Analiza naslova fotografske građe u javnim mrežnim katalozima

Šafar, Leon

Undergraduate thesis / Završni rad

2022

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:563213

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-04-20**

Filozofski fakultet
Sveučilišta u Zagrebu

Repository / Repozitorij:

ODRAZ - open repository of the University of Zagreb
Faculty of Humanities and Social Sciences

zir.nsk.hr

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
2021./2022.

Leon Šafar

Analiza naslova fotografske građe u javnim mrežnim katalozima

Završni rad

Mentor: dr.sc. Goran Zlodi, izv. prof.

Zagreb, rujan 2022.

Izjava o akademskoj čestitosti

Izjavljujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

Sadržaj

Sadržaj.....iv

1. Uvod.....1

2. Katalogizacija.....2

2.1. Djela i slike.....3

2.1.1. Djelo.....3

2.1.2. Slika.....3

2.1.3. Odnos između zapisa o djelu i zapisa o slici.....4

3. Naslov.....5

3.1. Opisni i nadomjesni naslovi.....6

3.2. Informacije o prikazu: iz CCO smjernica.....7

3.2.1. Opis prikaza.....7

3.2.2. Vrsta prikaza.....7

3.2.3. Prikaz predmeta.....7

3.2.4. Datum prikaza.....8

3.3. Analiza primjera.....8

3.4. Zaključak analize.....12

4. Wikifier.....13

4.1. Objašnjenje koda i definicije:.....14

4.2. Primjena Wikifiera:.....17

4.3. Primjena Wikifiera na primjerima.....18

4.4. Analiza rezultata Wikifiera na primjerima.....26

4.5. Usporedna analiza Wikifiera i NER-a.....28

5. Zaključak.....31

6. Literatura.....32

7. Popis slika.....33

8. Popis tablica.....34

Sažetak.....35

Summary.....36

1. Uvod

Tema ovog završnog rada je analiza naslova fotografske građe u javnim mrežnim katalozima. Razlog odabira ove teme je što ne postoje jasne upute za oblikovanje naslova za primjere koji se većinom odnose na nepokretne spomenike kulture, zgrade, građevine i ostale slične tvorevine, koje nemaju glavni stvarni naslov, već se sastavlja nadomjesni ili opisni naslov od strane osobe koja katalogizira građu. U poglavlju Katalogizacija je ukratko opisan i definiran proces katalogizacije te razlika između pojmova djela i slike te njihov međusobni odnos što je važno za nastavak rada. U idućem poglavlju, Naslov, ukratko su definirani naslovi, njihove svrhe i načini upotrebe. Fokus je na nadomjesnom i opisnom naslovu te na smjernicama iz CCO-a koje daju konstrukciju i opis istim naslovima. Zatim su analizirani primjeri iz domaćih i stranih javnih mrežnih kataloga te zabilježene usporedbe. U zadnjem poglavlju se govori o Wikifieru, web servisu napisanom u programskom jeziku Python 3, koji na brz i jednostavan način daje dodatne informacije o traženom tekstu. Definiran je kod pozivanja web servisa, opisan način na koji radi te na istim primjerima napravljena analiza i usporedba izlaznih podataka. Wikifier je uspoređen sa još dva web servisa na istim primjerima te su izlazni podaci analizirani.

Cilj ovog rada je predstaviti strukturu teksta naslova za djela koja nemaju određen naslov kako bi se dobili termini kojima se može ciljati na kontrolirana nazivlja za predmetnu obradu. Kako fotografska građa nema naslov kao takav, isti termini se mogu iskoristiti za poluautomatiziranu dodjelu, odnosno generiranje naslova prema uzorku (objekt, godina izgradnje, opis ili pogled, mjesto). Katalogizator bi naslov mogao doraditi nakon njegova predgeneriranja.

2. Katalogizacija

Katalogizacija je proces pohranjivanja građe u baze podataka raznih domena kao što su muzeji, knjižnice, arhivi i tako dalje, u svrhu inventarizacije radi čuvanja zapisa nabavljene građe te omogućavanje korisnicima pronalaženje željenih jedinica građe (Bowman, 2010: 4-5). Prema Pravilniku za opis i pristup građi u knjižnicama, arhivima i muzejima „jedinica građe je materijalni ili nematerijalni objekt nastao djelovanjem čovjeka ili prirode, koji ima društveni, povijesni, znanstveni, kulturni i/ili umjetnički značaj, što ga čini predmetom prikupljanja, pohrane, obrade, komuniciranja i zaštite u knjižnicama, arhivima i muzejima". (Nacionalna i sveučilišna knjižnica u Zagrebu, Hrvatski državni arhiv, Muzejski dokumentacijski centar, 2019: 8). Kako bi proces bio olakšan i izvediv, postoje razna pravila kojih se valja pridržavati na međunarodnoj razini. Nadovezujući se na postojeće standarde, knjiga *Katalogiziranje predmeta kulturne baštine: vodič za opis djela kulturne baštine i njihovih slika* (dalje u tekstu CCO smjernice) daje „smjernice za odabir, redoslijed i oblikovanje podataka koji se koriste za popunjavanje elemenata metapodataka u katalošku zapisu; ovaj priručnik osmišljen je za promicanje dobre opisne katalogizacije, zajedničke dokumentacije i poboljšanog pristupa krajnjem korisniku" (Baca et al., 2006). Iako dobro osmišljen priručnik, naglašava se da pojedine domene zahtijevaju detaljniju katalogizaciju, kao na primjer kod muzejskih registara gdje je potrebno podrobnije opisati stanje građe u kojoj se nalazi, način konzerviranja iste građe te mjere (dužina, visina, širina). CCO smjernice se fokusiraju na standarde sadržaja podataka za deskriptivnu katalogizaciju, to jest standarde koji usmjeravaju izbor pojmova i koji definiraju redoslijed, sintaksu i oblik u kojem su ti pojmovi, fraze, vrijednosti i narativni opisi zabilježeni. „Katalogizirati djelo (engl. *work*) znači opisati što je djelo, tko je napravio djelo, gdje je napravljeno, od kojih materijala je izrađeno i o čemu je riječ" (Baca et al., 2006). Prije početka procesa katalogizacije, katalogizator mora znati fokus kataloginiranja te jasno razumjeti parametre onoga što opisuje. Katalogiziranje se odvija po unaprijed određenim pravilima tezaurusa koji opisuje funkciju elemenata od kojih su neki obavezni, a to su vrsta djela (engl. *work type*), naslov, autor, mjere i materijali i tehnike. Nabrojani elementi moraju biti izrečeni, a u slučaju nedostataka pojedinih informacija postavlja se vrijednost NULL, ostavlja prazno ili upisuje „nepoznato" iako ti postupci nisu poželjni (Baca et al., 2006: 10).

2.1. Djela i slike

Potrebno je razlikovati sliku (engl. *image*) od djela (engl. *work*) u procesu katalogizacije jer se za opis i dokumentiranje djela i slike koriste isti tipovi podataka većinom u oba slučaja da ne bi došlo do zbunjivanja krajnjeg korisnika ili do netočnih rezultata traženja podataka (Baca et al., 2006: 4). Na temelju toga se stvara zapis o djelu (engl. *work record*) i zapis o slici (engl. *image record*) po zadanim smjernicama te su isti povezani na razne načine.

2.1.1. Djelo

Djelo je posebna intelektualna ili umjetnička kreacija ograničena primarno na objekte i strukture koje su izradili ljudi, uključujući izgrađena djela, vizualna umjetnička djela i kulturne artefakte. „Pod izgrađena djela spada arhitektura, druge strukture ili okruženje koje je stvorio čovjek, obično dovoljno veliko da ljudi mogu ući, najčešće služe praktičnoj svrsi te se smatra da ima estetsku vrijednost" (Baca et al., 2006: 4). Takva fotografirana djela su glavni fokus ovog završnog rada.

2.1.2. Slika

Slika je vizualni prikaz djela. Obično postoji u fotomehaničkom, fotografskom ili digitalnom formatu. U tipičnoj zbirci vizualnih izvora, slika je slajd, fotografija ili digitalna datoteka. Slike ne uključuju trodimenzionalne fizičke modele, crteže, slike ili skulpture, koji su djela sama po sebi. Fotografija djela također se može tretirati ili kao umjetničko djelo ili kao slika, ovisno o utjecaju fotografa i estetskoj ili povijesnoj vrijednosti fotografije. Na primjer, fotografije Eugenea Atgeta umjetnosti i arhitekture Pariza, fotografije Alison Frantz starogrčkih mjesta i artefakata ili fotografija La Tour Eiffel poznatog francuskog fotografa Brassaïja prikazuje Eiffelov toranj noću. Ova se fotografija tretira kao umjetničko djelo, a ne samo kao slika koja dokumentira Eiffelov toranj. Nasuprot tome, druga fotografija koja prikazuje istu strukturu vjerojatno bi se tretirala kao fotografska dokumentacija Eiffelovog

tornja, zabilježena u slikovnom zapisu i povezana s radnim zapisom za Eiffelov toranj kao arhitektonsko djelo (Baca et al., 2006: 5).

2.1.3. Odnos između zapisa o djelu i zapisa o slici

Zapis o slici je povezan poveznicom na zapis o djelu te bi tako bio povezan i sa informacijama o djelu. Djelo može biti povezano sa više slika (fotografije iz raznih kutova), a slika može biti povezana sa više djela (fotografija dvije crkve). Katalogizacija slika arhitekture se vrši na tri moguća načina. Jednim pristupom se stvara jedinstveni zapis o djelu građevine na koji se povezuje zapis o slici za eksterijer, interijer, detalje i tako dalje. Ova metoda se koristi najčešće za jednostavne građevine i tvorevine. Drugi način također kreira zapis o djelu zgrade koji su povezani sa zapisom o slici, ali na način da se stvara zapis o djelu za svaki plan, model ili drugi dokument vezan uz to. Svaki zapis o slici je vezan uz svoj zapis o djelu (na primjer, slika plana katedrale kao zapis o slici je vezana uz sliku plana katedrale kao zapis o djelu). Treći pristup stvara zapis o djelu građevine, ali ju također razlama i stvara novi zapis za svaki dio vezan uz nju (u primjeru crkve, oltar, kupola, atrij i tako dalje bi imali svoj vlastiti zapis o djelu) (Baca et al., 2006: 6).

3. Naslov

Iako spada pod kategoriju katalogizacije, ovdje ima posebno poglavlje jer je to primarni fokus ovog završnog rada. Naslov je obavezni podatak prilikom unosa jedinične građe u bazu podataka koji služi kao razlikovno obilježje te za identifikaciju i lakši pronalazak djela. Postoji nekoliko vrsta naslova: glavni stvarni naslov, alternativni naslov, usvojeni naslov, varijantni naslov, opisni naslov, nadomjesni naslov. Glavni stvarni naslov je riječ, izraz ili skup znakova na jedinici građe koji primarno imenuje jedinicu građe te on može biti i alternativni naslov u slučaju da je autorov naslov preferirani (Nacionalna i sveučilišna knjižnica u Zagrebu, Hrvatski državni arhiv, Muzejski dokumentacijski centar, 2019: 54). Naslov može biti pridodan od strane autora (na primjer u slikarstvu), takozvani autorov naslov, ali postoje slučajevi u kojima se naslov smatra neprikladnim za identifikaciju ili nije dan od strane autora. U tom slučaju se uzima tekstualni sadržaj te jedinične građe, u potpunom ili skraćenom obliku, sa odgovarajućim interpunkcijskim znakovima. Ako jedinica građe ne sadržava tekst, podatak o stvarnom naslovu se izostavlja te mu se dodjeljuje opisni ili nadomjesni naslov od strane osobe koja se bavi katalogiziranjem tog djela ili nadležne institucije. Mnogi od takvih primjera su razne fotografije, pogotovo arhitekture, to jest zgrade, gradovi, spomenici. Iako je gore napomenuto da se u slučaju manjka informacija može koristiti pojam „bez naslova”, to se ne odnosi na slučaj naslovljavanja jedinice građe, osim u primjeru gdje je autor naslovio svoje djelo sa „Bez naslova (engl. *Untitled*)”. U krajnjem slučaju može se navesti „nepoznat naslov (engl. *title unknown*)” malim slovima kako bi se razlikovalo od ostalih naslova (Baca et al., 2006: 66). Naslov se prevodi na materinji jezik po pravilima određenima za jezik prijevoda. Ako se ostavlja naslov na izvornom jeziku, on postaje alternativni naslov. Ovaj postupak omogućuje bolju povezanost sa drugim izvorima. Iznimka je ako je djelo službeno poznato pod izvornim naslovom kao na primjer „Notre Dame, Paris”.

Primjer: Grad Zvečaj

„Objašnjenje: Odnosi se na fotografiju na kojoj se javlja grafički istaknut glavni stvarni naslov Schloss Svečaj uz usporedne stvarne naslove Svečaji vár i Grad Zvečay. Budući da je glavni stvarni naslov na njemačkom jeziku, a usporedni stvarni naslov na hrvatskom jeziku pisan je starijim pravopisom, nijedan se ne smatra prikladnim za identifikaciju djela. Kao

usvojeni naslov djela odabire se nadomjesni naslov koji se sastoji od suvremene transkripcije glavnog stvarnog naslova na hrvatskom jeziku. Stvarni naslovi na jedinici građe navode se kao podatak o stvarnom naslovu: Schloss Svečaj = Svečaji vár = Grad Zvečay" (Nacionalna i sveučilišna knjižnica u Zagrebu, Hrvatski državni arhiv, Muzejski dokumentacijski centar, 2019: 190)

3.1. Opisni i nadomjesni naslovi

Opisni i nadomjesni naslovi moraju biti sažeti, jasni i lako razumljivi te se preporučuje dosljednost u uporabi nazivlja. Potrebno je izbjegavati izraze lokalnog govora, kratice, akronime i druge izraze koji mogu biti nerazumljivi korisnicima, a treba sadržavati podatke koji se smatraju značajnim za identifikaciju jedinice građe. Preporučeno je uvrstiti oblik ili žanr jedinice građe, temu jedinice građe (uključujući imena agenata, događaja, proizvoda i tako dalje), ime ili imena stvaratelja, ime ili imena mjesta povezanih s jedinicom građe. Ti podaci se ponavljaju i u drugim elementima u opisu. Opisni naslov mora sažeto opisati kakvo je to djelo ili što djelo prikazuje te navesti povijesne, religijske, mitološke i ostale subjekte gdje je prikladno. Ako je u pitanju mjesto ili vlasnik i njih treba navesti, a ako nema nikakva distinktivna obilježja, navodi se kratki fizički opis djela vezan uz vrstu djela. U slučaju građevina, opisni naslov bi se trebao referirati na vlasnika ako postoji, svrhu građevine (na primjer katedrala) ili adresu ulice u kojoj se nalazi, ovisno o potrebi. Mnoge građevine nemaju imena te se stvara naslov na temelju vrsti djela koje ona predstavlja (na primjer kuća) (Baca i Harpring, 2016). Kao razlikovno obilježje djela odabire se podatak koji se smatra najznačajnijim za identifikaciju djela, a dostupan je u propisanim izvorima podataka. Može se zabilježiti više razlikovnih obilježja djela kad se to smatra potrebnim za identifikaciju. Podatak koji je naveden kao razlikovno obilježje djela može se ponoviti i u drugim odgovarajućim elementima podataka u opisu, a navodi se u okrugloj zagradi na jeziku opisa kad je potrebno razlikovati različita djela s istim naslovom. Kada postoje više razlikovnih obilježja, odvajaju se zarezom, malim početnim slovom osim ako nisu vlastita imena u pitanju (Nacionalna i sveučilišna knjižnica u Zagrebu, Hrvatski državni arhiv, Muzejski dokumentacijski centar, 2019).

3.2. Informacije o prikazu: iz CCO smjernica

Informacije o prikazu (engl. *view information*) sadrže detalje o prikazu djela onako kako se prikazuje na slici. Elementi koji se koriste su opis prikaza (engl. *View Description*), vrsta prikaza (engl. *View Type*), prikaz predmeta (engl. *View Subject*) i datum prikaza (engl. *View Date*). Važno je zabilježiti podatke o prikazu slike bez obzira na format (fotografija, negativ, dijapozitiv i tako dalje) ili vrstu ustanove (knjižnica, muzej, arhiv). Ove informacije su vrlo korisne za detaljne opise jedinice građe koja fizički nije dostupna zbog udaljenosti ili nekog drugog razloga, to jest kada je fotografija tog djela jedini dostupni izvor, te pridonose razumijevanju tog djela. Na primjer, kod više fotografija iste građevine navesti iz kojeg je kuta gledano, to jest navesti rakurse.

3.2.1. Opis prikaza

Opis prikaza je polje slobodnog teksta koje razrađuje prostorne, kronološke ili kontekstualne aspekte djela snimljene u prikazu slike. Upotpunjuje kontekst i razrađuje perspektivu opisujući detalje, dijelove, strane svijeta pri pogledu i tako dalje. Pomaže krajnjem korisniku da razumije informacije unutar slike i razlikuje više slika istog djela. Terminologija bi trebala biti što je moguće dosljednija. Vlastita imena se pišu velikim početnim slovom, a ostale riječi malim početnim slovom (uključujući strane svijeta) te treba izbjegavati skraćenice. Piše se na jeziku kataloškog zapisa te se koristi prirodni red riječi.

3.2.2. Vrsta prikaza

Tip pogleda bilježi specifičnu točku gledišta ili perspektivu, kao što je profilni pogled, krupni plan ili unutarnji pogled. Pomaže korisniku da razlikuje više slika istog djela. Za razliku od opisa prikaza, ima određen i limitiran vokabular korištenja preporučen da je povezan s normativnom (autoritativnom) datotekom. Koristite pojmove koji označavaju položaj, kut, raspon, orijentaciju, opseg ili dio djela prikazanog u prikazu slike. Također, koriste se mala početna slova, izbjegavaju kratice te se piše na jeziku kataloškog zapisa.

3.2.3. Prikaz predmeta

Prikaz subjekta može uključivati pojmove ili izraze koji karakteriziraju predmet rada onako kako je prikazan na slici. Bilježenje predmeta (teme) pomaže u razlikovanju više slika istog

djela i omogućuje krajnjim korisnicima da identificiraju određene slike koje ilustriraju određeni koncept ili detalj. Posebno je koristan za detalje, složene radove, i izgrađena djela koja uključuju mnogo različitih prikaza i detalja. Međutim, postoje slučajevi slika gdje element prikaza subjekta ponekad nije potrebno zabilježiti. Prikaz predmeta je potreban kada je detalj prikazan na slici specifičan za sliku, ali nije primarni fokus djela. Terminologija bi trebala biti kontrolirana korištenjem normativne datoteke ili određenog popisa. Upisuje se malim početnim slovom osim ako nije riječ o vlastitom imenu, kratice se izbjegavaju te se piše na jeziku kataloškog zapisa.

3.2.4. Datum prikaza

Element prikaza datuma uključuje svaki datum ili raspon datuma povezanih sa stvaranjem ili proizvodnjom slike. Nije obavezan element, ali je preporučen kada je poznat točan datum. Na primjer, pojedina građevina je slikana u godini kada je izgrađena i pedeset godina kasnije. Sigurno postoji razlika u njihovom izgledu. Ako bi se dogodio potres, a ne postoji nijedan drugi prikaz zgrade, iskoristila bi se fotografija kasnijeg datuma radi rekonstrukcije građevine. Kod prikaza datuma valja pripaziti i ne pomiješati ga sa drugim datumima koji se zabilježe. Na primjer, neka slika snimljena 1973. godine u formatu dijapozitiva je kopirana u digitalni format 2008. godine. Datum digitalne slike je 2008., ali datum prikaza je 1973. godina. Takve razlike bi trebale biti zabilježene i razjašnjene u zapisu o slici (engl. *Image Record*). Datum izrade (engl. *Creation Date*) kopije općenito se bilježi s drugim administrativnim podacima i ne treba ga brkati s datumom prikaza. Informacije o datumima moraju biti dosljedno oblikovane kako bi se omogućilo učinkovito pretraživanje i dohvaćanje datuma. Moraju se pratiti lokalna pravila. Potrebno je zabilježiti svaki poznati datu, uključujući najraniji datum i najnoviji datum. U primjeru gdje je djelo restaurirano treba naglasiti da je fotografija nastala nakon restauracije.

3.3. Analiza primjera

Ovaj dio se bavi analiziranjem primjera, i uspoređivanje istih, iz javnih mrežnih kataloga u svrhu boljeg shvaćanja te možda lakšeg prevođenja na strani jezik koristeći gore navedena pravila. Primjeri se navode iz raznih izvora, engleskih i hrvatskih.

Korišteni izvori su:

- Kongresna knjižnica (engl. *Library of Congress*)
- Digitalna zbirka i katalog Hrvatske akademije znanosti i umjetnosti
- Zbirka fotografske dokumentacije Ministarstva kulture i medija
- Digitalni repozitorij Instituta za etnologiju i folkloristiku
- Digitalne zbirke Nacionalne i sveučilišne knjižnice u Zagrebu
- Digitalni fundus PPMHP (Pomorski i povijesni muzej Hrvatskog primorja Rijeka)

Korišteni primjeri su:

- 1. Exterior, Isherwood complex, general view - U.S. Naval Academy, Isherwood Hall, Annapolis, Anne Arundel County, MD https://www.loc.gov/resource/hhh.md0947.photos/?sp=1&q=isherwood+hall)
- Lutheran Church of St. Catherine (1768), southeast view, Arkhangelsk, Russia (https://www.loc.gov/item/2018681357/)
- Crkva svete Marije Magdalene (Sela, Sisak): pogled s jugozapada (https://dizbi.hazu.hr/a/?pr=i&id=2324813)
- Novska – Župna crkva sv. Luke – Vanjština – pogled na glavno pročelje sa zvonikom (http://www.fototeka.min-kulture.hr/hr/Predmet/18258/3)
- Seljačka kuća. (https://repozitorij.dief.eu/a/?pr=i&id=12162)
- Seljačka kuća. (https://repozitorij.dief.eu/a/?pr=i&id=12147)
- Zagreb: Katedrala sa Nadbiskupskim dvorom. (https://digitalna.nsk.hr/pb/?object=info&id=22218)
- Zagreb: Hrvatski sokol. (https://digitalna.nsk.hr/pb/?object=info&id=22196)
- Korzo (https://digitalni.ppmhp.hr/?pr=i&id=15381)

Uzet je minimalno jedan primjer iz svakog navedenog izvora.

Na prvi pogled se uočava da je većina naslova drugačije strukturirano. U primjeru 1. Exterior, Isherwood complex, general view - U.S. Naval Academy, Isherwood Hall, Annapolis, Anne Arundel County, MD; 1. (engl. *first*) označava redni broj jer se u toj zbirci nalazi još nekoliko primjeraka fotografija koje se odnose na fakultet U.S. Naval Academy i to na Isherwood Hall kako je nazvano jedno krilo fakulteta. Zatim slijedi opis pa građevina koja se opisuje te iz kojeg gledišta. Vanjština, kompleks Isherwood, opći pogled. Potom dolazi mjesto gdje se

nalazi opisana tvorevina, to jest već spomenuti fakultet i njegovo krilo, lokacija fakulteta ili grad Annapolis koji se nalazi u okrugu Anne Arundel u Marylandu. Drugi primjer iz izvora Kongresne knjižnice je Lutheran Church of St. Catherine (1768), southeast view, Arkhangelsk, Russia i pokazuje nam malo drugačije stvari. Nema opis ili pogled kao početak naslova već naslov djela koje je fotografirano, Luteranska crkva sv. Katarine, zatim godinu izrade iste crkve u okrugloj zagradi te pogled na crkvu, a na kraju mjesto i država gdje se nalazi. Iz ova dva primjera se zaključuje kako oboje prate neku normu naslovljavanja, iako pojedini elementi nedostaju u kojem naslovu (godina). Valja napomenuti da je Lutheran Church of St. Catherine (1768), southeast view, Arkhangelsk, Russia naslov sastavljen od strane autora te takav i zadržan prilikom katalogizacije. Također, bitna činjenica koju treba imati na umu pri ovoj usporedbi jest da je prvi primjer dio veće zbirke pod nazivom U.S. Naval Academy, Isherwood Hall, Annapolis, Anne Arundel County, MD. Stoga, pogled ili opis u naslovu primjera dolaze na prvo mjesto kako bi se korisniku olakšala pretraga za željenim rezultatom. Treći primjer dolazi iz Digitalne zbirke i kataloga Hrvatske akademije znanosti i umjetnosti. Crkva svete Marije Magdalene (Sela, Sisak) : pogled s jugozapada. Naslov djela je u središtu, za čime nastupa mjesto gdje se građevina nalazi u obloj zagradi i na kraju rakurs. Razlika je odmah prepoznatljiva u strukturi gdje u ovom slučaju ide djelo, mjesto, pogled dok u prva dva slučaja ide djelo, pogled, mjesto. Četvrti primjer, Novska – Župna crkva sv. Luke – Vanjština – pogled na glavno pročelje sa zvonikom, dolazi iz Zbirke fotografske dokumentacije Ministarstva kulture i medija te nam odražava malo drugačiji stil naslovljavanja. Počevši s mjestom, zatim na djelo i opis tog djela te na kraju na točku gledišta koja opisuje dodatni detalj fotografije. Detaljnije izraženije nego primjer Crkve svete Marije Magdalene, ali trebalo bi započeti naslovljavanje sa djelom koje je prikazano na fotografiji. Iduća dva primjera su istog naziva, Seljačka kuća, i dolaze iz Digitalnog repozitorija Instituta za etnologiju i folkloristiku, a snimio ih je naš poznati fotograf Tošo Dabac. Nema nikakvih drugih razlikovnih obilježja u naslovu te dolazi do lake zabune. Ovo nije primjer dviju istih seljačkih kuća, već različitih, sa poznatom lokacijom koja ovdje nije navedena već je navedena u elementu Regija i Lokalitet. S obzirom da je lokacija poznata, naslov bi bio mnogo razumljiviji i kvalitetniji da je ista uvrštena. Kad bi se radilo o fotografiji iste kuće, tada bi bilo prikladno staviti u naslov malo detaljniji opis ili smjer iz kojeg je snimljena fotografija. Iako ovaj naslov zadovoljava minimalni uvjet, to jest kratki fizički opis djela vezan uz vrstu djela, trebalo se to odraditi na malo detaljniji način. Zatim imamo dva primjera iz Digitalne zbirke Nacionalne i sveučilišne knjižnice u Zagrebu, Zagreb: Katedrala sa Nadbiskupskim dvorom i Zagreb: Hrvatski sokol. Ova dva primjera nisu fotografije, već

razglednice. U tom slučaju, glavni stvarni naslov bi bio tekst koji se na njima nalazi. Njega pronalazimo u primjeru Zagreb: Hrvatski sokol. Točan natpis na razglednici je ZAGREB: HRVATSKI SOKOL. TRG. Kako se Zagreb: Hrvatski sokol. smatra dovoljno razumljiv i sažet, odbacuje se ostatak te to postaje glavni stvarni naslov. U drugom primjeru nalazimo na prednjoj strani nalazimo natpis ORIENT ZAGREB. Taj natpis se odbacuje jer nije dovoljno kvalitetan, nema distinktivna svojstva te je to ime naklade koja je proizvodila razglednice. Na poleđini razglednice se nalazi tekst Zagreb: Katedrala sa Nadbiskupskim dvorom koji se uzima kao glavni stvarni naslov. Posljednji primjer je Korzo i on je iz Digitalnog fundusa PPMHP-a. Iako fotografija na sebi sadrži natpis S. Hering, Fiume; on nije uzet kao glavni stvarni naslov jer je Salamon Hering snimio tu fotografiju iz čega bi se dalo zaključiti da je ovo autorov naslov. U Digitalnom fundusu PPMHP-a ne piše čiji je naslov. U slučaju da je naslov nadomjesni, nedostatak informacija stvara zbunjenost i pogrešne zaključke. Trebalo je navesti ime autora fotografije, mjesto fotografije i kut gledanja. Dobra stvar je što je to sve navedeno u elementima koji se javljaju. U elementu opisa je detaljno opisano iz kojeg kuta je fotografirano te nadodan opis ulice, ljudi koji se tamo nalaze i građevina koji okružuju Korzo, a u elementu izrade je naglašen fotograf, mjesto i godina izrade fotografije. Zamislimo hipotetsku situaciju u kojoj imamo nekoliko različitih fotografija iste građevine. Na primjer, zgradu Filozofskog fakulteta Sveučilišta u Zagrebu fotografiranu iz različitih kutova te jedan primjerak sa fokusom knjižnice. Sve primjerke pri katalogiziranju nazovemo Filozofski fakultet. Takav naslov je nedorečen jer se lako može misliti na bilo koji Filozofski fakultet u Hrvatskoj te vrlo lako zbuni krajnjeg korisnika i dobije pogrešne rezultate. Kada bi se svaka fotografija nazvala Filozofski fakultet Sveučilišta u Zagrebu, to bi poboljšalo korisnikovo iskustvo i lakše bi došao do željenih rezultata, ali nije baš točno. Zato bi se trebale dodati točke gledišta ili detalj koji iskače radi lakšeg prepoznavanja i razlikovanja. Neki od naslova bi tada glasili Filozofski fakultet Sveučilišta u Zagrebu (pogled s istoka) ili Filozofski fakultet Sveučilišta u Zagrebu (pogled sa zapada na knjižnicu) te bi oni bili puno funkcionalniji. U ovim primjerima nije potrebno dodavati mjesto jer se nalazi u naslovu fakulteta. Da je riječ o nekoj drugoj građevini kao crkvi ili katedrali koja ima isto ime na više mjesta, tu bi bilo primjereno navesti lokaciju gdje se građevina nalazi. Međutim u zapisu o slici bi se trebalo navesti sve elemente od kojih se neki ponavljaju i u naslovu. Tako bi u opisu prikaza pisalo pogled sa zapada na knjižnicu i kafić Filozofija, u vrsti prikaza bi pisala vanjština, u prikazu predmeta bi pisalo knjižnica i kafić, u prikazu datuma bi pisala godina kada je slikana fotografija te na kraju povezana djela sa Filozofski fakultet Sveučilišta u Zagrebu gdje bi poveznica vodila na zapis o djelu i detaljnijim informacijama o samoj građevini.

3.4. Zaključak analize

Iz ovih primjera može se izvući nekoliko zaključaka. Kongresna knjižnica prati pravilnu strukturu nadomjesnih i opisnih naslova, a možda i ostali engleski javni mrežni katalozi. Postoji mogućnost da su samo ova dva primjera vođeni pravilima, a ostali pogrešni. Trebalo bi proći mnogo primjera iz Kongresne knjižnice i ostalih engleskih javnih mrežnih kataloga kako bi se izvukao valjani zaključak. Nadalje, hrvatski javni mrežni katalozi su malo drugačiji i nijedan od primjera ne prati pravila nadomjesnog ili opisnog naslova. Izuzetak su dva primjera razglednica iz Digitalne zbirke Nacionalne i sveučilišne knjižnice u Zagrebu, ali to je tema za glavni stvarni naslov. U više primjera se primjećuje da se prate nekakva pravila o opisivanju, rakursima, lokacijama i datumima, ali na svakom primjeru je drugačiji redoslijed. To nisu novi katalozi, a naša detaljnija pravila su tek uvedena u Nacionalnom pravilniku za katalogizaciju (NPK), iz čega se izvlači zaključak da jednostavno prije nije postojalo valjanih izvora i pravila za praćenje strukture nadomjesnih naslova i razlikovnih obilježja. Stoga, princip navođenja iz primjera iz Kongresne knjižnice se još čini i najlogičnijim. Navodi se redom: ime građevine koja je fotografirana (ako je poznato, ako nije navodi se njena funkcija; na primjer: kuća), godina izgradnje (ako je poznata), opis i pogled na građevinu (na primjer: vanjština, pogled s juga) te mjesto gdje se građevina nalazi. Ovako bi to izgledalo u strukturiranom obliku: Crkva svete Marije Magdalene, pogled s jugozapada, (Sela, Sisak). U nestrukturiranom obliku bi izgledalo malo drugačije: Pogled s jugozapada na Crkvu svete Marije Magdalene u Selima, Sisak. Radi preglednosti je preporučljivo koristiti strukturirani oblik. Također, takav univerzalni oblik bi pomogao u prevođenju na strane jezike.

4. Wikifier

Wikifier je web servis koji uzima ulazni tekstualni dokument i označava ga (anotira) te povezuje sa relevantnim stranicama na Wikipediji. Kod je dostupan svima, potrebna je samo registracija koja je besplatna. Ovdje je prikazana verzija koda za programski jezik Python 3 (Wikifier documentation, n.d., url):

```
import urllib.parse, urllib.request, json


def CallWikifier(text, lang="hr", threshold=1):
    # Prepare the URL.
    data = urllib.parse.urlencode([
        ("text", text), ("lang", lang),
        ("userKey", "ghxhaseywtnbvymajyixltevylnxvj"),
        ("pageRankSqThreshold", "%g" % threshold), ("applyPageRankSqThreshold", "true"),
        ("nTopDfValuesToIgnore", "200"), ("nWordsToIgnoreFromList", "200"),
        ("wikiDataClasses", "true"), ("wikiDataClassIds", "false"),
        ("support", "true"), ("ranges", "false"), ("minLinkFrequency", "2"),
        ("includeCosines", "false"), ("maxMentionEntropy", "3")
        ])
    url = "http://www.wikifier.org/annotate-article"
    # Call the Wikifier and read the response.
    req = urllib.request.Request(url, data=data.encode("utf8"), method="POST")
    with urllib.request.urlopen(req, timeout = 60) as f:
        response = f.read()
        response = json.loads(response.decode("utf8"))
    # Output the annotations.
```

```
    for annotation in response["annotations"]:
        print("%s (%s)" % (annotation["title"], annotation["url"]))
```

CallWikifier("Rak pluća je teška bolest kod pušača")

4.1. Objašnjenje koda i definicije:

Ispod su navedene funkcije, parametri, moduli i tako dalje, ukratko sve što se u kodu nalazi, te njihova kratka objašnjenja kako bi kod bio svima razumljiviji te njegov način rada.

URL, akronim za engleski pojam Uniform Resource Locator, je putanja do određenog sadržaja na internetu, to jest poveznica.

Urllib je paket koji sadrži nekoliko modula za rad s URL-ovima (poveznicama):

- urllib.request (za otvaranje i čitanje poveznica) modul definira funkcije i klase koje pomažu u otvaranju poveznica (najčešće HTTP) – osnovna i sažeta provjera autentičnosti, preusmjeravanja, kolačići i tako dalje. Parametar *timeout* je opcionalan, ali specificira vrijeme za blokiranje operacija (u ovom slučaju pokušaj spajanja) te se izražava u sekundama, a ako nije naveden, uzet će se globalno zadano vrijeme. Ovaj modul zapravo funkcionira za HTTP, HTTPS i FTP veze.
- urllib.parse (za raščlanjivanje URL-a) modul definira standardno sučelje za razbijanje URL *stringova* u komponente (shema adresiranja, mrežna lokacija, put i tako dalje), za kombiniranje komponenti natrag u *stringove* URL-a i pretvaranje „relativnog URL-a" u apsolutni URL s „osnovnim URL-om".
- urllib.parse.urlencode pretvara objekt mapiranja ili niz dvoelementnih ntorki, koji mogu sadržavati *str* ili *bytes* objekte, u kodirani ASCII tekstualni niz. Ako se rezultirajući niz koristi kao podatak za POST operaciju s funkcijom urlopen(), tada bi trebao biti kodiran u bajtove, inače bi rezultirao TypeError.

json – JSON (JavaScript Object Notation) koder i dekoder je format razmjene podataka inspiriran JavaScript sintaksom doslovnog objekta.

Tablica 1: Tablica konverzije

JSON	Python
object	dict
array	list
string	str
number (int)	int
number (real)	float
true	True
false	False
null	None

(Python documentation, n.d., url)

Parametri potrebni za razumijevanje koda:

- userKey: jedinstveni niz od 30 znakova koji identificira svakog korisnika. Ovaj parametar je obavezan.
- text: tekst dokumenta koji se želi anotirati. Koristi se UTF-8 način zapisa.
- lang: koristi se ISO-639 standard za jezik dokumenta i podržani su dvoslovne i troslovne oznake jezika.
- wikiDataClasses: vrijednost treba biti *true* ili *false*, određuje treba li uključiti listu wikidata parova (ID koncepta, naziv koncepta) za sve klase kojima ovaj koncept pripada (izravno ili neizravno) za svaku anotaciju.
- wikiDataClassIds: funkcionira kao i wikiDataClasses, ali generira listu samo ID-ova koncepta (rezultira kraćim JSON *output*-om).

- support: vrijednost treba biti *true* ili *false*, određuje treba li uključiti, za svaku anotaciju, popis podraspona (*subranges)* u ulaznom dokumentu koji podržavaju ovu određenu anotaciju.
- ranges: vrijednost treba biti *true* ili *false*, određuje treba li uključiti, za svaki podraspon u dokumentu koji izgleda kao mogući koncept, listu svih anotacija koji su kandidati za taj podraspon. To će značajno povećati veličinu rezultirajućeg JSON izlaza (*output-a),* pa bi se trebao koristiti samo ako postoji velika potreba za ovim podacima.
- includeCosines: vrijednost treba biti *true* ili *false*, određuje treba li uključiti, za svaku anotaciju, sličnost između ulaznog dokumenta i stranice Wikipedije koja odgovora toj anotaciji.
- maxMentionEntropy: vrijednost se postavlja na stvarni broj *x* kako bi se ignorirala česta dvosmislena ponavljanja (to jest neće pridodati nove kandidate anotacija u procesu). Zadana vrijednost je -1 i ona onemogućuje ovu heuristiku.
- minLinkFrequency: ako se veza sa određenom kombinacijom teksta i cilja pojavljuje na vrlo malo stranica Wikipedije (manje nego što je zadana vrijednost minLinkFrequency), poveznica je u potpunosti zanemarena te se ciljna stranica ne smatra kandidatom anotacije za izraz koji odgovora tekstu veze. Zadana vrijednost je -1 i ona onemogućuje ovu heuristiku.
- pageRankSqThreshold: vrijednost se postavlja na stvarni broj *x* kako bi se izračunao prag za odbacivanje anotacija na temelju njihovog rezultata ranga stranice. Wikifier će izračunati zbroj kvadrata svih anotacija, sortirati ih po padajućem redoslijedu rezultata ranga stranice te izračunati prag tako da zadrži anotacije čiji rang stranice premašuje ovaj prag (bi doveo zbroj kvadrata ranga stranice do SxX). Prema tome, niži *x* rezultira višim pragom i manje anotacija. Zadana vrijednost je -1 i ona onemogućuje ovaj mehanizam. Rezultirajući prag je prijavljen u minPageRank polju JSON objekta rezultata. Ako se žele odbaciti anotacije čiji je rang stranice manji od zadane vrijednosti minPageRank polja, postavlja se parametar applyPageRankSqThreshold na *True* (njegova zadana vrijednost je *False).*
- nTopDfValuesToIgnore: Ako se fraza sastoji od većinom čestih riječ, biti će ignorirana i neće generirati anotaciju. U ovom slučaju, riječ koja se smatra čestom ako je jedna od nTopDfValuesToIgnore najčešćih riječi (u smislu učestalosti dokumenta)

u Wikipediji odgovarajućeg jezika. Zadana vrijednost je 0 i ona onemogućuje ovu heuristiku. Preporuka autora je da se vrijednost postavi na 200 kao dobra polazna točka.

- nWordsToIgnoreFromList: funkcionira na isti način kao i nTopDfValuesToIgnore, samo uz razliku što ne pretražuje najčešće riječi od svih riječi, već koristi najčešće riječi s popisa koji se daje Wikifieru pri pokretanju. Trenutačno postoje popisi za otprilike polovicu podržanih jezika, dobiveni uzimanjem najčešćih riječi i ručnim uklanjanjem nekih koje su se činile korisnima te ne bi trebale biti zanemarene, kao što su na primjer imena mjesta. Ako su navedeni oba parametra, nTopDfValuesToIgnore i nWordsToIgnoreFromList, Wikifier će koristiti popis ako je dostupan za jezik ulaznog dokumenta, a ako je nedostupan, koristit će vrijednosti zadane u nTopDfValuesToIgnore. Automatska zadana vrijednost je -1 i ona omogućuje ovu heuristiku. Preporuka autora je da se vrijednost također postavi na 200 kao dobra polazna točka.

(Wikifier documentation, n.d., url)

4.2. Primjena Wikifiera:

U programski kod se upisuje tekst koji se želi označiti, to jest anotirati, te program ispisuje poveznice koje vode na stranicu Wikipedije (članke) za određene dijelove teksta. Odličan je za svaku upotrebu u potrazi za brzim, jednostavnim i besplatnim informacijama. Raznim funkcijama i parametrima određuje koje riječi će prihvatiti i vratiti u obliku poveznice, a koje ignorirati te koliko će ih vratiti. Rješava problem višeznačnosti riječi te pokušava vratiti točno željeni podatak na način globalnog pristupa. Globalni pristup funkcionira način da tekst shvaća kao cjelinu te se tako i odnosi prema riječima, to jest da su logički povezane (Brank, Leban i Grobelnik, 2017, url). Najbolje možemo prikazati ovim testiranim primjerom. Ulazni tekst je: „rak pluća je teška bolest kod pušača”, a izlazne poveznice na Wikipediju su: rak (bolest), rak pluća, bolest, pušenje. Nije izbačena poveznica za rak kao životinja već po logičkim povezanostima riječi zaključuje da se radi o bolesti. Iako poveznica za rak (bolest) nije direktna poveznica na stranicu raka kao bolesti, već je to poveznica na razdvojbenu stranicu na kojoj su poveznice za sva ostala značenja riječi rak (životinja, horoskopski znak, zviježđe i bolest), prepoznao je da se radi o bolesti te naglasio da bi se moglo raditi o tome.

Način na koji funkcionira je da sakuplja poveznice između stranica koje su međusobno povezane na Wikipediji. Na kraju, slaže ih po PageRank vrijednostima, to jest mjeri važnost web stranice te ispisuje najbolje, a ostale odbacuje. Podržava sve jezike koji se mogu pronaći na Wikipediji, ali nažalost nisu sve internetske stranice dostupne za sve jezike. Neki manje zastupljeni jezici broje tek nekoliko tisuća stranica, ali oni zastupljeniji broje i preko 100 000 stranica. Jezik se vrlo lako promijeni u kodu, za hrvatski jezik je oznaka „hr” (u primjeni: lang=”hr”). Koristeći se trenutno naslovima fotografske građe u javnim mrežnim katalozima ispisuje neke podatke, a nažalost neke ne jer ili nemaju svoj članak na Wikipediji ili još nisu zastupljeni ili nisu dovoljno povezani.

4.3. Primjena Wikifiera na primjerima

Isti primjeri će biti korišteni i u ovom poglavlju te će izlazni rezultati biti uspoređeni i svrstani u tablice. Cilj ovog poglavlja je utvrditi koliko dobro Wikifier funkcionira. Drugim riječima, koje rezultate ostvaruje i koliko su relevantni ili točni za primjere.

Tablica 2: Prvi primjer

1. Exterior, Isherwood complex, general view - U.S. Naval Academy, Isherwood Hall, Annapolis, Anne Arundel County, MD	
Naslov članka:	URl članka:
Exterior (topology)	http://en.wikipedia.org/wiki/Exterior_
Cristopher Isherwood	http://en.wikipedia.org/wiki/Christopher_Isherwood
Plantation complexes in the Southern United States	http://en.wikipedia.org/wiki/Plantation_complexes_in_the_Southern_United_States
General (United States)	http://en.wikipedia.org/wiki/General_
Panorama	http://en.wikipedia.org/wiki/Panorama
United States	http://en.wikipedia.org/wiki/United_States
United States Navy	http://en.wikipedia.org/wiki/United_States_Navy
United States Naval Academy	http://en.wikipedia.org/wiki/United_States_Naval_Academy
Academy Awards	http://en.wikipedia.org/wiki/Academy_Awards
Anne Arundel County, Maryland	http://en.wikipedia.org/wiki/Anne_Arundel_County,_Maryland
Arundel	http://en.wikipedia.org/wiki/Arundel
Maryland	http://en.wikipedia.org/wiki/Maryland

Iz ove tablice se može iščitati da je većina informacija relevantna za tekst koji je unešen u Wikifier iako postoji više informacija nego što je potrebno, a neke su pogrešne. Tako je poveznica za Exterior (topology) pogrešna jer vodi na matematički članak o vanjskim

skupovima. Međutim, zanimljivo je da prepoznaje da se radi o vanjskim skupovima te nas vodi na dio članka gdje je to opisano. Izvorni naslov članka je Interior (topology). Cristopher Isherwood je također pogrešan rezultat. Naime, Cristopher Isherwood je bio književnik, a po njemu nije nazvan kompleks Isherwood, već po Benjaminu Franklinu Isherwoodu koji je bio inženjerski časnik u Američkoj mornarici, a on nije ovdje naveden. Još jedan pogrešan navod je Plantation complexes in the Southern United States što nema veze sa izvornim tekstom. Zadnja pogreška je Academy Awards, članak o dodjeli za Oscara. Manje pogrešni navodi, ali ipak ne i točni bi bili General (United States) i Panorama. General (United States) vodi na članak o općem pojmu generala u vojsci. S obzirom da u tekstu ima mnogo poveznica sa vojskom i mornaricom, naveden je kao izlazni rezultat. Panorama vodi na članak o panorami, vođen riječju *view*. Ostali navodi su točni i smisleno izvedeni, United States od skraćenice U.S., United States Navy kao izvedenica iz United States Naval Academy. Okrug, grad i država su točno navedeni, jedina primjerka je što Annapolis nije naveden kao glavni grad američke savezne države Maryland.

Tablica 3: Drugi primjer

Lutheran Church of St. Catherine (1768), southeast view, Arkhangelsk, Russia	
Naslov članka:	URL članka:
Lutheranism	http://en.wikipedia.org/wiki/Lutheranism
Russian Orthodox Church	(http://en.wikipedia.org/wiki/ Russian_Orthodox_Church)
Church of St. Catherine (Saint Petersburg)	http://en.wikipedia.org/wiki/Church_of_St %2e_Catherine_
Saint	http://en.wikipedia.org/wiki/Saint
Catherine of Alexandria	http://en.wikipedia.org/wiki/ Catherine_of_Alexandria
1768	http://en.wikipedia.org/wiki/1768
Panorama	http://en.wikipedia.org/wiki/Panorama
Arkhangelsk	http://en.wikipedia.org/wiki/Arkhangelsk
Russia	http://en.wikipedia.org/wiki/Russia

Ova tablica nam prikazuje slične rezultate kao i prethodni primjer. Pogrešan navod Ruske pravoslavne Crkve jer je u tekstu spomenuta luteranska crkva. Katarina Aleksandrijska je poveznica samo sa Crkvom svete Katarine te je ona također pogrešan navod. Crkva svete Katarine (Sankt Peterburg), svetac, panorama i 1768 su polovično točni navodi. Možda nepotrebni, ali nisu pogrešni. Crkva svete Katarine (Sankt Peterburg) vodi na članak gdje su pronađene sve Crkve svetih Katarina u svijetu o kojima postoji članak na Wikipediji. O traženoj crkvi iz Arkhangelska ne postoji zabilježen članak na Wikipediji te je iz tog razloga preporučena crkva iz Sankt Peterburga. Svetac, panorama i 1768 su općeniti člancima o pojmovima koje predstavljaju. Članak o godini govori što se sve dogodilo te godine. Točni i najkorisniji navodi su Luteranizam, Arkhangelsk i Rusija. Vode na prave članke te daju sve potrebne informacije.

Tablica 4: Treći primjer

Crkva svete Marije Magdalene (Sela, Sisak) : pogled s jugozapada	
Naslov članka:	URL članka:
Katoličanstvo	http://hr.wikipedia.org/wiki/Katoli%c4%8danstvo
Svetac	http://hr.wikipedia.org/wiki/Svetac
Marija (majka Isusova)	http://hr.wikipedia.org/wiki/Marija_
Marija Magdalena	http://hr.wikipedia.org/wiki/Marija_Magdalena
Magdalena (rijeka)	http://hr.wikipedia.org/wiki/Magdalena_
Sela	http://hr.wikipedia.org/wiki/Sela
Sisak	http://hr.wikipedia.org/wiki/Sisak
Jugozapad	http://hr.wikipedia.org/wiki/Jugozapad

Kao što je prikazano, postoje sličnosti između engleskih i hrvatskih primjera u izlaznim podacima. I ovdje postoji pogreška u vidu Magdalena (rijeka) gdje poveznica šalje korisnika na razdvojbenu stranicu Wikipedije sa izborima koji bi mogli biti točni, a uz to navodi da je najčešći rezultat rijeka. Poveznice koje su smislene i daju širu sliku korisniku su katoličanstvo, svetac, Marija (majka Isusova) i Sela. Nijedan od ovih podataka nije izričito naveden u ulaznom tekstu, osim Sela. Kod Sela se ne javlja pogreška, već poveznica korisnika također šalje razdvojbenu stranicu Wikipedije gdje je prva poveznica za Sela u Sisku. Marija (majka Isusova) je također poveznica na razdvojbenu stranicu Wikipedije gdje opisuje njena značenja. Katoličanstvo i svetac su članci o općenitim pojmovima. Marija Magdalena, Sisak i jugozapad se javljaju u istom obliku kao i u ulaznom tekstu te sa sigurnošću znamo da je o tim pojmovima riječ. Crkva svete Marije Magdalene nema svoj članak na Wikipediji nego je spomenuta u članku o Selima.

Tablica 5: Četvrti primjer

Novska - Župna crkva sv. Luke - Vanjština - pogled na glavno pročelje sa zvonikom	
Naslov članka:	URL članka:
Novska	http://hr.wikipedia.org/wiki/Novska
Župa	http://hr.wikipedia.org/wiki/%c5%bdupa
Župna crkva	http://hr.wikipedia.org/wiki/ %c5%bdupna_crkva
Crkva	http://hr.wikipedia.org/wiki/Crkva
Crkva sv. Luke u Mostaru	http://hr.wikipedia.org/wiki/Crkva_sv %2e_Luke_u_Mostaru
Svetac	http://hr.wikipedia.org/wiki/Svetac
Sveti Luka	http://hr.wikipedia.org/wiki/Sveti_Luka
Zvonik	http://hr.wikipedia.org/wiki/Zvonik

Uzorak je vidljiv i u ovoj tablici izlaznih podataka. Greška je Crkva svetog Luke u Mostaru jer ovaj tekst se odnosi na Novsku, ali jasna je putanja te razlog zašto je ovaj izlazni podatak naveden. Ostali izlazni podaci su relativno točni i smisleni. Župa, župna crkva, crkva, svetac, sveti Luka i zvonik su međusobno povezani pojmovi te daju detaljnije informacije o ulaznom tekstu. Novska je tražen grad te je i to ostvareno. Župna crkva svetog Luke nema svoj članak na Wikipediji te zato i nije pronađen, ali postoji nekoliko natuknica o crkvi u članku o Novskoj.

Tablica 6: Peti primjer

Seljačka kuća	
Naslov članka:	URL članka:
Kuća	http://hr.wikipedia.org/wiki/Ku%c4%87a

Ovaj primjer ima očekivano malo izlaznih podataka, točnije samo jedan, jer su samo dvije riječi ulaznog teksta. Kuća kao članak je točan i jedini koji ovdje može postojati, svaki drugi izlazni podatak bi bio pogrešan. Poveznica vodi na članak koji u općem smislu opisuje kuću.

Tablica 7: Šesti primjer

Zagreb: Katedrala sa Nadbiskupskim dvorom.	
Naslov članka:	URL članka:
Zagreb	http://hr.wikipedia.org/wiki/Zagreb
Katedrala	http://hr.wikipedia.org/wiki/Katedrala

Iako je na slici Zagrebačka katedrala sa Nadbiskupskim dvorom, u ulaznom tekstu ona se ne navodi. Iz tog razloga u izlaznim podacima nema poveznice za Zagrebačku katedralu, već samo za Zagreb kao grad i katedralu kao građevinu u općem smislu. Kada bi ulazni tekst bio Zagreb: Zagrebačka katedrala sa nadbiskupskim dvorom, u izlaznim podacima bi se našla i Zagrebačka i Đakovačka katedrala, ali ne i katedrala kao opći pojam građevine. Nadbiskupski dvor nema svoj članak na Wikipediji, a ni dvor kao pojam. Postoji članak za Dvor kao općinu, ali se ne povezuje sa ulaznim tekstom te je ta mogućnost odbačena.

Slika 1: Zagreb: Katedrala sa Nadbiskupskim dvorom.

Tablica 8: Sedmi primjer

Zagreb: Hrvatski sokol.	
Naslov članka:	URL članka:
Zagreb	http://hr.wikipedia.org/wiki/Zagreb
Hrvati	http://hr.wikipedia.org/wiki/Hrvati
Hrvatski sokol	http://hr.wikipedia.org/wiki/Hrvatski_sokol
Sokolovi	http://hr.wikipedia.org/wiki/Sokolovi

Izlazni podaci ovog primjera su relativno točni. Hrvati kao narod koji se mnogo puta pojavljuje u članku za Hrvatski sokol, a poveznica za sokolove označuje ptice. Iako je na slici građevina koja je još poznata i pod imenom Zagrebačko tjelovježbeno društvo „Hrvatski sokol” (ZTD „Hrvatski sokol”), ona nije u ovom slučaju izlazni podatak nego je Hrvatski sokol kao organizacija. To je u ovom slučaju greška, ali u članku postoji poveznica za ZTD „Hrvatski sokol” iako je to gimnastičarski klub. Sama građevina nema svoj članak na

Wikipediji. Kada bi ulazni tekst glasio Zagreb: ZTD Hrvatski sokol, jedina razlika u izlaznim podacima bi bila zamjena Hrvatski sokol (organizacija) za ZTD Hrvatski sokol (gimnastičarski klub).

Slika 2: Zagreb: Hrvatski sokol.

Zadnji primjer ulaznog teksta je Korzo, ali on ne daje nikakve izlazne podatke. Nije razlog u tome što je jedna riječ (testirano sa Rijeka i daje podatke), već članak na Wikipediji nema dovoljno velik rang stranice (PageRank).

4.4. Analiza rezultata Wikifiera na primjerima

Usporednom analizom engleskih i hrvatskih rezultata, zaključuje se da Wikifier dobro radi, ali potrebno je još poboljšanja. Pokazalo se da izbacuje ne vezane pojmove uz upisani tekst u oba slučaja te krajnjeg korisnika navodi na pogrešnu informaciju. U ovom slučaju je korištena verzija preporučena od strane proizvođača uz jednu iznimku. Vrijednost threshold parametra je postavljena na 1 dok je u originalu postavljena na 0.8. To je napravljeno iz razloga jer vrijednost 0.8 ne donosi mnogo izlaznih podataka. Na primjer, u tekstu Lutheran Church of St. Catherine (1768), southeast view, Arkhangelsk, Russia, izlazni podaci su članci za luteranizam, Arkhangelsk i Rusiju. U drugom primjeru, ali za hrvatski jezik, Crkva svete Marije Magdalene (Sela, Sisak): pogled s jugozapada, izlazni podaci su katoličanstvo, svetac

i Sisak. Parametri koji još utječu na izlazne podatke se mogu mijenjati, međutim ne postoji savršena kombinacija parametara za sve tekstove koji su unešeni. Drugim riječima, kombinacija koja savršeno odgovara jednom tekstu, možda neće nimalo odgovarati drugom tekstu te bi njegovi izlazni podaci bili netočni ili manjkavi ili preopširni. Iz tog razloga odlučeno je napraviti usporedbu i analizu podataka gdje su i netočni i opširni i točni. U jasnim tekstovima gdje ima manje višeznačnih riječi se lakše dolazi do točnih podataka. Taj dio Wikifiera bi se još trebao malo unaprijediti. Također, opširni podaci nisu i netočni, stoga ih je poželjno imati jer su možda u interesu krajnjeg korisnika. Ove analize nas dovode do zaključka da Wikipedija nije sveobuhvatna te da nisu sve tražene informacije zapisane u člancima ili su ti članci vrlo slabo pohranjeni te irelevantni. Zato, ako se dogodi da se nešto ne može pronaći, to ne mora značiti da je Wikifier pogriješio, već je možda traženi članak nepostojeći. Ovaj program je vrlo koristan za pretragu brzih i jednostavnih informacija o stvarima koje zanimaju korisnike. Ima svojih mana, ali i prednosti. Kada bi se usavršio i funkcionirao savršeno za sve unešene tekstove, mogao bi biti svugdje i sve informacije bi bile lako dostupne. U svakoj bazi podataka muzeja, arhiva, digitalnih zbirki bi bio koristan. Pored naslova svakog arhiviranog djela stoji simbol koji se pokreće korisnikovim pritiskom na njega te vraća rezultate kao poveznice na članke Wikipedije kako bi korisnik dobio više korisnih i jednostavnih informacija ako to želi.

4.5. Usporedna analiza Wikifiera i NER-a

Prepoznavanje imenovanih entiteta, ili NER (engl. *named-entity recognition*) je proces klasificiranja entiteta odabranog teksta u unaprijed definirane kategorije. Uspoređeni su rezultati Wikifiera i dva online web servisa za prepoznavanje imenovanih entiteta, RelDIanno za hrvatske tekstove i GATE za engleske.

ReLDIanno je web servis za označavanje teksta koji omogućuje obradu triju južnoslavenskih jezika: slovenski, hrvatski i srpski. Označivač (engl. *The Tagger*) je online alat za obradu teksta koji omogućuje izvođenje četiriju različitih vrsta lingvističkih anotacija: morfosintaktičko označavanje, lematizacija, prepoznavanje imenovanih entiteta te raščlanjivanje. ReLDIanno pri prepoznavanju imenovanih entiteta za označeni tekst prepoznaje pet kategorija: osoba (PER), posvojni pridjev (DERIV-PER), lokacija (LOC), organizacija (ORG) i razno (MISC) (Zupan, Ljubešić i Erjavec, 2017, url). Tako u primjeru Crkva svete Marije Magdalene (Sela, Sisak): pogled s jugozapada prepoznaje iduće entitete:

Marija (B-per), Magdalena (I-per), selo (B-org) i Sisak (B-loc), a ostali entiteti su označeni sa oznakom O. Ovaj web servis koristi IOB2 (inside-outside-beginning) format, drugim riječima, prva stavka u dijelu je označena oznakom B (*beginning)* što znači početak, a sve sljedeće u istom dijelu sa I (*inside*) oznakom. Tokeni koji nisu prepoznati su označeni sa O (*outside*) oznakom, što se prikazuje i u primjeru. Prva primjedba je što je selo označeno kao organizacija i to rezultira pogreškom u ovom slučaju jer je u ovom primjeru Selo naziv sela pokraj Siska. Ovdje se uočava da je manji prag izlaznih rezultata nego kod Wikifiera, a greška je i ovdje postojana. U drugom primjeru, Novska – Župna crkva sv. Luke – Vanjština – pogled na glavno pročelje sa zvonikom, prepoznaje samo: Župna (B-loc), crkva (I-loc), sv. (I-loc), Luke (I-loc). Ne prepoznaje Novsku kao mjesto što je pogreška. Luka nije označen kao osoba jer je ugniježđen, to jest, Župna crkva sv. Luke je označena kao cjelina što je točno. U primjeru seljačke kuće ne prepoznaje nijedan entitet. Zagreb: Katedrala sa Nadbiskupskim dvorom. prepoznaje samo Zagreb kao lokaciju, a također je isti slučaj i sa primjerom Zagreb: Hrvatski sokol. Za primjer Korzo se ne prikazuju nikakvi rezultati.

Za engleske primjere usporedbe odabran je web servis YODIE koji se nalazi u sklopu GATE Cloud-a iz razloga jer funkcionira na slični način kao Wikifier. Cilj mu je rješavanje problema višeznačnosti riječi, to jest tumačenje riječi na samo jedan mogući način. Također, kao i Wikifier, u izlazne podatke šalje poveznicu na stranicu Dbpedija čiji projekt gradi veliku višejezičnu bazu znanja izvlačenjem strukturiranih podataka iz izdanja Wikipedije na 111 jezika (Lehmann et al., 2012, url). U prvom primjeru, 1. Exterior, Isherwood complex, general view - U.S. Naval Academy, Isherwood Hall, Annapolis, Anne Arundel County, MD; YODIE vraća poveznice za Isherwood, U.S. Naval Academy, Annapolis i Anne Arundel County, MD (kao jednu poveznicu). Pri označavanju također navede pouzdanost primjerka, dbpInterestingClasses te dbpSpecificClasses. Kada pogledamo primjer za oznaku Isherwood, dbpInterestingClasses ima vrijednost osoba, a dbpSpecificClasses ima vrijednost pisac. Klikom na poveznicu, šalje nas na internetsku stranicu Dbpedije o Cristopheru Isherwoodu što je pogreška, isto kao i u slučaju Wikifiera jer se ovdje radi o kompleksu koji je nazvan po drugoj osobi. Ostali podaci su točno navedeni. Iz ovog primjera se vidi da postoje pogreške kao i kod Wikifiera, ali ne daje više izlaznih rezultata nego što je potrebno. Fokusira se na tekst i ne traži poveznice po drugim člancima. Zanimljiva je struktura stranice koja sadrži sve poveznice sa Wikipedije i mnoge druge podatke. Drugi primjer iz engleskog jezika, Lutheran Church of St. Catherine (1768), southeast view, Arkhangelsk, Russia donosi slične podatke kao i prvi primjer. Prepoznaje lutheran church, Arkhangelsk i Russia te vodi na točne

stranice. Kako uzima podatke iz Wikipedije, nije pronađena točno tražena i određena crkva jer ne postoji članak Wikipedije o njoj. Razlika između ova dva web servisa je što YODIE uzima samo ulazni tekst u obradu, a ne i ostale moguće poveznice te stoga izbacuje manje rezultata. Izbacivanjem manje rezultata su oni i točniji iako se i ovdje dogodila pogreška, međutim to se nadoknađuje dobrom povezanošću sa ostalim sadržajima i informacijama. Trenutno YODIE podržava engleski, njemački, španjolski i francuski jezik za razliku od Wikifiera koji podržava sve jezike. Međutim, ako bi u kod Wikifiera bilo dopisano nekoliko podataka u predzadnjem redu te bi tada izgledao ovako: print("%s (%s) (%s) (%s)" % (annotation["title"], annotation["url"], annotation["lang"], annotation["dbPediaTypes"])), rezultati bi bili drugačiji, sličniji YODIE web servisu. Annotation [„lang”] nam pokazuje koji je jezik korišten, a annotation [„dbPediaTypes”] nam pokazuje kojeg je tipa izlazni tekst. U 1. Exterior, Isherwood complex, general view - U.S. Naval Academy, Isherwood Hall, Annapolis, Anne Arundel County, MD primjeru su isti izlazni podaci uz navedene dodatke. Za neke izlazne neke podatke ne postoji tip teksta (Exterior, Plantation complexes in the Southern United States, Panorama, Academy Awards). Kako su navedeni podaci većinom netočni i irelevantni za ulazni tekst, on ih ne prepoznaje što je korisno. Pogreška koja se ovdje javlja je već spomenuti Cristopher Isherwood kojem su dodane oznake person, agent, artist, writer. Oznake su točne te se podudaraju sa oznakama iz YODIE-a, ali je osoba pogrešna. Dalje, za svaki rezultat postoje oznake koje su korisne. Kako je već napomenuto, izlazni tekst je opširniji nego kod YODIE-a te tako označava točno ostale primjere. Neki su navedeni kao: Anne Arundel County, Maryland ([„PopulatedPlace”, „Place”, „AdministrativeRegion”, „Region”]); Arundel ([„City”, „Settlement”, „PopulatedPlace”, „Place”]); Maryland ([„PopulatedPlace”, „Place”, „AdministrativeRegion”, „Region”, „State”]). Iz ovih primjera se najbolje prikazuje način na koji su označeni te kojim redoslijedom. Kreće se od osnove kao što je nastanjeno mjesto, zatim mjesto pa sve do termina koji bi obuhvatio i ostale navedene. Iz ovih podataka se iščitava da je Maryland država, Arundel grad, a Anne Arundel County, Maryland regija, to jest poveznice vode na stranicu države, grada i regije. Ovaj dodatak uvelike pomaže krajnjem korisniku u brzoj pretrazi za rezultatima. Uz ovaj dodatak kodu, Wikifier nudi više mogućnosti nego YODIE zbog više izlaznih podataka koji se brzo i jednostavno mogu iščitati te pronaći rezultati koji su potrebni krajnjem korisniku. Također, funkcionira i na hrvatskim primjerima iako su dodatni izlazni podaci napisani na engleskom jeziku.

5. Zaključak

Iz analiza naslova primjeraka na engleskom jeziku prikupljenih iz Kongresne knjižnice se zaključuje da su bolje i dosljednije strukturirani nego primjeri prikupljeni iz hrvatskih izvora. Treba napomenuti da je korišteno više hrvatskih izvora, a samo jedan engleski stoga postoji mogućnost da su primjeri iz nekih drugih engleskih izvora gori nego ovi obrađeni. Međutim, to ne mijenja činjenicu da je svaki primjerak iz drugog hrvatskog izvora imao drugačije strukturiran naslov. U jednima su napisane sve bitne činjenice koje čine razliku između dva dijela, a u drugima su izostavljane iako su znane što je propust u pohrani. Kako je pravilnik o naslovljavanju relativno novijeg datuma, postoji mogućnost da su svi primjeri pohranjeni prije njegova postajanja te se osoba ili institucija služila slobodnom voljom. Ubuduće bi valjalo pripaziti na distinktivna obilježja koja treba navoditi i na koji način. Redoslijed koji bi se mogao pratiti može izgledati ovako: ime građevine ili njena svrha ako je ime nepoznato, zatim datum izgradnje ako je poznat, opis to jest pogled na građevinu te na kraju lokacija gdje se nalazi.

Wikifier se pokazao vrlo korisnim i brzim rješenjem za dobivanjem informacija ulaznog teksta. Iako funkcionalan, ima grešaka na koje treba pripaziti. Donosi više izlaznih podataka nego što je ključno za tekst, a neki su i pogrešni. Iako ima svojih mana, mogu se ispraviti mijenjajući vrijednosti parametara, ali trenutno još ne postoji idealan omjer vrijednosti parametara koji bi donio točne izlazne podatke za svaki tekst. Još je u fazi razvoja, ali je dobar alat za brze i jednostavne informacije iz članaka Wikipedije. Drugi testirani alat, YODIE, iako pravi greške koristan je za dobivanje brzih i osnovnih informacija. Obrađuje samo ulazni tekst te na temelju njega vraća podatke koji su izvučeni iz članaka Wikipedije. Za razliku od ova dva alata, ReLDIanno ne prepoznaje mnogo entiteta ulaznog teksta i ima pogrešaka.

6. Literatura

Baca, M. i Harpring, P. (2016) *Categories for the Description of Works of Art.* Los Angeles: J. Paul Getty Trust i College Art Association. URL: https://www.getty.edu/research/publications/electronic_publications/cdwa/index.html [pristup: 06.09.2022.]

Baca et al. (2006) *Cataloging Cultural Objects: A Guide to Describing Cultural Works and Their Images.* Chicago: American Library Association.

Bowman, J. H. (2010) *Osnove katalogizacije.* Zagreb: Naklada Nediljko Dominović

Brank, J., Leban, G. i Grobelnik, M. (2017) *Annotating documents with relevant Wikipedia concepts.* Ljubljana: Jožef Stefan Institute. URL: https://ailab.ijs.si/dunja/SiKDD2017/Papers/Brank_Wikifier.pdf [pristup: 04.09.2022]

Lehmann et al. (2012) *Dbpedia – A large-scale, Multilingual Knowledge Base Extracted from Wikipedia.* IOS Press. URL: http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf [pristup: 04.09.2022]

Pravilnik za opis i pristup građi u knjižnicama, arhivima i muzejima. Zagreb: Nacionalna i sveučilišna knjižnica u Zagrebu, Hrvatski državni arhiv, Muzejski dokumentacijski centar, 2019.

Python 3.10.6 Documentation. URL: https://docs.python.org/3/ [pristup: 04.09.2022.]

Zupan, K., Ljubešić, N. i Erjavec, T. (2017) *Annotation guidelines for Slovenian named entities. Janes-NER.* URL: https://nl.ijs.si/janes/wp-content/uploads/2017/09/SlovenianNER-eng-v1.1.pdf [pristup: 04.09.2022.]

Wikifier documentation. URL: https://www.wikifier.org/info.html [pristup: 04.09.2022.]

7. Popis slika

Tablica slika

Slika 1: Zagreb: Katedrala sa Nadbiskupskim dvorom.....25

Slika 2: Zagreb: Hrvatski sokol.....26

8. Popis tablica

Indeks tablica

Tablica 1: Tablica konverzije.....15

Tablica 2: Prvi primjer.....19

Tablica 3: Drugi primjer.....21

Tablica 4: Treći primjer.....22

Tablica 5: Četvrti primjer.....23

Tablica 6: Peti primjer.....24

Tablica 7: Šesti primjer.....24

Tablica 8: Sedmi primjer.....25

Analiza naslova fotografske građe u javnim mrežnim katalozima

Sažetak

Ovaj završni rad se može podijeliti na dva dijela. U prvom dijelu se detaljno objašnjavaju svi potrebni pojmovi čije je razumijevanje bitno za daljnji rad na primjerima. Ti pojmovi uključuju katalogizaciju, razliku između djela i slike te naslove. Naslovi su primarni fokus prvog dijela završnog rada. Analizirani su i uspoređeni između nekoliko primjeraka iz hrvatskih i engleskih izvora. Korišteni su javni mrežni katalozi te primjeri nepokretnih spomenika kulture, građevina i ostalih sličnih tvorevina gdje naslov nije dodijeljen od strane autora već se stvara nadomjesni ili opisni naslov.

Drugi dio završnog rada se bavi Wikifierom, to jest programskim kodom u Python 3 programu. Kod je objašnjen, opisan i definiran te korišten u izvornom obliku uz jednu preinaku, kako bi izlazni rezultati bili opširniji radi kvalitetnije analize i usporedbe. Korišteni primjeri su isti kao i u prvom dijelu završnog rada. Također, uspoređen je sa načinom rada ReLDIanno-a i YODIE-a koji služe za prepoznavanje imenovanih entiteta.

Ključne riječi: katalogizacija, naslov, Wikifier

Title analysis of photographic contents in public online catalogs

Summary

This paper can be divided in two parts. In the first part, all necessary terms are explained in detail, because their understanding is essential for further work on the examples. These terms include cataloging, the difference between a work and an image, and titles. Titles are the primary focus of the first part. They were analyzed and compared between several examples from Croatian and English sources. Examples such as immovable cultural monuments, buildings and other similar creations were used from public online catalogs where the title was not assigned by the author, but a descriptive one was created.

The second part deals with Wikifier, that is, the programming code in Python 3 program. The code is explained, described, defined and used in its original form with one modification, so that the output results are more extensive for better analysis and comparison. The used examples are the same as in the first part. Also, it is compared with ReLDIanno and YODIE, which are used to recognize named entities.

Key words: cataloging, title, Wikifier