontekst, npr. ko je literarni lik (PERS) tudi naslov literarnega dela (OTH): PERS[Faust] išče smisel življenja vs. OTH[Faust] je dramska pesnitev).
4. Če za določeno entiteto ni nobene primerne (pod)kategorije, a gre nedvomno za imensko entiteto, jo uvrstimo v kategorijo Drugo (MISC).
5. Če je IE kljub temu pomensko nedvomno PER, ORG ali LOC (npr. Indijska podcelina), jo uvrstimo v najustreznejšo kategorijo.

# 3 TIPOLOGIJA IMENSKIH ENTITET

|  |  |  |  |
| --- | --- | --- | --- |
| **Kategorija** | **Podkategorija** | **Primeri** | **Ne spada v to kategorijo**  |
| **PER** | Oseba (ime in/ali priimek) | Janez Novak, da Vinci, Ludvik XIV. | dr., gospa, sv. |
| Ime domače živali | Fifi |   |
| Umetniško ime, psevdonim | Madonna, mati Terez(ij)a, Banksy |   |
| Imena fiktivnih oseb (iz filmov, knjig ipd.) | Ana Karenina, Rdeča kapica |   |
| Vzdevki | (Boštjan Gorenc -) Pižama, Zvezdica89 |   |
|  | Poimenovane skupine ljudi (družina ali lokalno omejena skupnost) | Angleži, Nemec, Ljubljančan; Novakovi |  |
|  | Omembe oz. sklici na Twitterju | @pizama, @Nike |  |
| **DERIV-PER** | Svojilni pridevnik iz osebnega imena  | Novakov (pes) | Alzheimerjeva (bolezen) |
| **ORG** | Organizacije | EU, Nato, Rimskokatoliška cerkev |  parlament, vlada |
| Podjetja  | Microsoft, Pasadena d.o.o. |   |
| Upravljalci letališč | Aerodrom Ljubljana | Letališče Jožeta Pučnika |
| Izobraževalne ustanove  | Filozofska fakulteta |   |
| Instituti  |  Institut “Jožef Stefan” |   |
| Muzeji, knjižnice |  Prirodoslovni muzej |   |
| Gledališča, kinematografi ipd. |  MGL, Kinodvor |   |
| Mediji (TV, radio, časopisi) |  Dnevnik, Delo, Radio Center |   |
| Restavracije, hoteli, lokali |  Kavarna Zvezda, [hH]otel Lev |   |
| Zdravstvene ustanove  |  [zZ]dravstveni dom Ribnica |   |
| Glasbene skupine |  U2, Beatli, [aA]nsambel Avsenik |   |
| Institucije  |  [oO]bčina Piran, NPK |   |
| Politične stranke in druga civilna združenja |  DeSUS, Zveza potrošnikov Slovenije |   |
| Športni klubi, društva in združenja |  (HDD SIJ) Acroni Jesenice, (FC) Barcelona |   |
| Kulturne organizacije (tudi amaterske) |  [mM]ešani pevski zbor Divača |   |
| **LOC** | Nebesna telesa (planeti ipd.) |  Mars, Andromeda, Halleyjev komet |   |
| Celine |  Južna Amerika |   |
| Države, dežele (pretekle in sedanje) |  Slovenija, Združene države (Amerike) | EU |
| Regije |  Primorska, Valonija, Nova Anglija |   |
| Mesta in predeli mest, kraji in deli krajev |  Ljubljana, Šiška, Vrhnika, Na klancu |   |
| Ulice, trgi |  Jamova cesta 39 |   |
| Nakupovalna središča |  Citypark, Supernova |   |
| Letališča  |  Letališče Jožeta Pučnika |   |
| Cerkve (kot poimenovane stavbe) |  [cC]erkev sv. Nikolaja |  Rimskokatoliška cerkev |
| Krajevne znamenitosti (kulturne, naravne) |  Tromostovje, Triglavski narodni park |   |
| Druge poimenovane zgradbe (brez org. strukture) | [kK]ulturni dom Ljubno, WTC 2 |  Cankarjev dom (ima org. strukturo, npr. direktorja) |
| Gore, jezera, reke in drugi poimenovani geografski objekti |  Triglav, Blejsko jezero, Sava, Logarska dolina |   |
| **MISC** | Sistemi, programi, aplikacije |  Windows 10, Word, Android 5.1 Lollipop |  .docx, pdf, OCR |
| Naslovi knjig, filmov, nanizank, slik in drugih umetniških del; naslovi dokumentov ipd. |  Vojna in mir, Ko jagenjčki obmolknejo, Sopranovi, Guernica; Uradni list RS |   |
| Registrirana imena ali modeli naprav (avti, mobiteli, računalniki, igre ipd.) in drugi komercialni izdelki (znamke) | Galaxy Note 7, Nokia Lumia 950, Toyota RAV4, Minecraft, Človek ne jezi se |   |
| Imena prireditev in drugih dogodkov  | Oskarji, Zlata lisica, 10. mednarodna konferenca Jezikovne tehnologije  |  shod nacifašistov |
| Imena projektov  | Obzorje 2020 |   |
| Borzni indeksi | SBI20,  Dow Jones, Nasdaq |   |

# 4 DODATNE USMERITVE PO POSAMEZNIH KATEGORIJAH

1. PER
	1. Ne označujemo funkcij ter strokovnih, častnih ipd. nazivov pred imeni.
	2. Začetnice imena in/ali priimka ter tujejezične predloge v priimkih (e.g. van, da, von) pa obravnavamo kot del IE.
	3. Lastnoimenska poimenovanja skupin ljudi, povezanih v družino ali širšo krajevno skupnost (npr. Novakovi; Slovenci, Korošci, Angleži, Londončani), se tudi uvrščajo med PER. Pridevniki so del IE le, če so izpeljani iz drugega lastnega imena in kot taki zapisani z veliko začetnico, npr. Beneški Slovenci, ne pa tudi koroški Slovenci.
	4. Omembe oz. sklice na uporabniške račune na Twitterju (@XYZ) v vsakem primeru označimo kot PER, tudi če se nanašajo na organizacije.
2. DERIV-PER
	1. Ta kategorija vključuje svojilne pridevnike v širšem smislu, vključno z opisnim nanašanjem na osebo, kadar to “razkriva” kakršne koli informacije o tej osebi (ne le svojine), tudi če je fiktivna, npr. PER[Faustovo] mnenje. Če pa je neka predmetnost (bolezni, nagrade ipd.) samo poimenovana po določeni osebi in se ne nanaša na dotično osebo (npr. Alzheimerjeva bolezen kot vrsta bolezni in ne bolezen osebe po imenu Alzheimer), je ne označimo kot izpeljano IE.
3. LOC
	1. Ključni identifikator lokacij so *entitete, ki so naravno ali umetno umeščene v prostor in so kot take tipično označene na zemljevidih*.
	2. Države kot geopolitične enote vedno obravnavamo kot lokacije, da se izognemo težavam pri interpretaciji pri geopolitičnih kontekstih in metonimičnih pomenih.
	3. Če se športno društvo, šola ipd. imenuje po kraju, lahko izjemoma uporabimo oznako ORG, npr. ORG[Barcelona] je premagala ORG[Madrid] in ORG[Bayern].
4. ORG
	1. Ključni identifikator organizacije je *prisotnost neke organizacijske strukture, tj. ljudi, ki uradno vodijo/upravljajo organizacijo*.
	2. Kljub temu pa občnoimenskih poimenovanj posameznih organov v organizaciji ne vključimo kot del IE, ker ne ločujejo dveh različnih organizacij, npr. uprava ORG[Gorenja], občinski svet ORG [občine Piran], razen če so zapisani z veliko začetnico, npr. ORG[Oddelek za anglistiko Filozofske fakultete].
	3. Preferiramo uradno poimenovanje organizacije. Če v zapisu ni velike začetnice (npr. ORG[evropska komisija]) ali je del poimenovanja izpuščen, npr. ORG[ministrstvo za kulturo] namesto celotnega naziva ORG[ministrstvo za kulturo Republike Slovenije], jo še vedno lahko obravnavamo kot imensko entiteto, če je referiranje nanjo lastnoimensko. Če je referiranje opisno ali dvoumno, kar se tipično kaže v “popridevljenju”, pa je ne označujemo (npr. kulturno ministrstvo, piranska občina).
5. MISC

 5.1. Latinska poimenovanja rastlin in delov telesa niso IE.

5.2. URL-naslovi, e-poštni naslovi in ključniki (ang. hashtags) na Twitterju niso IE.

# LITERATURA

* Marc Reznicek: **Linguistische Annotation von Nichtstandardvarietäten** — Guidelines und „Best Practices" Guidelines NER (version 1.5): <https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5>
* **MUC-6 Named Entity Task Definition**: <http://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html>
* **CONLL 2003**: <http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt>
* **BSNLP 2017 shared task:** <http://bsnlp-2017.cs.helsinki.fi/shared_task.html>