 | DOI: https://doi.org/10.3986/9789610508861_01

Ajda Pretnar Žagar

# Digitalne tehnologije v folkloristiki
# Digital technologies in folklore studies

Članek pregleda digitalne tehnologije, ki podpirajo folklorne raziskave, tako da sledi celotnemu ciklu podatkov. Cilj je navdihniti začetnike v digitalni humanistiki in računalniški folkloristiki s predstavitvijo ključnih tehnologij, ki pomagajo digitalnemu raziskovanju. Članek opiše proces zbiranja podatkov, predobdelave, analize in shranjevanja. Zbiranje podatkov vključuje tradicionalno terensko delo, zbiranje obstoječih podatkov s spleta ali preko specializiranih repozitorijev. Koraki zbiranja podatkov so predstavljeni in pojasnjeni v načrtu upravljanja podatkov. Nato se podatki predobdelajo z digitalizacijo analogne vsebine z optičnim prepoznavanjem znakov in pretvorbo govora v besedilo. Predstavljena sta formata TEI XML in CoNNL-U za shranjevanje jezikovnih podatkov ter naštete ključne tehnike obdelave naravnega jezika, od katerih jih je mnogo na voljo za slovenščino. Računalniške metode organizirajo digitalizirane podatke, pri čemer se uporabljajo statistične analize in strojno učenje za označevanje, segmentacijo in semantično analizo. Predstavimo brezplačno dostopna spletna orodja za analizo podatkov, ki segajo od manj tehničnih konkordančnikov do bolj naprednih orodij strojnega učenja. Nazadnje obravnavamo shranjevanje podatkov ter naštejemo ključne repozitorije za jezikovne podatke. Kljub naprednim orodjem ostaja človeško razumevanje ključno za kontekstualizacijo ugotovitev in določanje pomena pri raziskovanju folklore.

• **Ključne besede**: digitalna folkloristika, življenjski cikel podatkov, digitalne tehnologije, računalniška folkloristika

This article reviews the digital technologies that support folklore research by tracing the full data cycle. The aim is to inspire beginners in digital humanities and computational folklore studies by presenting key technologies that support digital research. The article describes the process of data collection, pre-processing, analysis and storage. Data collection involves traditional fieldwork, collecting existing data from the web or via specialised repositories. The steps of data collection are presented and explained in the data management plan. The data are then pre-processed by digitising the analogue content through optical character recognition and speech-to-text conversion. The TEI XML and CoNNL-U formats for storing linguistic data are presented and key natural language processing techniques are listed, many of which are available for Slovene. Computational methods organise the digitised data, using statistical analysis and machine learning for labelling, segmentation and semantic analysis. We present freely available online tools for data analysis, ranging from less technical concordancers to more advanced machine learning tools. Finally, we discuss data storage and list key repositories for linguistic data. Despite advanced tools, human understanding remains key to contextualising findings and determining meaning in folklore research.

• **Keywords**: digital folklore, data lifecycle, digital technologies, computational folklore

## 1 Uvod

Vsako znanstveno raziskovanje temelji na podatkih, bodisi kvalitativnih, kvantitativnih ali nekaj vmes. V folkloristiki zajemajo raziskovalni podatki besedila, avdio posnetke, slike, video posnetke in materialno kulturo ljudskih pravljic, zgodb, šal in drugih primerov ljudskega izražanja. Tipičen cikel podatkov v sodobnem raziskovanju vključuje zbiranje podatkov, predobdelavo (pripravo), analizo in raziskovanje, interpretacijo ter ohranjanje. Veliko teh korakov je bilo nekoč opravljenih ročno, medtem ko jih danes večinoma podpirajo sodobne strojne in programske rešitve. Zvok se ne snema več na trakove, ampak v digitalni obliki. Posnetki se ne prepisujejo več ročno, ampak s pomočjo orodij na osnovi umetne inteligence. Podatki se ne štejejo več ročno, ampak se lahko interaktivno prikažejo. Nazadnje, podatki ne ležijo več (samo) v fizični obliki v arhivu, ampak so shranjeni digitalno.

Podatke običajno pridobivamo z organiziranimi raziskovalnimi procesi (digitaliziranimi ali ne). Vendar to ni edini vir podatkov, ki je na voljo za raziskovanje folklore. V zadnjih letih se uveljavlja digitalna folkloristika. Ta veja folkloristike se osredotoča na raziskovanje spletnega ljudskega izražanja in vključuje študije spletnih šal, memov, GIF-ov, internetnih legend in mitov (Shifman 2013; Chiaro 2018; Yogarajah 2022). Prednost digitalne folkloristike je dostopnost podatkov, ki jih je mogoče enostavno zajeti ali prenesti s spleta. Takšni podatki se imenujejo »podatki iz divjine«, če si izposodimo izraz Susan Halford (Halford 2017; Halford opozarja, da izraz prinaša svoj niz težav). V razdelku 2 se posvetimo pridobivanju podatkov ter načrtu za upravljanje podatkov v digitalni folkloristiki.

Programska orodja zahtevajo, da so podatki na voljo v digitalni obliki. Vsi podatki, avdio in video posnetki, slike in besedilo, so lahko v analogni ali digitalni obliki (zvočni trak vs. zvočna datoteka, papirnata fotografija vs. digitalna fotografija, papir vs. besedilna datoteka). Pretvorba prvih v drugo ni nič kaj enostavna. Dva običajna pristopa za digitalizacijo analognih besedilnih in zvočnih podatkov sta optično prepoznavanje znakov (OCR) in tehnologije pretvorbe govora v besedilo (STT).[1] V razdelku 3 so predstavljeni sodobni pristopi k digitalizaciji folklornih podatkov ter tehnike za obdelavo naravnega jezika.

Folkloristika je po svoji naravi interdisciplinarna veda. Zato je ena najmočnejših članic na področju digitalne humanistike, ki uporablja računalniške pristope za študije humanističnih podatkov. Računalniško raziskovanje folklore sega od ustvarjanja zbirk ljudskih pravljic (Karsdorp idr. 2015; Meder idr. 2023) do uporabe statistike in strojnega učenja za identifikacijo motivov (Thuillard idr. 2018; Eklund idr. 2023). V razdelku 4

[1] Tudi predmeti se vse bolj digitalizirajo s 3D tehnologijo, kar raziskovalcem omogoča študij predmetov na daljavo.

opišemo orodja za računalniško-podprto analizo besedilnih podatkov, od enostavnejših konkordančnikov do orodij na osnovi umetne inteligence.

Shranjevanje podatkov je pomemben sestavni del življenjskega cikla podatkov. Spletne zbirke podatkov spodbujajo ponovno uporabo, primerjalno raziskovanje in razvoj novih analitičnih pristopov. Na voljo je vse več skrbno urejenih zbirk folklore in kulturne dediščine. V razdelku 5 obravnavamo FAIR izhodišča za shranjevanje podatkov, na katerih so osnovane odprte spletne zbirke za raziskovanje folklore. V članku je poudarek na folkloristiki, vendar večina pristopov velja za humanistiko na splošno. Članek obravnava vsako stopnjo cikla podatkov in podaja primere digitalnih tehnologij, ki jih podpirajo. Kjer je mogoče, so navedena najnovejša orodja za slovenski jezik.

## 2 Stopnja 1: Zbiranje podatkov

Folklorni podatki vključujejo besedila, avdio in video posnetke, slike ter predmete. Predmeti so nekoliko posebni, zato jih bomo izpustili iz pregleda, čeprav sodobni pristopi tridimenzionalnega modeliranja omogočajo digitalizacijo materialne dediščine (Kingsland 2020). V grobem obstajajo trije pristopi k zbiranju podatkov v folkloristiki: s terenskim delom, z uporabo obstoječih virov in preko spleta. Tudi pri terenskem delu se podatki digitalno beležijo v avdio (ali video) formatu. Avdio in video posnetke je mogoče analizirati v njihovi izvirni obliki ali pretvoriti v besedilo, kar bomo obravnavali v naslednjem razdelku. Druga možnost, še posebej v digitalni folkloristiki, je zbiranje podatkov na spletu. Danes obstaja veliko spletnih repozitorijev, ki vsebujejo takoj dostopne strukturirane folklorne podatke (Babič idr. 2022, Fišer idr. 2018, Verdonik in Zwitter Vitez 2011).

Spletni podatki obsegajo razprave na forumih, objave na družbenih omrežjih, bloge in vloge, klepetalnice, niti na Redditu, meme, komentarje in druge vrste spletnih podatkov. Ang idr. (2013) temu pravijo »najdeni podatki«, medtem ko Krawczyk-Wasilewska (2016) to imenuje »e-folklora«. Za Susan Halford (2017) gre za t. i. »podatke iz divjine«, čeprav opozarja, da izraz prinaša svoj niz težav. Glavna značilnost takih podatkov je, da niso bili zbrani z določenim raziskovalnim vprašanjem (ali vprašanji) v mislih.

Takšni podatki še vedno lahko predstavljajo vrednost za raziskovanje folklore. Vsebujejo odgovore na določena vprašanja, na primer, »Kaj je ljudem smešno na spletu?«, »Kako se s širjenjem po spletu spreminjajo spletne pripovedi?« ali »Kateri motivi obstajajo zgolj v internetnih skupnostih?«. So takoj dostopni, še posebej za raziskovalce z nekaj programerskimi spretnostmi. So izvorno digitalizirani in zato je po njih enostavno iskati. Vendar pa imajo spletni podatki več težav, med drugim pristranskost vzorca, zanesljivost in lastništvo.

Do pristranskosti vzorca pride, ko podatki (vzorec) ne predstavljajo populacije, kar pogosto velja za spletne podatke. Nekateri ljudje ne sodelujejo v spletnih razpravah

in zato niso vključeni v takšne študije. Čeprav bi lahko trdili, da le-ti posledično niso populacija študije, je to treba izrecno navesti v raziskovalnem načrtu in upoštevati pri analizi. Poleg tega je zelo težko določiti demografsko porazdelitev vzorca. Nekateri deli populacije so lahko pod- ali preveč zastopani v izbranem podatkovnem vzorcu.

Spletni podatki imajo vprašljivo lastništvo. Hkrati so predmet avtorskih pravic, zaščite zasebnosti in pogojev uporabe. Ko so podatki »najdeni« na spletu, niso podani z obrazcem za soglasje in udeleženci niso seznanjeni, da sodelujejo v raziskavi. Ne morejo se umakniti iz raziskave, preprosto zato, ker bi bilo preverjanje vsakega subjekta logistično nemogoče. Čeprav so spletni podatki javni podatki, če je za dostop do njih potreben račun, je javna narava podatkov že postavljena pod vprašaj, saj jih lahko vidijo le člani določene skupnosti. Spletni podatki so torej podvrženi istim etičnim merilom kot terenski podatki - ohranjanju človeškega dostojanstva, zaščiti udeležencev raziskave, maksimiranju koristi, minimiziranju tveganj ter zagotavljanju spoštovanja in pravičnosti (Halford 2017).

Faza zbiranja podatkov mora vključevati načrt upravljanja podatkov (NUP). Sestavljanje takega načrta lahko raziskovalcem zagotovi, da se bo s podatki učinkovito upravljalo, jih ohranjalo in delilo, kar poveča njihovo vrednost in vpliv. Načrt upravljanja podatkov opredeli vrsto podatkov, ki bodo zbrani (tj. format, obseg in standarde), kako se bodo podatki zbirali (tj. metode in orodja), kako se bodo podatki opisali (tj. metapodatkovne standarde), kje se bodo shranili (tj. varnostne kopije in varnost podatkov) ter kako se bodo delili (tj. repozitoriji ali arhivi). Prav tako načrt opredeli lastniške pravice in licence, etične in pravne vidike glede podatkov ter kdo je odgovoren za kuriranje podatkov.

Načrt za upravljanje podatkov je na prvi pogled videti zastrašujoč, vendar je lahko izredno koristen za raziskovalca, saj mora ta upoštevati vse vidike rokovanja s podatki. Hkrati taki načrti postajajo vseprisotna dobra praksa v mnogih raziskovalnih disciplinah, vključno s folkloristiko. Kung (2022), na primer, predlaga zaokrožen NUP za jezikovne podatke, obstajajo pa tudi specializirane metapodatkovne sheme, ki obravnavajo heterogenost podatkov v folkloristiki (Lourdi idr. 2007).

## 3 Stopnja 2: Predobdelava podatkov

Terensko delo v folkloristiki običajno privede do zbirke avdio in video datotek. Če se datoteke nanašajo na pesmi in plesne predstave, se lahko datoteke analizirajo neposredno z uporabo tehnik avdio segmentacije (Marolt idr. 2019), pridobivanja glasbene informacije (Wiering idr. 2009), prepoznavanja slik (Lee 2022) ali multimodalne analize (Smits in Wevers 2023). Vendar pa, če nas zanima vsebina, morajo biti datoteke digitalizirane v zaporedje znakov, drugače povedano, v besedilo.

### 3.1 Avtomatska transkripcija

Besedilo je eden najpogostejših tipov podatkov v digitalni humanistiki. Običajno, ko je ljudski izraz posnet, se avdio datoteka transkribira in postane besedilo. Tehnologija pretvorbe govora v besedilo (STT) je prebojna za sodobno folkloristiko, saj prihrani dragocen čas pri transkripciji. Avtomatsko prepoznavanje govora (ASR) je model strojnega učenja, ki je usposobljen za prepoznavanje besed in fraz v govorjenem jeziku ter njihovo pretvorbo v besedilo.

Priljubljeni modeli STT vključujejo zasebne Google DeepSpeech, Microsoft Azure Speech to Text in IBM Watson Speech to Text. Trije odprtokodni modeli so Kaldi (Povey idr. 2011), Whisper (Radford idr. 2023) in Slovene Conformer[2] (Lebar Bajec idr. 2022). Whisper, Slovene Conformer, Azure Speech to Text in DeepSpeech podpirajo slovenščino. Čeprav so sodobni modeli STT nepopolni, imajo povprečno 12-odstotno stopnjo napake pri besedah (Radford idr. 2023), kar običajno zagotavlja uporabno začetno transkripcijo. Služijo lahko kot začetna točka za raziskovalca, ki popravlja napake, namesto da bi celotno besedilo vpisoval ročno.

Pretvorba govora v besedilo je močno odvisna od kakovosti podatkov. Natančnost transkripcijskega modela se zmanjšuje z nižjo akustično kakovostjo, hrupom v ozadju, prekrivanjem govora, besedami zunaj slovarja in narečji ter mešanjem jezikov. Nižja akustična kakovost je običajno prisotna pri starejših posnetkih, medtem ko je hrup v ozadju pogost pri neformalno posnetem gradivu (npr. v lokalu, na festivalu ali na zabavi). Prekrivanje govora se nanaša na problem določanja govorcev, imenovan tudi diarizacija govorca (Park idr. 2022). Pred poskusom analize je treba avdio posnetek razdeliti po govorcih, da model STT pravilno deluje. Modeli imajo zmanjšano točnost tudi zaradi slabe učinkovitosti pri besedah izven slovarja in narečjih. Modeli so naučeni na določenem naboru podatkov, ki vsebuje omejeno (čeprav veliko) število besed. Če model naleti na neznano besedo (besedo zunaj slovarja), ne more določiti, kako jo transkribirati. Narečje prav tako spremeni zvok besede onkraj prepoznavanja modela. Nazadnje, modeli STT običajno zahtevajo, da raziskovalec določi jezik avdio posnetka. Če je jezik mešan, je težko zanesljivo transkribirati dialog. Na primer, izposojene angleške besede v slovenskem govoru predstavljajo težavo.

Transkripcija za pridobivanje glasbene informacije obsega transkripcijo instrumenta (na podlagi not), transkripcijo petja (na podlagi not) in transkripcijo besedila (na podlagi besedila). Za folkloristiko je avtomatska transkripcija besedila (ALT) najbolj relevantna. Najzgodnejši modeli so lahko identificirali vsako peto zapeto besedo (Mesaros in Virtanen 2010). Čeprav sodobni modeli dosegajo veliko višjo natančnost (Demirel idr. 2020), ALT še vedno dosega slabše rezultate kot standardni modeli STT zaradi pomanjkanja anotiranih baz učnih podatkov in kompleksnih testnih vzorcev, saj (ljudsko) petje v posnetkih vsebuje več hrupa, nejasno tonalnost, zdrs tona in neenakomerno ritmiko.

[2] Dostopno na: https://www.slovenscina.eu/razpoznavalnik.

Za slovenščino portal slovenscina.eu[3] našteva več odprtokodnih jezikovnih tehnologij, vključno s STT modelom (v času pisanja ni modela ALT).

**Zimska.**

Gora, dol in vsa planjava
pod odejo belo spava. —
Potok, prej tako glasan,
v trdi led je zakovan.

Medved gleda iz brloga,
glad, si misli, je nadloga,
brž umakne spet se, in
že zaspal je kosmatin...

Vihra razigrane volje
s snegom se vrti čez polje;
potnik, ki čez plan hiti,
z vihro, s snegom se bori...

Mrak se k zemlji bliža leno,
glavo ima vso sneženo. —
Kajžež — visok gore tri —
išče, a ga ne dobi...

Mimo oken, mimo duri
išče ga še v pozni uri,
z burjo, snegom se bori
ves ta čas, ko dete spi...

*Fran Žgur.*

»
Zimska.

Gora, dol in vsa planjava Medved gleda iz brloga,
pod odejo belo spava. — glad, si misli, je nadloga,
Potok, prej tako glasan, brž umakne spet se, in

v trdi led je zakovan. že zaspal je kosmatin...
Vihra razigrane volje Mrak se k zemlji bliža leng,
s snegom se vrti čez polje; — glavo ima vso sneženo. —
potnik, ki čez plan hiti, Kajtež — visok gore tri —

z vihro, s snegom se bori... išče, a ga ne dobi,..

Mimo oken, mimo duri
išče ga še v pozni uri,

z burjo, snegom se bori
ves ta čas, ko dete spi...

Fran Žgur.

SE |
P E

Tabela 1: Primer optičnega prepoznavanja znakov (OCR) z orodjem Tesseract za slovensko pesem (revija Zvonček, 1. 12. 1913). Model ima visoko točnost transkripcije besed, ne upošteva pa dvostolpične stave besedila. Transkript ima dve črkovni napaki (odebeljeno in podčrtano). Model poskuša transkribirati tudi del besedilnih ornamentov.

Prav tako je mogoče avdio posnetke analizirati neposredno. Avdio posnetek je treba predhodno obdelati z uporabo ločenih tehnik obdelave zvoka. Ena takšna tehnika je segmentacija podatkov, kjer se avdio posnetek razdeli na posamezne enote, kot so govor, solo petje, zborovsko petje in instrumentalna glasba. Segmentacija podatkov omogoča delo s posameznimi avdio enotami in primerjavo le-teh med različnimi posnetki. Obstaja veliko tehnik predobdelave zvoka; opis vseh presega obseg tega članka.

### 3.2 Digitalizacija tiska

Ko so izvorni podatki že na voljo kot besedilo, a v tiskani obliki, je treba uporabiti drugačno tehnologijo. Natisnjene pravljice, pripovedi, zgodovinska pričevanja in ročno pisani zapiski zahtevajo optično prepoznavanje znakov (OCR) za digitalizacijo. Tehnologija OCR prebere digitalno sliko besedila in sliko pretvori v zaporedje znakov z uporabo modela strojnega učenja (Tabela 1). Najbolj znano odprtokodno OCR orodje je Tesseract (Smith 2007), ki ga je nekaj časa razvijal Google. Kasneje je Google razvil svoje lastno OCR orodje, zapakirano v Google Cloud, medtem ko je Tesseract ostal odprtokoden. ABBYY je še ena priljubljena komercialna rešitev OCR.

[3] Dostopno na: https://slovenscina.eu.

OCR orodja običajno pričakujejo dobro strukturirano PDF datoteko, vendar pomanjkanje enotnega PDF standarda včasih preprečuje, da bi bili tudi preprosti dokumenti digitalizirani s 100-odstotno natančnostjo. Da bi premagali to težavo, se za določene vrste dokumentov razvijajo specializirani modeli OCR – na primer, za dvostolpične časopise, mešanico tiskanih in ročno napisanih opomb ali kompleksno postavitev s številnimi slikami (Berg-Kirkpatrick in Klein 2014).

Ključni korak pri digitalizaciji podatkov je ustrezno organiziranje podatkov in izbira primerne oblike. Avdio, video in slike so shranjeni v binarni obliki,[4] medtem ko so besedilni podatki običajno kodirani v formatu TEI (TEI Consortium 2023), priznani kodirni shemi v skupnosti digitalnih humanistov. TEI temelji na formatu XML, ki omogoča označevanje delov besedila (Slika 1). Oznake v TEI so vnaprej določene in vključujejo lastnosti, kot so datum, čas, ime, avtor, kraj in naslov. V takšnih podatkih je mogoče programsko iskati dano lastnost, na primer vse kraje.

Medtem ko je TEI predvsem format za shranjevanje podatkov, format CoNLL-U zagotavlja dodatne funkcije za obdelavo naravnega jezika. CoNLL-U besedilo je organizirano tako, da je v vsaki vrstici po ena beseda, ki so ji dodane jezikovne oznake. Predhodno označeni podatki omogočajo specializirano iskanje na podlagi jezikovnih značilnosti, na primer luščenje zgolj lematiziranih samostalnikov.

### 3.3 Obdelava naravnega jezika

Ko so podatki digitalizirani in pretvorjeni v želeni format, jih je treba pred nadaljnjo analizo predobdelati z uporabo tehnik obdelave naravnega jezika (NLP). Na primer, CoNNL-U že zagotavlja besedilo v visoko predobdelani obliki, zato je ta format še posebej uporaben za nadaljnjo analizo. Sicer je treba uporabiti dodatne tehnike NLP.

NLP predobdelava obsega kombinacijo tokenizacije, lematizacije in oblikoslovno označevanje (POS). Tokenizacija razdeli besedilo na osnovne enote analize, običajno besede. Lematizacija te enote preoblikuje v osnovno obliko, npr. v ednino imenovalnika. Oblikoslovno označevanje dodeli vsaki besedi v danem korpusu oznako besedne vrste na podlagi njene slovnične vloge in konteksta v stavku. Sorodna naloga je prepoznavanje imenskih entitet (NER), ki doda imensko entiteto dani besedi. Na primer, po opisanem postopku bi se stavek »*Jutri grem v Ljubljano,«* pretvoril v »*jutri_Adverb*, *iti_Verb*, *v_Adposition*, *Ljubljana_Noun_loc*«, pri čemer je vsaka beseda pretvorjena v osnovno obliko, priprisana ji je besedna vrsta, beseda »Ljubljana« pa je označena kot kraj.

Alternativa predobdelavi besedila je pretvorba vsake besede v vektorsko predstavitev na podlagi njenega pomena. Ta pristop se imenuje vložitev besed (angl. *embedding*). Pristop vzame vsako besedo iz besedila in ji dodeli vektor na podlagi predhodno naučenega modela vložitev.

[4] Glej smernice CLARIN: https://www.clarin.si/repository/xmlui/page/data?locale-attribute=sl.

Eden prvih lematizatorjev za slovenščino je bil LemmaGen (Juršič idr. 2010), razvit na Inštitutu Jožefa Stefana. Malce kasneje je bilo objavljeno orodje za oblikoslovno označevanje Obelisk (Grčar idr. 2012). Danes so bolj razširjena večnamenska orodja. Cevovod CLASSLA (Ljubešić in Dobrovoljc 2019) podpira tokenizacijo, lematizacijo, oblikoslovno označevanje in prepoznavanje imenskih entitet. Slovenski model Trankit (Krsnik in Dobrovoljc 2023) ponuja podobno funkcionalnost. Za vložitev besed je primerna SloBERTa (Ulčar in Robnik-Šikonja 2021), najnovejši model vložitve besed za standardno slovenščino, in njena mlajša sestra SloBERTa-SlEng za slovenski sleng. 1

## 4 Stopnja 3: Analiza in raziskovanje podatkov

Ko so podatki zbrani, organizirani in predobdelani, jih lahko analiziramo. Korpusno jezikoslovje, ki ima analognega predhodnika iz 13. stoletja (McCarthy in O'Keeffe 2010), je danes skoraj popolnoma računalniško. To je študij jezika z uporabo velikih zbirk besedil, imenovanih korpusi (ednina: korpus). Ker so podatki tako obsežni, korpusno jezikoslovje uporablja računalniške in statistične metode za njihovo preučevanje.

Eden prvih korakov pri analizi podatkov je raziskovanje podatkov, ki vključuje opazovanje frekvenc izrazov in meta(podatkov), osnovno statistiko, primerjavo kategorij in podobno. Bodisi za izhodiščno raziskovanje podatkov bodisi za obsežno jezikovno analizo so primerna možnost za začetnike konkordančniki. Konkordančnik je orodje za raziskovanje frekvenc besed in fraz, sopojavitev besed ter iskanje ključnih besed v kontekstu (KWIC). Obstaja veliko programske opreme za lokalno analizo konkordanc, medtem ko postajajo spletni konkordančniki vedno bolj razširjeni. Med komercialnimi orodji je najbolj znan SketchEngine (Kilgarriff idr. 2014), medtem ko je noSketchEngine priljubljena brezplačna različica za upravljanje korpusov.[5] CLARIN.SI navaja splošno dostopne slovenske konkordančnike.[6]

Naslednji korak za konkordančniki, ki temeljijo na ključnih besedah, je obdelava naravnega jezika in rudarjenje besedil. To vključuje analizo celotnih korpusov in prepoznavanje vzorcev v njih z metodami, ki temeljijo na umetni inteligenci, kot so tematsko modeliranje, gručenje ali klasifikacija dokumentov, analiza sentimenta in podobno.

Med bolj priljubljenimi spletnimi orodji za analizo besedil je Voyant Tools[7] (Miller 2018), brezplačno orodje za analizo besedil. Ponuja več statističnih vizualizacij, kot so oblaki besed, grafi sopojavitev, spremembe frekvenc besed in analize trendov. Glavna prednost Voyant-a, poleg tega, da je brezplačno dostopna spletna storitev, je njegova enostavnost uporabe, kjer raziskovalec naloži dokumente in interaktivno opazuje statistiko in grafe.

---

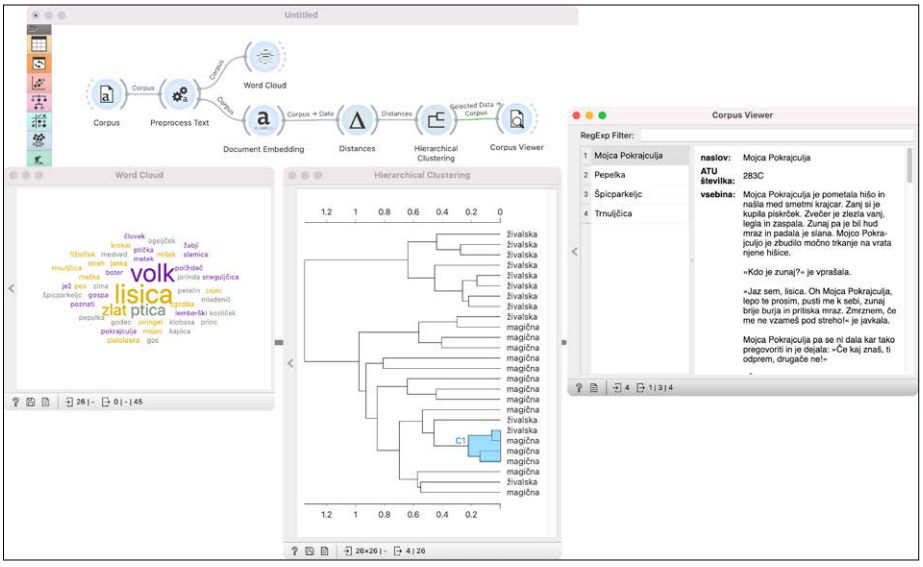
[5] Seznam konkordančnikov je dostopen na: https://corpus-analysis.com/tag/concordancer. Stran navaja tudi druga orodja za analizo korpusov.

[6] Slovenski konkordančniki: https://www.clarin.si/info/concordances.

[7] Dostopno na: https://voyant-tools.org/.

```
</seg>
      <bibl corresp="#bib23.1">
        <date type="reprint" when="1999">1999</date> (<date type="firstEdition" when="1789">1789</date>
        <biblScope unit="page">113</biblScope>
      </bibl>
      <bibl corresp="#bib7.1">
        <date when="1789">1789</date>
        <biblScope unit="page">408</biblScope>
      </bibl>
    </ab>
    <ab xml:id="PREG-00-00012" n="13">
      <seg xml:lang="sl-bohoric" xml:id="PREG-00-00012.dipl" type="dipl">Sarezhen kruh fe rad rieshe.</seg>
      <seg xml:lang="sl" xml:id="PREG-00-00012.crit" type="crit">Sarečen kruh se rad rieže.</seg>
<seg xml:id="PREG-00-00012.crit.norm" type="crit.norm" xml:lang="sl">Zarečen kruh se rad reže.</seg>
<seg xml:id="PREG-00-00012.crit.norm.ana" type="crit.norm.ana" xml:lang="sl">
<s xml:id="PREG-00-00012.crit.s1">
<w norm="Zarečen" ana="mte:Appmsnn" msd="UPosTag=ADJ|Case=Nom|Definite=Ind|Degree=Pos|Gender=Masc|Number=Sing|VerbForm=Part"
lemma="zarečen" xml:id="PREG-00-00012.crit.s1.t1">Sarečen</w>
<w ana="mte:Ncmsn" msd="UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Sing" lemma="kruh" xml:id="PREG-00-00012.crit.s1.t2">kruh</
w>
<w ana="mte:Px------y" msd="UPosTag=PRON|PronType=Prs|Reflex=Yes|Variant=Short" lemma="se" xml:id="PREG-00-00012.crit.s1.t3">se</w>
<w ana="mte:Rgp" msd="UPosTag=ADV|Degree=Pos" lemma="rad" xml:id="PREG-00-00012.crit.s1.t4">rad</w>
<w norm="reže" ana="mte:Vmpr3s" msd="UPosTag=VERB|Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin"
lemma="rezati" join="right" xml:id="PREG-00-00012.crit.s1.t5">rieže</w>
<pc ana="mte:Z" msd="UPosTag=PUNCT" xml:id="PREG-00-00012.crit.s1.t6">.</pc>
<linkGrp type="UD-SYN" targFunc="head argument" corresp="#PREG-00-00012.crit.s1">
 <link ana="ud-syn:amod" target="#PREG-00-00012.crit.s1.t2 #PREG-00-00012.crit.s1.t1"/>
 <link ana="ud-syn:nsubj" target="#PREG-00-00012.crit.s1.t5 #PREG-00-00012.crit.s1.t2"/>
 <link ana="ud-syn:expl" target="#PREG-00-00012.crit.s1.t5 #PREG-00-00012.crit.s1.t3"/>
 <link ana="ud-syn:advmod" target="#PREG-00-00012.crit.s1.t5 #PREG-00-00012.crit.s1.t4"/>
 <link ana="ud-syn:root" target="#PREG-00-00012.crit.s1 #PREG-00-00012.crit.s1.t5"/>
 <link ana="ud-syn:punct" target="#PREG-00-00012.crit.s1.t5 #PREG-00-00012.crit.s1.t6"/>
</linkGrp>
</s>
```

Slika 1: Primer pregovora (Babič 2022), označenega po shemi TEI.




Slika 2: Primer delotoka hierarhičnega gručenja pravljic bratov Grimm v orodju Orange. Delotok prikazuje gručo (C1), v kateri so različne oznake ATU. Mojca Pokrajculja je živalska pravljica, ostale pa so magične pravljice. Kljub temu si te pravljice delijo podobne besede.

Za pristop s poudarkom na umetni inteligenci orodja za vizualno programiranje ponujajo nizek vstopni prag za raziskovalce, ki želijo spoznati rudarjenje besedil. Ta orodja namreč vizualno prikažejo delotok, dajejo takojšnje (in vmesne) rezultate, spodbujajo interakcijo in so na splošno zasnovana, da so preprosta. Tak primer je odprtokodno orodje za rudarjenje podatkov Orange[8] (Demšar idr. 2013), kjer uporabnik sestavlja analitične delotoke s povezovanjem komponent. Modularno zlaganje komponent omogoča prilagodljive, a sposobne delotoke, ki ne zahtevajo programerskih veščin (Slika 2). Rezultati so predstavljeni v interaktivnih vizualizacijah, ki omogočajo raziskovanje podatkov. Orodje ponuja razširitev Text za analizo besedil. Podobna orodja za vizualno programiranje so KNIME, RapidMiner in Weka.

Za analizo folklornih podatkov obstaja veliko tehnik na osnovi umetne inteligence, kot so analiza sentimenta, časovna analiza, razreševanje koreferenc, dodeljevanje razrednih oznak (klasifikacija) ali samodejno luščenje tem (tematsko modeliranje) iz dokumentov. Ker zahtevajo računalniške in znanstvene veščine, lahko navdušen digitalni humanist začne z zmogljivimi konkordančniki in nadaljuje od tam. Projekti digitalnih humanistov so izrazito primerni za interdisciplinarno sodelovanje.

## 5 Stopnja 5: Ohranjanje podatkov

Zadnja stopnja v življenjskem ciklu podatkov je shranjevanje in ohranjanje podatkov. Dobre prakse deljenja podatkov omogočajo ponovljivost rezultatov in ponovno uporabo podatkov, kar nas pripelje nazaj k prvemu koraku – zbiranju podatkov. Seveda so primerni za deljenje le nekateri raziskovalni podatki, še posebej, če vključujejo občutljive informacije. Kljub temu se večina podatkov v folkloristiki lahko deli v neki obliki.

FAIR (Wilkinson idr. 2016) je najbolj znan in široko sprejet standard za upravljanje podatkov. To je kratica za najdljivost (***f****indability*), dostopnost (***a****ccessibility*), interoperabilnost (***i****nteroperability*) in ponovno uporabnost (***r****eusability*). To so štiri splošna načela, ki povzemajo idejo, da mora biti raziskovalne podatke enostavno najti, pridobiti in delati z njimi. To pomeni, da morajo podatki imeti trajne identifikatorje, bogate metapodatke, biti indeksirani v iskalnem viru (najdljivost), jih je mogoče pridobiti z odprtim protokolom (dostopnost), so v splošno sprejetem formatu (interoperabilnost) in imajo podrobne (meta)podatke in jasne, odprte licence (ponovna uporabnost). Uporaba FAIR načel za humanistiko je opisana v Harrower idr. (2020).

FAIR načela so splošno vključena v odprte podatkovne repozitorije. Ti repozitoriji, na primer Danski Folkloristični Makroskop (Tangherlini 2013) ali bolj splošni Virtualni Jezikovni Observatorij, vključujejo strukturirane podatke iz folkloristike.

[8] Dostopno na: https://orangedatamining.com/.

Slovenska nacionalna veja evropske infrastrukture CLARIN, CLARIN.SI[9], omogoča dostop do različnih jezikovnih virov predvsem (a ne izključno) v južnoslovanskih jezikih. Vsi podatki niso v besedilni obliki – podatki iz folkloristike vključujejo tudi avdio (Matsuura idr. 2020) in video datoteke[10] (Nonhebel idr. 2004), medtem ko so digitalni podatki kulturne dediščine dostopni na spletnem portalu Europeana (Purday 2009). Za shranjevanje podatkov je primeren Zenodo, multidisciplinarni repozitorij odprtih raziskovalnih podatkov, ki podpira samoarhiviranje in splošno upravljanje in shranjevanje podatkov (za pregled glej Peters idr. 2017).

## 6 Zaključek

Seth Long trdi, da podatki v digitalni humanistiki kličejo po ponovni uporabi (Long in Williams 2014). Obstoječi korpusi in podatkovni nizi se lahko uporabijo za primerjalno analizo, analizo z modernimi pristopi, dopolnijo z dodatnim terenskim delom ali pa se obravnavajo na nove načine. Spletni podatki lahko usmerjajo raziskave z ustvarjanjem raziskovalnih vprašanj, dopolnjevanjem terenskih podatkov ali kot samostojen raziskovalni material.

Kje naj torej začne navdušeni digitalni humanist ali računalniški folklorist? Najprej s podatki. Digitalizacija podatkov iz folkloristike je ključna za zagotavljanje boljše kakovosti podatkov, kopiranje in razmnoževanje brez izgube kakovosti, lažje iskanje prek indeksiranja ter lažji dostop do podatkov za primerjalne raziskave in ponovno uporabo podatkov. Mnogi raziskovalci si prizadevajo za zagotavljanje široko dostopnih podatkov iz folkloristike, nekatera od teh prizadevanj so privedla do zglednih repozitorijev. Nekatere institucije s celovitimi spletnimi aplikacijami, ki omogočajo indeksirano iskanje, časovno usklajeno transkripcijo in vizualizacijo (Boyd 2019), ponujajo več kot zgolj običajen digitalni dostop. To je tisto, kar Timothy Tangherlini imenuje » folklorni makroskop«, tj. sistem, ki omogoča celostno uporabo institucionalnih zbirk (Tangherlini 2013). Primer takšnega makroskopa je Sinhronizator Metapodatkov Ustnih Zgodovin, ki poravnava avdio datoteke s časovnimi žigi transkriptov in izboljšuje segmente z metapodatki (tema, ključne besede, povzetek) (Boyd 2019).

Digitalni podatki zahtevajo digitalno usposobljeno osebje in ustrezne zmogljivosti za shranjevanje podatkov, kar ne sme biti samoumevno (glej Mosweu 2011). Usposabljanje na področju korpusnega jezikoslovja, rudarjenja besedil in obdelave naravnega jezika je osnova za napredno besedilno raziskovanje. Vendar to ne vključuje nujno učenja programiranja, kot je razvidno v razdelku 4. Mnoga orodja imajo nizko vstopno

[9] Dostopno na: https://www.clarin.si.

[10] Video datoteke se tipično uporablja za znakovni jezik.

raven, kjer strokovnjaki za folkloristiko lahko uporabijo svoje široko domensko znanje za raziskovanje podatkov na nove načine.

Z obilico digitalnih orodij, ki so na razpolago raziskovalcu, se pojavi vprašanje, kaj potem ostane vlogi raziskovalca? Pazljiv bralec bo opazil, da v članku manjka ena stopnja cikla podatkov; stopnja 4: interpretacija podatkov. Čeprav so računalniki in sodobne tehnike na osnovi umetne inteligence odlični pri organizaciji in čiščenju podatkov, so daleč od človeške sposobnosti interpretacije podatkov. Raziskovalec razume širši kontekst raziskovalnega problema, zgodovinske temelje, razmerja moči, socialno hierarhijo in jezikovne posebnosti. Sinteza izhaja iz združevanja obstoječega znanja z novimi vpogledi. Digitalna orodja so ravno to – orodja, ki nam pomagajo prepoznati zanimive vzorce ali odstopanja. Identifikacija vzročnosti, razlaga ugotovitev in določanje pomena ostaja v celoti v človeški domeni.

**Zahvala**

Članek je nastal v okviru financiranja Javne agencije za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije P6-0436: *Digitalna Humanistika: viri, orodja in metode* (2022–2027).

## Viri in literatura

Ang, Chee Siang; Bobrowicz, Ania; Schiano, Diane J.; Nardi, Bonnie, 2013: Data in the Wild: Some Reflections. *Interactions* 20/2, 39–43. DOI: https://doi.org/10.1145/2427076.2427085.

Babič, Saša; idr., 2022: Collection of Slovenian paremiological units Pregovori 1.0, Slovenian language resource repository CLARIN.SI. Na spletu: http://hdl.handle.net/11356/1455 (dostop 13. 2. 2024).

Berg-Kirkpatrick, Taylor; Klein, Dan, 2014: Improved Typesetting Models for Historical OCR, V: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore: ACL, 118–23.

Boyd, Doug, 2019: OHMS: Enhancing Access to Oral History for Free. *The Oral History Review* 40/1, 95–106. DOI: https://doi.org/10.1093/ohr/oht031.

Chiaro, Delia, 2018: *The language of jokes in the digital age: Viral humour*. Oxon: Routledge.

Demirel, Emir; Ahlbäck, Sven; Dixon, Simon, 2020: Automatic Lyrics Transcription Using Dilated Convolutional Neural Networks with Self-Attention, V: *International Joint Conference on Neural Networks*. Glasgow: IEEE, 1–8.

Demšar, Janez; Curk, Tomaž; Erjavec, Aleš idr., 2013: Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research* 14/avg, 2349−2353.

Eklund, Johan; Hagedorn, Josh; Darányi, Sándor, 2023: Teaching Tale Types to a Computer: A First Experiment with the Annotated Folktales Collection. *Fabula* 64/1–2, 92–106. DOI: https://doi.org/10.1515/fabula-2023-0005.

Fišer, Darja; Ljubešić Nikola; Erjavec Tomaž, 2018: The Janes project: language resources and tools for Slovene user generated content. *Language Resources & Evaluation*. DOI: https://doi.org/10.1007/s10579-018-9425-z.

Grčar, Miha; Krek, Simon; Dobrovoljc, Kaja, 2012: Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik (Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene). V: *Proceedings of the 8th Language Technologies Conference* C. Ljubljana: IJS, 89–94.

Halford, Susan, 2017: The Ethical Disruptions of Social Media Data: Tales from the Field, V: Woodfield, Kandy (ur.), *Ethics of Internet-mediated Research and Using Social Media for Social Research*. Bristol: Emerald, 13–25.

Harrower, Natalie; Immenhauser, Beat; Lauer, Gehrard; Maryl, Maciej; Orlandi, Tito; Rentier, Bernard; Wandl-Vogt, Eveline, 2020: *Sustainable and FAIR Data Sharing in the Humanities*. Berlin, Germany: ALLEA - All European Academies. DOI: https://doi.org/10.7486/DRI.tq582c863.

Juršič, Matjaž; Mozetič, Igor; Erjavec, Tomaž; Lavrač, Nada, 2010: Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science* 16/9, 1190–1214. DOI: https://doi.org/10.3217/jucs-016-09-1190.

Karsdorp, Folgert; van der Meulen, Marten; Meder, Theo; van den Bosch, Antal, 2015: MOMFER: A Search Engine of Thompson's Motif-Index of Folk Literature. *Folklore* 126/1, 37–52. DOI: https://doi.org/10.1080/0015587X.2015.1006954.

Kilgarriff, Adam; Baisa, Vít; Bušta, Jan. idr., 2014: The Sketch Engine: ten years on. *Lexicography ASIALEX* 1, 7–36. DOI: https://doi.org/10.1007/s40607-014-0009-9.

Kingsland, Kaitlyn, 2020: Comparative Analysis of Digital Photogrammetry Software for Cultural Heritage. *Digital Applications in Archaeology and Cultural Heritage* 18. DOI: https://doi.org/10.1016/j.daach.2020.e00157.

Krawczyk-Wasilewska, Violetta, 2016: *Folklore in the Digital Age: Collected Essays*. Łódź: Jagiellonian University Press.

Krsnik, Luka; Dobrovoljc, Kaja, 2023: The Trankit model for linguistic processing of standard Slovenian, Slovenian language resource repository CLARIN.SI. Na spletu: http://hdl.handle.net/11356/1870 (dostop 16. 2. 2024).

Kung, Susan S., 2022: Developing a data management plan. V: Berez-Kroeker, Andrea L.; McDonnell, Bradley; Koller, Eve; Collister, Lauren B. (ur.), *The Open Handbook of Linguistic Data Management*. Cambridge, Massachusetts: The MIT Press, 101–115.

Lebar Bajec, Iztok; Bajec, Marko; Bajec, Žan; Rizvič, Mitja, 2022: Slovene Conformer CTC BPE E2E Automated Speech Recognition model RSDO-DS2-ASR-E2E 2.0, Slovenian language resource repository CLARIN.SI. Na spletu: http://hdl.handle.net/11356/1737 (dostop 13. 2. 2024).

Lee, Benjamin Charles Germain, 2022: The Digital Humanities and the Ladino Press: Using Machine Learning to Extract and Analyze Visual Content in Historic Ladino Newspapers. *Studies in Digital History and Hermeneutics*, 191. DOI: https://doi.org/10.1515/9783110744828-010.

Ljubešić, Nikola; Dobrovoljc, Kaja, 2019: What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian. V: *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*. Florence: Association for Computational Linguistics, 29–34.

Long, Seth; Williams, Sierra, 2014: "Re-Purposing" Data in the Digital Humanities: Data Beg to Be Taken from One Context and Transferred to Another. Na spletu: https://blogs.lse.ac.uk/impactofsocialsciences/2014/04/02/re-purposing-data-in-the-digital-humanities/ (dostop 2. 4. 2014).

Lourdi, Irene; Papatheodorou, Christos; Nikolaidou, Mara, 2007: A multi-layer metadata schema for digital folklore collections. *Journal of Information Science* 33/2, 197–213. DOI: https://doi.org/10.1177/0165551506070711.

Marolt, Matija; Bohak, Ciril; Kavčič, Alenka; Pesek, Matevž, 2019: Automatic Segmentation of Ethnomusicological Field Recordings. *Applied Sciences* 9/3, 439. DOI: https://doi.org/10.3390/app9030439.

Matsuura, Kohei; Ueno, Sei; Mimura, Masato; Sakai, Shinsuke; Kawahara, Tatsuya, 2020: Speech Corpus of Ainu Folklore and End-to-End Speech Recognition for Ainu Language, V: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 2622–28.

McCarthy, Michael; O'Keeffe, Anne, 2010: Historical perspective: What are corpora and how have they evolved? V: O'Keeffe, Anne; McCarthy, Michael (ur.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 3–13.

Meder, Theo; Himstedt-Vaid, Petra; Meyer, Holger, 2023: Collecting International Narrative Heritage in a Multilingual Search Engine. *Fabula* 64/1–2, 107–27. DOI: https://doi.org/10.1515/fabula-2023-0006.

Mesaros, Annamaria; Virtanen, Tuomas, 2010: Automatic Recognition of Lyrics in Singing. *EURASIP Journal on Audio, Speech, and Music Processing* 2010/1. DOI: http://dx.doi.org/10.1155/2010/546047.

Miller, Alissa, 2018: Text Mining Digital Humanities Projects: Assessing Content Analysis Capabilities of Voyant Tools. *Journal of Web Librarianship* 12/3, 169–197. DOI: https://doi.org/10.1080/19322909.2018.1479673.

Mosweu, Tshepho, 2011: Digitising the Oral History Collection at Botswana National Archives and Records Services: Problems and Prospects. *Journal of the South African Society of Archivists* 44, 124–30.

Nonhebel, Annika; Crasborn, Onno; van der Kooij, Els, 2004: Sign Language Transcription Conventions for the ECHO Project. *Version* 9, 20.

Park, Tae Jin; Kanda, Naoyuki; Dimitriadis, Dimitrios; Han, Kyu J.; Watanabe, Shinji; Narayanan, Shrikanth, 2022: A Review of Speaker Diarization: Recent Advances with Deep Learning. *Computer Speech & Language* 72. DOI: https://doi.org/10.1016/j.csl.2021.101317.

Peters, Isabella; Kraker, Peter; Lex, Elisabeth; Gumpenberger, Christian; Gorraiz, Juan Ignacio, 2017: Zenodo in the Spotlight of Traditional and New Metrics. *Frontiers in Research Metrics and Analytics* 2. DOI: http://dx.doi.org/10.3389/frma.2017.00013.

Povey, Daniel; Ghoshal, Arnab; Boulianne, Gilles; Burget, Lukas; Glembek, Ondrej; Goel, Nagendra; Hannemann, Mirko; idr., 2011, The Kaldi Speech Recognition Toolkit, V: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hawaii: IEEE Signal Processing Society.

Purday, Jon, 2009: Think Culture: Europeana. Eu from Concept to Construction. *The Electronic Library* 27/6, 919–937. DOI: https://doi.org/10.1108/02640470911004039.

Radford, Alec; Kim, Jong Wook; Xu, Tao; Brockman, Greg; Mcleavey, Christine; Sutskever, Ilya, 2023: Robust Speech Recognition via Large-Scale Weak Supervision, V: Krause, Andreas; Brunskill, Emma; Cho, Kyunghyun; Engelhardt, Barbara; Sabato, Sivan; Scarlett, Jonathan (ur.), *Proceedings of the 40th International Conference on Machine Learning*. 28492–518.

Shifman, Limor, 2013: *Memes in digital culture*. Cambridge, Massachusetts: MIT press.

Smith, Ray, 2007: An Overview of the Tesseract OCR Engine, V: *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*. Curitiba: IEEE, 629–633.

Smits, Thomas; Wevers, Melvin, 2023: A Multimodal Turn in Digital Humanities: Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections. *Digital Scholarship in the Humanities* 38/3, 1267–1280. DOI: https://doi.org/10.1093/llc/fqad008.

Tangherlini, Timothy R., 2013: The Folklore Macroscope: Challenges for a Computational Folkloristics. *Western Folklore*, 7–27.

TEI Consortium. 2023: TEI P5: Guidelines for Electronic Text Encoding and Interchange. Na spletu: https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html (dostop 16. 11. 2023).

Thuillard, Marc; Le Quellec, Jean-Loïc; d'Huy, Julien, 2018: Computational Approaches to Myths Analysis: Application to the Cosmic Hunt. *Nouvelle Mythologie Comparée/New Comparative Mythology* 4, 123–154.

Ulčar, Matej; Robnik-Šikonja, Marko, 2021: SloBERTa: Slovene monolingual large pretrained masked language model. V: *Proceedings of SI-KDD within the Information Society 2021*. 17–20.

Van Uytvanck, Dieter; Stehouwer, Herman; Lampen, Lari, 2012: Semantic Metadata Mapping in Practice: The Virtual Language Observatory, V: Calzolari, Nicoletta; Choukri, Khalid; Declerck, Thierry, idr. (ur.), *LREC 2012: 8th International Conference on Language Resources and Evaluation*. Istanbul: European Language Resources Association (ELRA), 1029–34.

Verdonik, Darinka; Zwitter Vitez, Ana, 2011: Slovenski govorni korpus Gos. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Wiering, Frans; Veltkamp, Remco C.; Garbers, Jörg; Volk, Anja; van Kranenburg, Peter; Grijp, Louis P., 2009: Modelling Folksong Melodies. *Interdisciplinary Science Reviews* 34/2–3, 154–71. DOI: https://doi.org/10.1179/174327909X441081.

Wilkinson, Mark; Dumontier, Michel; Aalbersberg, IJsbrand J. idr., 2016: The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3. DOI: https://doi.org/10.1038/sdata.2016.18.

Yogarajah, Yathukulan, 2022: 'Hodling' on: Memetic Storytelling and Digital Folklore within a Cryptocurrency World. *Economy and Society* 51/3, 467–88. DOI: https://doi.org/10.1080/03085147.2022.2091316.