УНИВЕРЗИТЕТ У НИШУ
ЕЛЕКТРОНСКИ ФАКУЛТЕТ

Милена (Братислав) Фртунић Глигоријевић

УНАПРЕЂЕЊЕ УПОТРЕБЉИВОСТИ ОТВОРЕНИХ ПОДАТАКА ДЕФИНИСАЊЕМ МЕТОДЕ КАТЕГОРИЗАЦИЈЕ ЗАСНОВАНЕ НА МЕТАПОДАЦИМА ПОРТАЛА ОТВОРЕНИХ ПОДАТАКА

ДОКТОРСКА ДИСЕРТАЦИЈА

Ниш, 2023.

UNIVERSITY OF NIŠ
FACULTY OF ELECTRONIC ENGINEERING

Milena (Bratislav) Frtunić Gligorijević

THE ENHANCEMENT OF OPEN DATA USABILITY BY DEFINING A CATEGORIZATION METHOD BASED ON THE OPEN DATA PORTAL METADATA

DOCTORAL DISSERTATION

Niš, 2023.

Подаци о докторској дисертацији

Ментор:	др Леонид Стоименов, редовни професор Електронски факултет Ниш, Универзитет у Нишу
Наслов:	Унапређење употребљивости отворених података дефинисањем методе категоризације засноване на метаподацима портала отворених података
Резиме:	Услед бројних иницијатива које су за циљ имале обезбеђивање отворености и транспарентности података јавних и приватних институција, велики број података је објављен на порталима отворених података. Ради лакшег приступа и боље видљивости ових података, портали су увели могућност претраге података по критеријума као што су претрага по категорији, таговима, формату и организацији. Ове информације о подацима се памте у метаподацима података и уписују се приликом објаве података. Међутим, у пракси метаподаци нису увек потпуни. Недостатак информације о категоријама којима податак припада има велики утицај на видљивост тог податка и доводи до отежаног приступа и коришћења информација. Са растом броја података на порталима све је већи проблем проналажење и препознавање жељених информација уколико категорија није дефинисана. У оквиру ове докторске дисертације урађена је анализа метаподатака на порталима отворених података, анализа употреба категорија и тагова као и њихове повезаности на порталима отворених података. Након тога, адресиран је проблем недостатка информација о категоријама података предлогом методологије за категоризацију података на основу комбинације тагова. У оквиру методологије урађено је дефинисање хијерархијске организације тагова једне категорије у зависности од начина њихове употребе у већ категорисаним подацима, приказан је алат за визуалну анализу хијерархијске организације тагова и дат је предлог поступка за категоризацију података на основу комбинације тагова. Приказани поступак категоризације се ослања на начин употребе тагова већ категорисаних података, односно њихову хијерархијску организацију. У оквиру поступка је дефинисано одређивање сличности између два тага, две комбинације тагова, као и параметара на основу којих се извршава категоризација комбинације тагова са категоријама на порталу. Након тога, дефинисан је алгоритам за предлог категорија којима податак описан комбинацијом тагова треба да припада. За предложени поступак категоризације урађена је евалуација употребом података са канадског портала отворених података. На крају, у оквиру докторске дисертације предложен је модел за допуну метаподатака докумената на порталима.

Научна област:	Електротехничко и рачунарско инжењерство (Рачунарство и информатика)
Научна дисциплина:	Категоризација отворених података
Кључне речи:	отворени подаци, категоризација отворених података, анализа формалних концепата, категоризација
УДК:	(3:001):004.738.5
CERIF класификација:	T 120 - Системски инжењеринг, рачунарска технологија
Тип лиценце Креативне заједнице:	**CC BY-NC-ND**

Data on Doctoral Dissertation

Doctoral Supervisor:	PhD Leonid Stoimenov, full professor, Faculty of Electronic Engineering, University of Niš
Title:	The enhancement of open data usability by defining a categorization method based on the open data portal metadata
Abstract:	Due to numerous data transparency and open government initiatives, a large volume of data was published on open data portals. To make it more accessible and visible, these portals have introduced data filtering by category, tags, format, organization, etc. This information is stored as metadata and provided when publishing the data. However, the metadata is not always complete. The lack of data categories has a great impact on the data visibility, accessibility, and usability of information. As the data increases on the portals, it becomes harder to find and identify the wanted information when the category is missing. Within this doctoral dissertation, an analysis of metadata on open data portals, as well as an analysis of categories and tags usage, and their connections on open data portals was performed. Afterward, the problem of missing data categories was addressed by proposing a methodology for data categorization based on the combination of tags. Within the methodology, the hierarchical organization of tags in a category was defined based on their usage in categorized data. Then, a tool was presented for visual analysis of the hierarchical organization of tags, and a proposal was given for the data categorization based on the combination of tags. The presented categorization relies on the way tags are used in categorized data, i.e. their hierarchical organization. The approach calculates the similarity between two tags, and two combinations of tags, as well as defines the parameters for categorizing the combination of tags with categories on the portal. Afterward, an algorithm was defined that proposes the categories for a dataset with a given combination of tags. For the proposed categorization, an evaluation was performed using the data from the Canadian open data portal. Lastly, within the doctoral dissertation, a model was proposed for supplementing the datasets' metadata on open data portals.

Scientific Field:	Electrical and Computer Engineering (Computer Science)
Scientific Discipline:	Open data categorization
Key Words::	open data, formal concept analysis, open data categorization, cetegorization
UDC:	(3:001):004.738.5
CERIF Classification:	T 120 - Systems engineering, computer technology
Creative Commons License Type:	**CC BY-NC-ND**

Велика захвалност ментору проф. др Леониду Стоименову,
доценту др Милошу Богдановићу, колегама и породици.

САДРЖАЈ

СПИСАК ТАБЕЛА

СПИСАК СЛИКА

СПИСАК КОРИШЋЕНИХ СКРАЋЕНИЦА И ПОЈМОВА

Скраћеница	Попис
ОДП	Портал отворених података
АПИ	Апликациони програмски интерфејс
CKAN	Comprehensive Knowledge Archive Network
DKAN	Drupal Knowledge Archive Network
DCAT	Data Catalog vocabulary
ФЦА	Анализа формалних концепата
URL	Uniform Resource Locator
JSON	JavaScript Object Notation
REST	Representational State Transfer
ЛОД	Linked Open Data

1. УВОД

Велики број иницијатива које би требало да обезбеде отвореност и транспарентност јавних и приватних институција и њихових података [1], довео је до тога да данас отворени подаци представљају један од најзначајнијих фактора за развој и подстицај истраживања у различитим областима. Све ове иницијативе резултирале су стварањем портала отворених података (ОДП), на којима је током година објављена значајна количина података, са намером да се подаци свима учине доступним за коришћење, употребу на иновативне начине и стварање додатне вредности [2].

Број портала отворених података као и количина и разноврсност података која је на њима објављена, у константном је порасту, што узрокује настанак нових изазова за њихову обраду и анализу. Посебно су значајни државни портали отворених података због велике количине података који се на њима објављују као и широког спектра тема и категорија којима ти подаци припадају [3]. На овим порталима се објављују подаци јавних институција из различитих сектора попут привреде, образовања, здравства, транспорта и друге евиденције релевантне за друштво.

Портали отворених података обично су организовани као каталози код којих један скуп података обједињује групу података (ресурса) којима се може приступити или који се могу преузети. Сваки скуп података на порталу праћен је метаподацима који садрже дескриптивне информације о скупу података, организоване у облику парова кључ-вредност, при чему кључ означава које се својство описује, док вредност садржи нумерички или текстуални податак тог својства. Различити портали отворених података организују метаподатке на различите начине, али сви користе шеме унапред дефинисаних поља за памћење одређених информација о подацима који се постављају. Елементи метаподатака могу варирати од шеме до шеме, али увек постоји неки уобичајени скуп елемената попут назива, описа, групе, издавача, ознаке, ресурса, категорије итд.

Корисници приступају подацима коришћењем Веб портала или програмских интерфејса (АПИ). Ради лакшег приступања подацима, сваки портал нуди неки начин груписања и претраге података. Отворени подаци су на порталима врло често груписани по формату у којем су објављени, институцији која их је објавила, категорији којој припадају или на основу кључних речи које ближе описују податке (тагови). Претрага података на основу категорије је интуитиван начин претраге

података пошто омогућава лако проналажење, откривање и комбиновање података из исте области са различитих портала отворених података или у оквиру једног портала. За добијање уског скупа врло прецизних докумената користи се претрага на основу тагова. Међутим, за разлику од категорија чији је скуп обично ограничен, тагови нису унапред дефинисани. Тагове дефинишу корисници приликом постављања сетова података на портал као лични избор речи за које сматрају да најбоље описују неки сет података. Из тог разлога, слични сетови података могу да буду означени различитим вредностима тагова. Самим тим, приликом претраге на основу тагова, сетови података који би требало да буду део резултата претраге, могу бити изостављени из скупа резултата уколико нису означени истим скупом тагова. Последично, резултати претраге на основу тагова не морају увек да дају комплетне резултате.

Приликом објављивања података на порталима није обавезно уношење вредности за све мета-кључеве који су део метаподатака, те се из тог разлога често дешава да неке вредности недостају. Недостатак вредности кључева метаподатака директно утиче на квалитет претраге и доводи до отежаног приступа и коришћења жељених информација на порталима, чиме се смањује потенцијал и корисност података. Ово се посебно односи на недостатак информације о категорији којој податак припада.

Портали отворених података се често употребљавају за преглед скупова података који припадају некој одређеној категорији или ради преузимања свих докумената једне категорије ради касније анализе и обраде. Самим тим, услед недостатка ове информације, део података сигурно неће бити у резултатима претраге, те ће такве податке бити много теже пронаћи и издвојити. Уколико се неки податак не јавља у резултатима претраге када се то очекује и тешко га је пронаћи, тај податак не може бити употребљен за генерисање нових података и информација, чиме вредност и корисност траженог податка значајно опада.

Из тог разлога, дефинисање категорије за сваки скуп података игра изузетно важну улогу у проналажењу жељених података и њихову употребу. Стога, постоји потреба за развојем решења које би на ефикасан и веродостојан начин могло предложити вредности за мета-кључ који представља категорију. Такво решење би обезбедило значајно унапређење квалитета претраге и проналажења свих жељених података.

1.1. Циљеви научног истраживања докторске дисертације

Генерални циљ научног истраживања је предлог решења за категоризацију отворених података на основу вредности тагова који их описују. Предложено решење може да омогући допуну метаподатака некатегоризованих докумената, са акцентом на допуну информација о категорији којој би податак требало да припада. Главни циљ докторске дисертације је да се омогући побољшање доступности података и олакшавање претраге доступних докумената на порталима отворених података.

За реализацију наведеног циља дефинисано је више подциљева на које се ово научно истраживање ослања:

- Анализа употребе описа који се користе у метаподацима за описивање информација о категоријама и таговима на различитим порталима отворених података.
- Анализа доступности и потпуности метаподатака на различитим порталима отворених података са акцентом на проблем недостатка вредности мета-кључева који представљају категорију и ближе описе самих података (тагова).
- Анализа зависности вредности мета-кључева који се користе за описивање тагова и категорија у оквиру које је анализирана употреба различитих вредности тагова у категоријама, популарност тагова у категоријама и њихов значај за описивање докумената исте категорије.
- Дефинисање хијерархијске организације вредности тагова једне категорије у зависности од њиховог начина појављивања у оквиру докумената те категорије.
- Предлог архитектуре и имплементација алата за визуалну анализу хијерархијске организације вредности тагова. Ова анализа може да помогне у сагледавању веза које постоје између тагова и препознавању популарности и важности конкретног тага међу таговима који се користе за описивање докумената исте категорије, као и сагледавању његове позиције у оквиру хијерархије.
- Предлог методологије за категоризацију података на основу метаподатака који описују отворене податке.
- Предлог и имплементација поступка за категоризацију података на основу вредности тагова који описују некатегорисани документ.

- Предлог модела за допуну метаподатака докумената на порталима отворених података са акцентом на допуни информација о категоријама којима одређени документ може да припада.

1.2. Методе истраживања

У оквиру ове докторске дисертације, ради остваривања наведених циљева овог научног истраживања, примењено је више научних метода.

- На почетку је коришћен метод анализе за анализу доступне литературе из области отворених података и портала на којима се они објављују.
- Након тога, коришћен је метод прикупљања података за екстраховање, обраду и анализу података објављеним на порталима отворених података, као и доступних метаподатака сетова података на порталима.
- Метод екстракције биће искоришћен за избор скупа података над којим ће бити извршавани експерименти и евалуација.
- Главни метод истраживања заснован је на експерименталном раду са прикупљеним подацима.
- Резултати експеримената о предложеним категоријама података упоређивани су коришћењем компаративне методе са стварним стањем и категоријама којима они припадају.
- Коришћењем статистичких анализа над добијеним резултатима проверавана је успешност категоризације.
- Метод синтезе искоришћен је за креирање поступка категоризације података на основу вредности тагова који описују отворене податке, као и креирање предлога модела за допуну метаподатака докумената на порталима отворених података.

1.3. Преглед докторске дисертације

Докторска дисертација организована је у десет поглавља. У уводном поглављу описује се предмет и циљеви ове докторске дисертације. Такође, дат је преглед метода који су коришћене у овом раду.

У другом поглављу уводи се појам отворених података, објашњава се њихово значење и значај који имају. Након тога, уводи се појам портала отворених података,

као места на коме се објављују отворених подаци и објашњава се изглед и структура метаподатака који се памте за сваки од објављених података на порталу.

У трећем поглављу приказана је анализа метаподатака на порталима отворених података. У оквиру овог поглавља приказани су резултати анализе метаподатака 40 портала отворених података и урађено је поређење стања на порталима 2020. и 2021. године са акцентом на информације о броју доступних података на порталима, информације о категоријама и таговима који се користе, као и информације о потпуности ових метаподатака.

Након анализе метаподатака, у четвртом поглављу, приказа је анализа употребе тагова у оквиру категорија. Ова анализа је урађена над истих 40 портала отворених података и праћена су стања 2020. и 2021. године. Фокус овог дела докторске дисертације је на праћењу навика додељивања вредности таговима, те су анализиране информације о дужини тагова, понављању истих вредности тагова, као и колико се исте вредности тагова појављују у различитим категоријама.

У петом поглављу је дат увод у категоризацију података на порталима отворених података. Након тога, приказан је предлог методологије за категоризацију података у оквиру кога су приказани неопходни кораци за одређивање категорија подацима на основу вредности њихових тагова.

У шестом поглављу, приказано је креирање хијерархије употребе тагова у оквиру једне категорије. На почетку овог поглавља, дат је преглед Анализе формалних концепата. Након тога, приказано је како се овај метод може употребити за креирање хијерархије тагова на порталима отворених података на основу начина њиховог појављивања у оквиру категорија. На крају поглавља, приказан је значај визуалне анализе и представљен је интерактивни алат за визуализацију мрежа концепата креираних Анализом формалних концепата који омогућава лак преглед и анализу сложених мрежа концепата.

Након тога, у оквиру седмог поглавља приказан је поступак категоризације података на основу тагова којима су описани. Детаљно је објашњен процес претпроцесирања и креирања базе знања на основу које се врши категоризација. Након тога, приказана је категоризација и објашњен је поступак за рачунање сличности између појединачних тагова, као и комбинација тагова. У овом делу су дефинисани и параметри на основу којих се врши категоризација података и начин како се они употребљавају у овом процесу. На крају овог поглавља приказана је имплементација

предложеног поступка категоризације и пример употребе АПИ-ја и Веб апликације који имплементирају приказани поступак.

Евалуација представљеног поступка категоризације приказана је у поглављу осам. Урађена је детаљна анализа употребе овог поступка над подацима из 2022. године са канадског портала отворених података.

У деветом поглављу је предложен модел за допуну метаподатака докумената на порталима отворених података. Предложен је начин за припрему система за категоризацију, објашњено је како приказани поступак категоризације може да се примени и предложен је начин верификације и ажурирања промена.

Десето поглавље садржи дискусију и закључак ове докторске дисертације. У оквиру овог поглавља је урађен кратак резиме докторске дисертације и приказане су неке од карактеристика приказане методологије за категоризацију података. Након тога, дати су правци даљег истраживања.

На крају дисертације налазе се два додатка. У оквиру прилога А дат је преглед портала отворених података који су коришћени у анализама приказаним у поглављима четири и пет. Прилог Б садржи један пример метаподатака у JSON формату које корисник може да добије са портала који користи CKAN платформу.

2. ОТВОРЕНИ ПОДАЦИ И ПОРТАЛИ ОТВОРЕНИХ ПОДАТАКА

На почетку ове докторске дисертације потребно је увести појмове отворених података и портала отворених података. Из тог разлога, у наставку овог поглавља објашњен је концепт отворених података, концепт портала отворених података и основних карактеристика портала отворених података.

2.1. Концепт отворених података

Концепт отворених података у смислу дељења информација и знања није нов али је добио нови смисао и значење са развојем интернета и појавом различитих иницијатива које промовишу бољу информисаност, транспарентност и отворену управу.

Постоји више дефиниција отворених података али све оне теже концепту да отворени подаци могу слободно да се користе, употребљавају и деле. Open Knowledge фондација је у оквиру свог пројекта Open Definition дефинисала отвореност када се говори о подацима и садржају на следећи начин: „Отворено значи да свако може слободно да приступа, користи, мења и дели у било коју сврху (подложно, једино, захтевима који чувају порекло и отвореност)“, односно сажетије „Отворене податке и садржај свако може слободно да користи, мења и дели у било коју сврху” [5]. У приручнику за отворене податке (Open Data Handbook) [6] истакнуте су три најбитније карактеристике отворених података:

- Доступност и приступ подацима – подаци морају бити доступни у целини, у формату који омогућује њихово мењање и по могућности преузимање преко интернета.
- Поновна употреба и дистрибуција података – подаци треба да буду доступни тако да је дозвољена њихова поновна употреба и дистрибуција, укључујући и повезивање и комбиновање са другим скуповима података.
- Свако мора да буде у могућности да користи, поново користи и дистрибуира податке. Не треба да постоји дискриминација према областима употребе ових података или према особама или групама које те податке могу да користе.

Отворени подаци управа представљају податке од јавног значаја који су у власништву управе и који могу слободно да се користе без додатних ограничења. Подаци управе обично се састоје од докумената који се односе на привреду, здравство, транспорт, образовање и друге области релевантне за друштво.

Радна група је на радионици 2007. године одредила осам принципа отворених података управа које дефинишу основне карактеристике које би подаци управе требало да имају да би били „отворени“ [7]. Данас су ови принципи глобално прихваћени и представљају главне смернице отворених података управе:

- комплетни – сви јавни подаци треба да буду доступни.
- примарни – подаци треба да буду објављени у свом изворном облику са што већом грануларношћу, без претходне агрегације или измена.
- правовремени – треба да се објаве довољно брзо како би се очувала њихова вредност.
- машински-читљиви – подаци треба да буду структуирани тако да омогуће аутоматску обраду.
- доступни – доступни што већем броју корисника како би могли да буду употребљени за што већи број намена.
- не-власнички – формат у коме се подаци објављују треба да буде не-власнички како не би било додатних ограничења у погледу приступа и коришћења података, као и да би подаци остали употребљиви у будућности. Овај принцип подразумева да формат података треба да буде добро прихваћен или да се подаци објаве у више формата како би били доступни свима.
- без дискриминације – подацима може свако анонимно да приступи.
- са отвореном лиценцом – подаци не подлежу ауторским правима, патентима и пословним тајнама. Могу да буду дозвољена само разумна ограничења приватности, безбедности и привилегија.

Такође, закон о електронској управи Републике Србије, објављен у "Службеном гласнику РС", број 27, 6. априла 2018. дефинисао је да су отворени подаци они подаци који су доступни за поновну употребу, заједно са метаподацима, у машински читљивом и отвореном облику [8].

Отворени подаци добијају на све већем значају услед бројних захтева за отварањем како јавних, тако и приватних организација [4]. Предности отварања података су

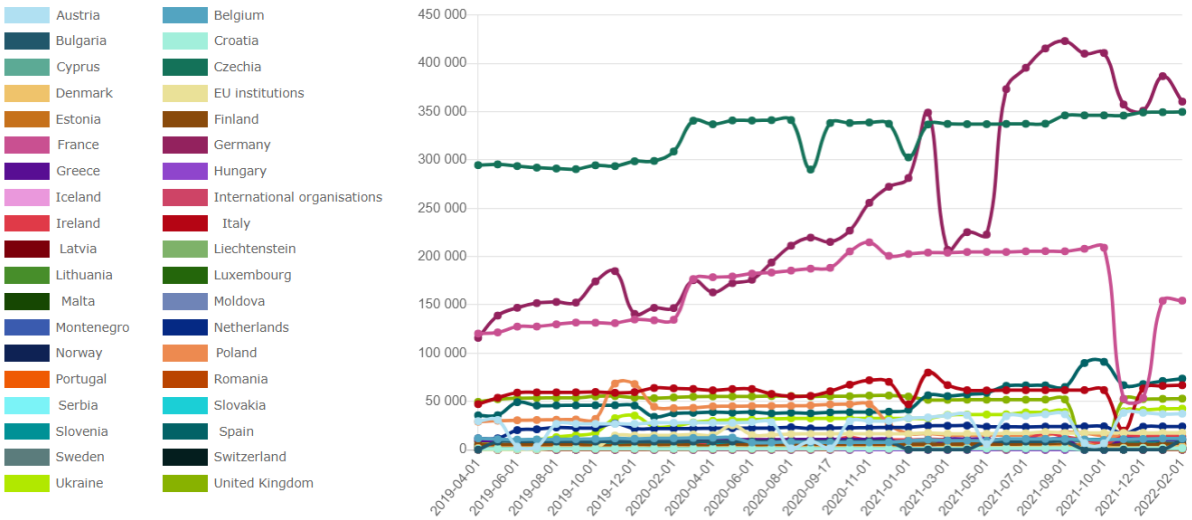
многобројне, како за институције које их објављују, тако и за привреду, академски сектор и све грађане. Сматра се да су отворени подаци главни покретачи отворене управе [9]. Управе широм света уочиле су да отварање података може да подстакне привредни раст, повећава квалитет својих услуга и учини јавну управу ефикаснијом, економичнијом и транспарентном за грађане [10].

Отварањем јавних података повећава се учешће грађана којима се даје могућност увида у рад управе, праћење активности управе и постиже се боља информисаност друштва. Тиме се постиже боље разумевање рада институција, повећање транспарентности, а са тим и повећање поверења грађана, смањује се простор за корупцију и подржава се демократско друштво. Поред тога, отварањем података подстиче се и повећање квалитета јавних услуга, креирање нових услуга, смањење трошкова, као и повећање ефикасности рада управа и смањење времена неопходног за размену података између управа.

Отворени подаци су постали велики покретач за економски раст. Јавне институције производе и обрађују највише скупова података из различитих области. Отворени подаци могу допринети процесима доношења пословних одлука, развоју иновативних услуга и креирању нових сервиса и пословних модела, који позитивно могу да допринесу отварању нових компанија, нових радних места и порасту економије. На пример, у извештају Европске комисије из 2015. године предвиђено је да ће 75.000 радних места бити директна последица отварања података, и да ће тај број достићи 100.000 до 2020. године [11]. У извештају из 2017. године који су припремили *IDC* и *Open Evidence* за Европску комисију наводи се да је вредност европске економије података порасла са 285 милијарди евра у 2015. години на 300 милијарди евра у 2016, а очекује се да ће порасти на 739 милијарди евра у 2020 [12]. У студији из 2018. године, коју је финансирала Европска комисија, на тему ревизије Декларативе 2003/98/ЕЦ о поновној употреби информација из јавног сектора, процењено је да је директна економска вредност информација јавног сектора 2018. износила око 52 милијарде евра [13]. У овом извештају предвиђа се да ће 2028. године, у зависности од тога да ли ће бити промена у регулативама, укупна вредност информација јавног сектора бити од 150 до 215 милијарди евра.

2.2. Концепт портала отворених података

Вредност отворених података лежи у њиховој даљој употреби, па је за организације и појединце који објављују податке битно да се обезбеде управо механизми који омогућавају лак и ефикасан приступ подацима и њихово проналажење [14]. Из тог разлога, државне и јавне институције објављују своје податке на порталима отворених података (ОДП) који су специјално дизајнирани да испуне ове захтеве. Као један од резултата великог броја иницијатива за отварање података и управа креиран је велики број портала отворених података. Број портала као и количина података која је на њима објављена је у константном порасту. На пример, у истраживању на тему квалитета отворених података, аутори рада [2] су разматрали 259 портала отворених података на којима се налазило више од милион скупова података. Поред тога, на званичном порталу отворених података Европе, августа 2019. године, налазило се мање од 900.000 сетова података, док се данас прикупљају подаци из 36 земаља и налази се преко 1.300.000 сетова података [15]. На слици 1 дат је приказ промена у бројевима сетова података по државама на порталу отворених података Европе у периоду од 01.04.2019. године до 01.02.2022. године [16]. Може се приметити да је на већини портала забележен пораст количине података која је на њима објављена. Поред тога, портал https://dataportals.org/ садржи метаподатке о 551 порталу отворених података различитих међународних организација, држава, региона и локалних самоуправа.



Слика 1 – Број сетова података на Европском порталу отворених података по земљама у периоду од 01.04.2019. до 01.02.2022. године [16]

Како би се омогућио што лакши и ефикаснији приступ подацима, корисницима се нуди приступ коришћењем Веб портала или апликациони програмских интерфејса (АПИ). Обично портали отворених података користе неке од постојећих платформи креираних за ову намену. Неке од платформи које се најчешће користе су Comprehensive Knowledge Archive Network (CKAN), Drupal Knowledge Archive Network (DKAN), Socrata и Opendatasoft.

Портали отворених података су обично организовани као каталози који садрже сетове података, где сваки сет агрегира групу података – ресурса. Сваком сету података и његовим ресурсима може да се приступи и могу да се преузму. Сетови података на порталима отворених података су праћени метаподацима који их ближе описују. Метаподатке о подацима уносе корисници приликом објављивања података и они представљају дескриптивне информације у структуираном формату које олакшавају даљу употребу и коришћење података. Метаподаци су организовани као парови кључ-вредност, при чему кључ представља својство податка које се памти док је вредност нумерички или текстуални податак који одговара кључу.

Различити портали отворених података организују метаподатке на различите начине. Сваки портал дефинише сопствену структуру и ограничења по питању метаподатака који прате скупове података, али сви садрже уобичајене информације о подацима попут наслова, описа, групе, издавача, ресурса и слично. Имена кључева у оквиру шема варирају, али се обично структура метаподатака ослања на Data Catalog vocabulary (DCAT) [17], речник који омогућава издавачу да опише скупове података у каталогу користећи стандардни модел чиме се олакшава употреба и агрегација метаподатака из више каталога [18]. Међутим, поред стандардних информација, шеме могу да садрже и неке додатне кључеве који могу да се користе за памћење додатних информација о подацима, карактеристичних за неке податке и портале. Детаљан преглед и анализа шеме метаподатака имплементираних у оквиру четири платформе за објављивање отворених података CKAN, DKAN, Socrata и OpenDataSoft аутори су приказали у раду [17].

3. АНАЛИЗА МЕТАПОДАТАКА НА ПОРТАЛИМА ОТВОРЕНИХ ПОДАТАКА

Подаци који се објављују на порталима отворених података као и њихови метаподаци нису увек потпуни [24]. Из тог разлога, многе земље широм света помериле су свој фокус са акцента на количину података која је јавно доступна на квалитет самих података и њихових метаподатака. Очекује се да ће побољшање квалитета метаподатака осигурати правилну примену и употребу података и подстаћи развој висококвалитетних апликација које могу да користе доступне отворене податке.

Европска комисија, у свом извештају о зрелости отворених података за 2021. годину [18], посматра отворене податке кроз четири димензије: политика, утицај, портал и квалитет. У овом извештају, посебна пажња је посвећена четвртој димензији, квалитету, која се односи на „квалитет отворених података“, односно, обезбеђивању квалитета како података тако и метаподатака. Како се наводи у извештају, Комисија покушава да „обезбеди систематско и благовремено прикупљање метаподатака, као и механизме праћења како би се обезбедило висококвалитетно објављивање метаподатака, у складу са стандардом DCAT-AP, и у складу са неколико захтева за квалитетно објављивање“ [18]. У овом извештају се истиче да је димензија квалитета побољшана само ограничено, што је чини најмање зрелом димензијом 2021. године.

Научна заједница наглашава важност метаподатака за откривање и поновну употребу података. Различита истраживања сугеришу да ће недовољно метаподатака значајно омети аутоматско откривање скупова података [20], да исправна интерпретација и употреба отворених података неће бити могућа ако скупови података нису спојени са тачним метаподацима [21] и да је откривање скупа података уско везано за метаподатке [22].

Обзиром да је количина отворених података у константном порасту и да она са собом доноси разноликост података, захтевају се методе за ефикасан приступ и извршавање упита над подацима. Портали отворених података су најчешће само приступна тачка ка подацима. Стога је истраживачка заједница преусмерила свој фокус на испитивање перформанси портала отворених података. Ова истраживања су углавном фокусирана, како на квалитет метаподатака, тако и на њихове недостатке. Квалитет метаподатака захтева стално праћење, те је из тог разлога пожељно

коришћење платформи за аутоматску евалуацију квалитета података [2][4][22]. Имплементирањем и коришћењем аутоматских алата за процену квалитета метаподатака, перформансе портала отворених података би биле побољшане кроз побољшану комплетност метаподатака.

У оквиру овог докторске дисертације посебна пажња дата је делу метаподатака који се односи на претрагу података на порталима. Како би се обезбедило лакше проналажење жељених резултата, портали отворених података обично нуде више критеријума за претрагу података. Неки од уобичајених критеријума су претрага на основу категорије којој податак припада, на основу тагова који ближе описују податке, по организацији која је податке објавила, формата податка и слично. Претрага на основу категорије и тагова је посебно значајна, обзиром да представља природне механизме за проналажење жељених информација.

Портали отворених података обично имају ограничен скуп унапред дефинисаних категорија којима подаци могу да припадају. У зависности од портала, податак може да припада једној или више категорија. Са друге стране, тагови представљају листу кључних речи и фраза које ближе описују скуп података и могу се користити за прецизну претрагу скупова података. Тагове уносе корисници приликом објављивања података, њихове вредности нису унапред дефинисане и представљају лични одабир термина за које корисник сматра да су битни за означавање податка. Из тог разлога, аутори су у раду [22] дефинисали да се категорије бирају из контролисаног речника који омогућава интуитивно претраживање скупова података и приказује тренутни преглед доступних података у каталогу, док таговима недостаје тај капацитет због њихове ниске конзистентности.

У оквиру ове докторске дисертације урађена је анализа 40 портала отворених података и поређење стања на овим порталима у јулу 2020. године и децембру 2021. године, са циљем да се утврди тренд раста доступних података као и да се анализира употреба категорија и тагова у оквиру метаподатака [24]. Из тог разлога, фокус анализе стављен је на:

- анализу употребе мета-кључева за чување информација о категоријама и таговима,
- промене у укупном броју сетова података,
- промене у броју категорија на порталу,
- промене у укупном броју тагова на порталу,

- промене у броју сетова података којима није додељена категорија,
- промене у броју сетова података који нису описани таговима,
- промене у броју категорија којима један сет припада,
- промене у броју тагова који се употребљавају за описивање једног сета података на порталу.

У оквиру ове анализе обухваћено је 10 портала који употребљавају Socrata платформу и 30 портала који употребљавају CKAN платформу. Информације о свим порталима који су обухваћени овом анализом дат је у Прилогу А ове докторске дисертације.

Иако се структура метаподатака на овим платформама разликује, слични мета-кључеви се користе за описивање информација које су предмет анализе у овом истраживању. Корисници Socrata платформе употребљавају мета-кључ *category* за чување информације о категорији и мета-кључ *tags* за чување информације о таговима. На CKAN платформи, примењују се мета-кључеви *tags* за чување информације о таговима и *groups* за чување информације о категорији. Међутим, анализом је установљено да не користе сви корисници CKAN платформе предвиђене мета-кључеве за ову намену, већ да употребљавају и неке друге ознаке за чување ових информација. Ово је посебно изражено за чување информација о категорији где 26% анализираних корисника користи друге ознаке, док за чување информације о таговима само канадски портал отворених података користи другачији мета-кључ. Преглед ознака које се употребљавају на анализираним порталима дат је у табели 1.

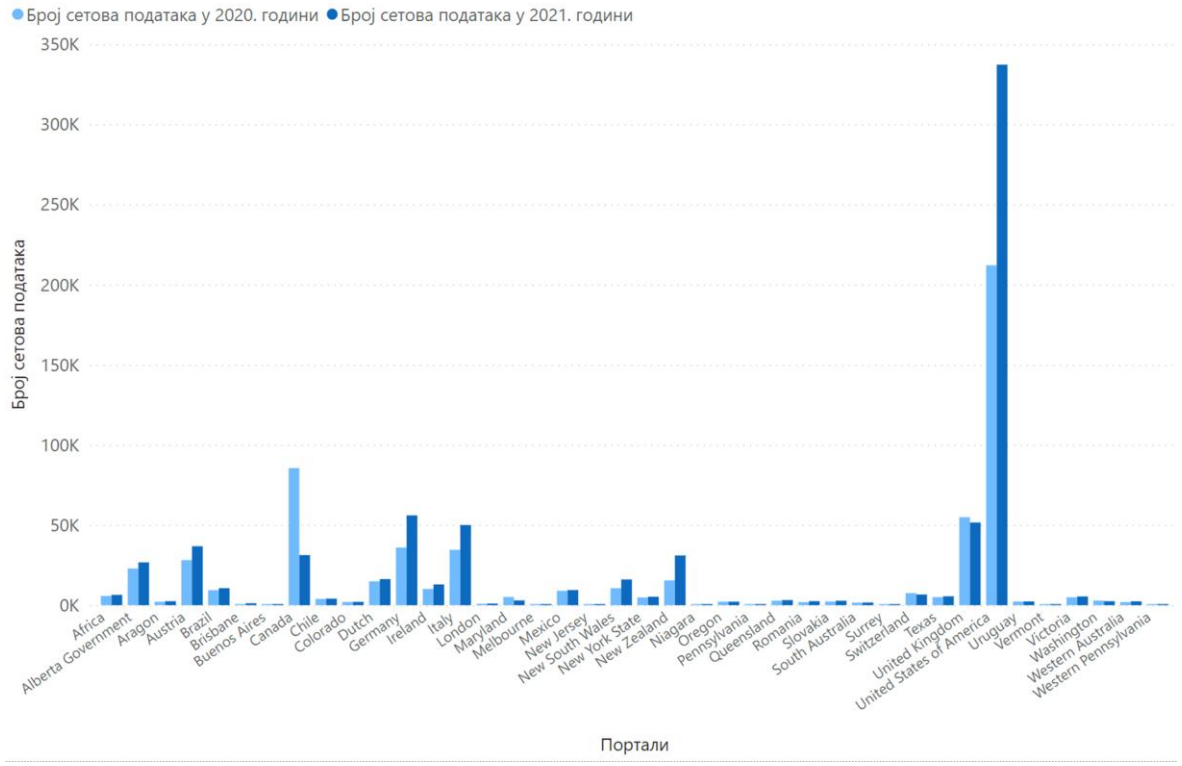
Табела 1 – Мета-кључеви који се употребљавају за чување информација о категоријама и таговима на анализираним порталима отворених података.

Платформа	Мета-кључ за категорије	Мета-кључ за тагове
CKAN	groups sector theme topic theme-primary; theme-secondary subject categorization	tags keywords
Socrata	category	tags

3.1. Промене у укупном броју сетова података

У анализираном периоду већина портала отворених података забележила је повећање укупног броја сетова података на порталу. Преглед броја података по порталима у 2020. и 2021. години приказан је на слици 2.

На пет анализираних портала отворених података број сетова је увећан преко 50% у овом периоду. Са друге стране, осам портала је забележило смањење укупног броја сетова података. Процентуално највеће смањење забележено је на канадском порталу отворених података на коме је укупан број сетова смањен за 63,45%, услед консолидације података [25]. Просечно, на нивоу свих портала у анализираном периоду забележен је раст броја сетова података од 20,77%. Међутим, упоређивањем идентификационих бројева сетова података на порталима, примећено је да неки идентификациони бројеви сетова података из 2020. године нису били присутни приликом преузимања података 2021. године. На 23 портала преко 90% сетова података преузетих 2020. године је било доступно и приликом преузимања наредне године. С друге стране, на 8 портала отворених података тај проценат је износио до 50%, чиме је просечно 80,1% скупова података било доступно приликом оба преузимања података на нивоу свих анализираних портала.

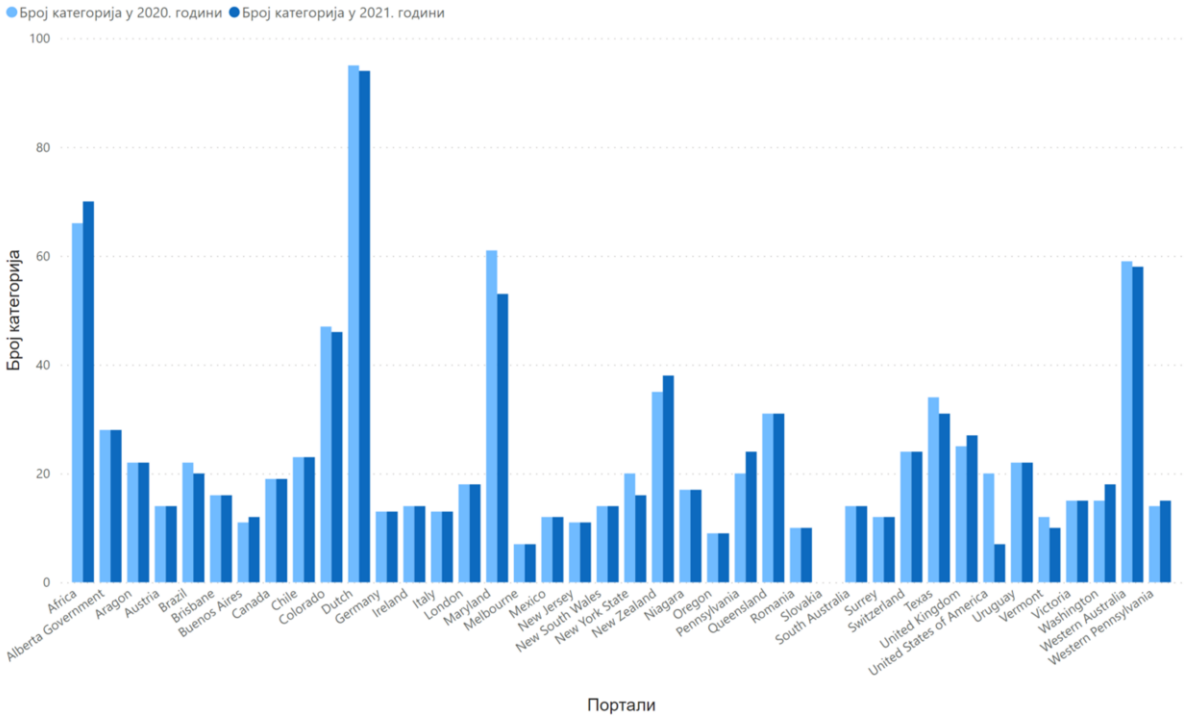


Слика 2 – Преглед броја података на порталима отворених података у 2020. и 2021. години

3.2. Промене у броју категорија на порталу

У оквиру анализе промена у броју категорија на порталима, приликом анализе броја доступних категорија на порталима отворених података примећено је да само један од анализираних портала нема дефинисане категорије, док је на осталим порталима дефинисан већи број категорија. Број доступних категорија варира од портала до портала. На неким порталима је тај број прилично велики док је на другим релативно мали. Анализом је установљено да је средња вредност броја категорија на порталима износила 17,5 категорија 2020. године а 16,5 2021. године.

Уколико се посматрају промене у броју категорија на порталима може да се примети да је у посматраном периоду број доступних категорија остао исти на 60% анализираних портала. На неколико портала је забележено смањење укупног броја категорија, при чему је највећа разлика у броју категорија забележена на порталу Сједињених Америчких Држава. Такође, на 7 портала забележено је повећање броја категорија. Детаљан преглед броја категорија на порталима у анализираном периоду приказан је на слици 3.



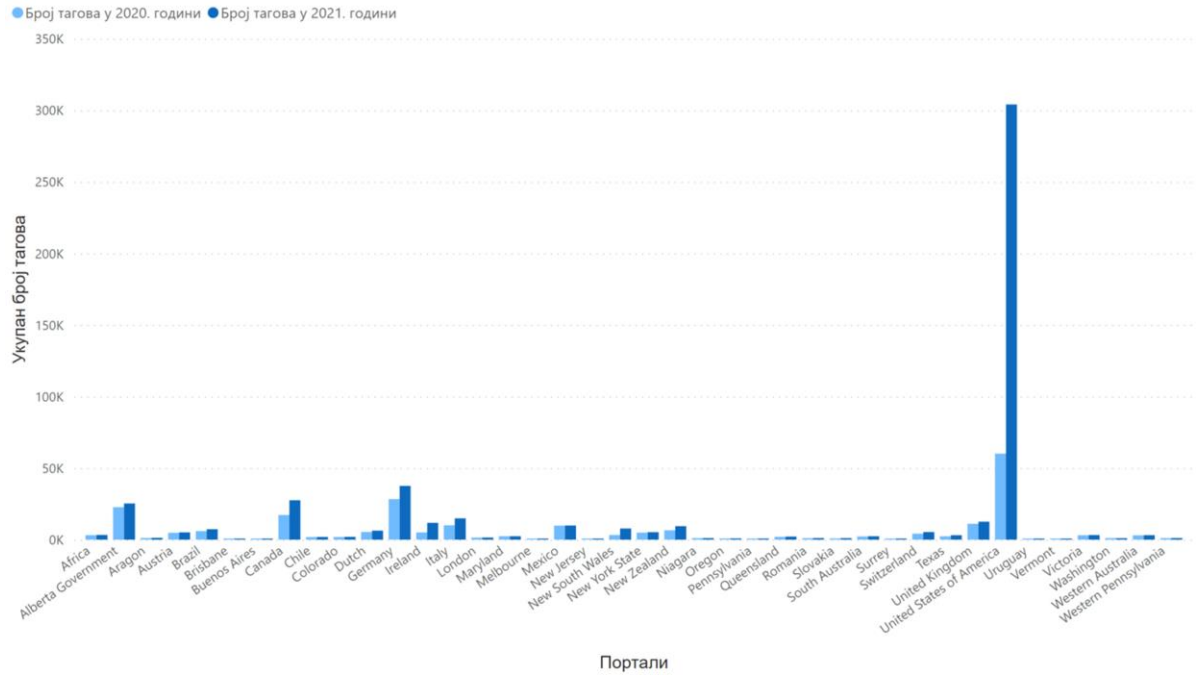
Слика 3 – Преглед броја категорија на порталима отворених података у 2020. и 2021. години

3.3. Промене у укупном броју тагова на порталу

Приликом посматрања промена у укупном броју тагова на анализираним порталима, примећено је да се у анализираном периоду број тагова на порталима повећао у просеку за 28,48%.

Четири портала је забележило пад укупног броја тагова, при чему је портал Буенос Ајреса једини забележио велико смањење броја тагова од 52%. На осталим порталима тај проценат је износио мање од 10%, док је на 2 портала био испод 1%.

Са друге стране, већина портала је имала раст броја тагова, при чему се посебно истичу портали Ирске, Новог Јужног Велса и Сједињених Америчких Држава на којима је број тагова значајно увећан. Преглед броја тагова на порталима у анализираном периоду приказан је на слици 4.

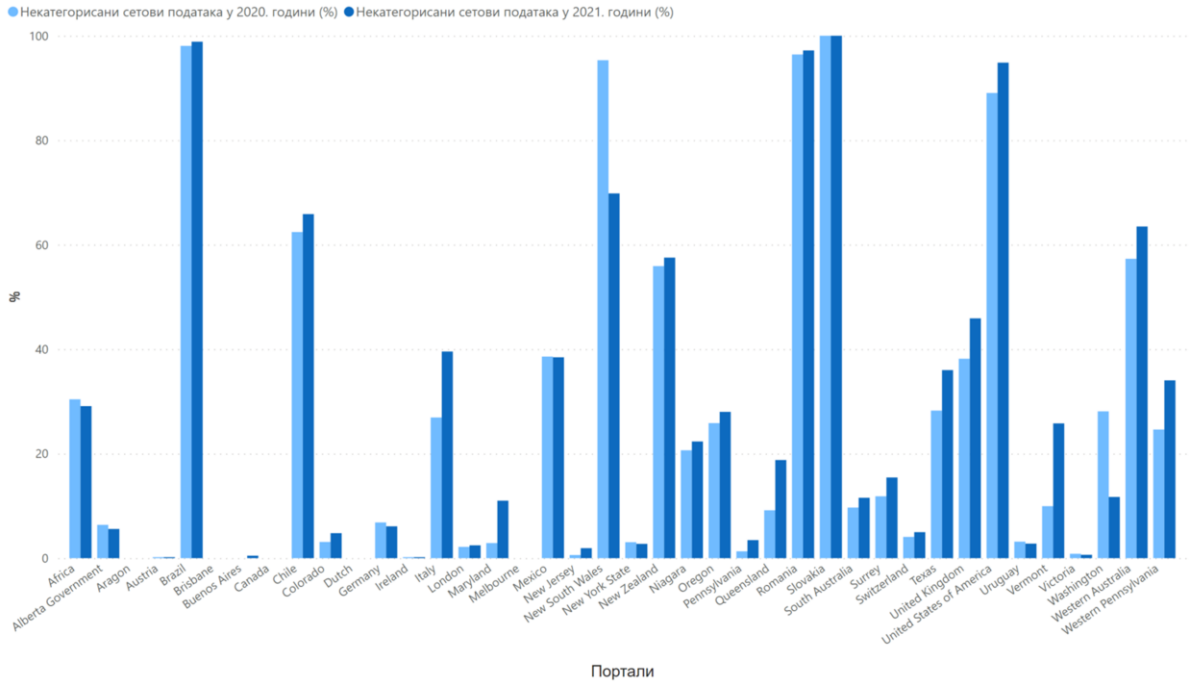


Слика 4 – Преглед броја тагова на порталима отворених података у 2020. и 2021. години

3.4. Промене у броју сетова података којима није додељена категорија

Анализом промена у броју сетова података којима није додељена категорија, закључено је да од укупно 40 анализираних портала отворених података, само 6 портала имало додељене категорије свим сетовима података 2020. године. Притом је пет од шест наведених портала задржало тај тренд и 2021, док на шестом порталу два

сета података нису имала додељену категорију. Са друге стране, четири портала са дефинисаним категоријама је имало јако слаб проценат категорисаности података, са 89% и више података без додељене категорије 2020. године. Словачки портал отворених података, као портал без дефинисаних категорија је једини портал са 100% некатегорисаних података. Од ових портала само је портал Новог Јужног Велса смањио број некатегорисаних података 2021, док је на осталим порталима проценат остао сличан или је повећан. У просеку, проценат некатегорисаних података 2020. је износио 24,78%, док је 2021. тај проценат износио 26,28%. Преглед процената некатегорисаних података на порталима у анализираном периоду приказан је на слици 5.

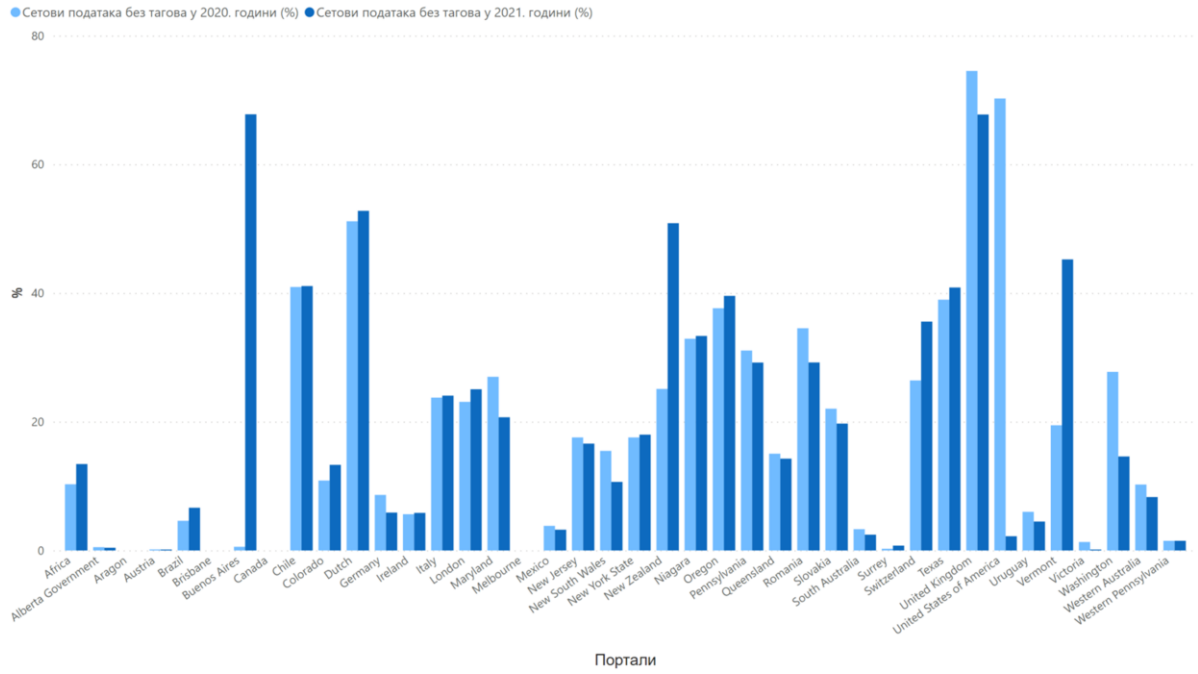


Слика 5 – Проценат података којима није додељена категорија по порталима отворених података у 2020. и 2021. години

3.5. Промене у броју сетова података који нису описани таговима

Анализом промена у броју сетова података који нису описани таговима, примећено је да повећање укупног броја тагова није утицало на број скупова података без пратећих кључних речи. Неколико портала отворених података, попут портала Сједињених Америчких Држава, имало је значајно смањење броја података без тагова. Портали који су већ имали добар проценат означених скупова података задржали су тај тренд. Међутим, скоро половина портала је имала повећање броја скупова података без тагова, при чему је укупан проценат неозначених скупова података у просеку порастао

са 18,5% на 19,14%. Преглед процената података на порталима који нису описани таговима у анализираном периоду приказан је на слици 6.

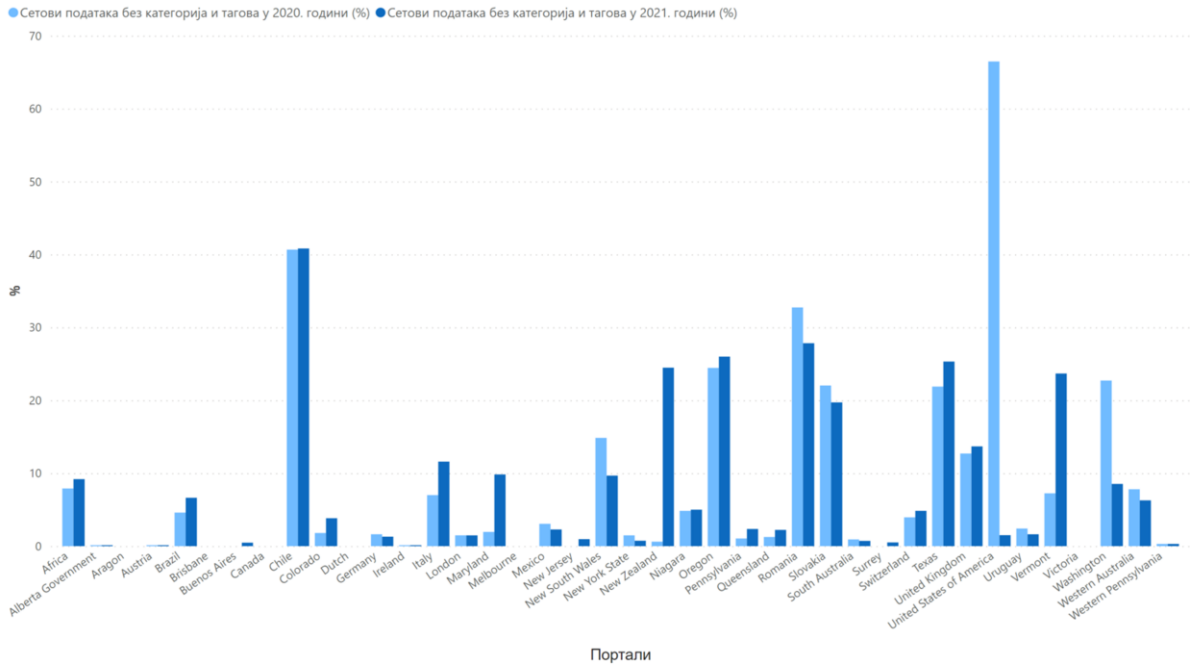


Слика 6 – Проценат података који нису описани таговима по порталима отворених података у 2020. и 2021. години

3.6. Промене у броју сетова података који нису описани таговима и нису додељени категоријама

Анализом портала 2020. године закључено је да се на већини портала јављају подаци који нису описани таговима а нису ни додељени категоријама. На само девет портала није било ових података. Од тих девет портала, 2021. године шест портала је задржало овај тренд, док су се на преостала 3 портала појавили подаци без категорија и тагова. Преглед процената података на порталима који нису описани таговима и немају додељене категорије у анализираном периоду приказан је на слици 7.

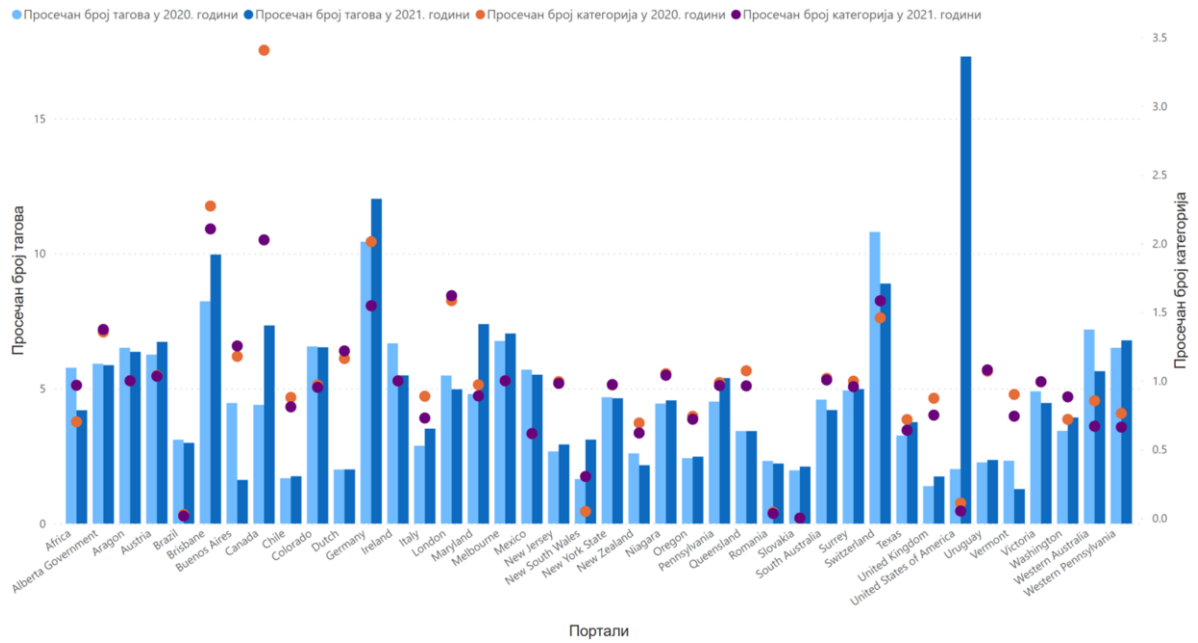
Може да се примети да је на неким порталима проценат ових података повећан у 2021.години. Један од ретких портала на којима је забележено велико смањење ових података је портал Сједињених Америчких Држава, док су остали портали задржали сличне проценте. С друге стране, просечан проценат скупова података без додељених категорија и дефинисаних тагова, смањен је са 8% у 2020. на 7,34% у 2021. години.



Слика 7 – Проценат података који нису описани таговима и немају дефинисану категорију по порталима отворених података у 2020. и 2021. години

3.7. Промене у броју категорија којима један сет података припада

У зависности од портала отворених података један сет података може да припада једној или више категорија. У оквиру ове анализе на 14 портала отворених података сетови података припадају само једној категорији. Међу овим порталима налазе се сви корисници Socrata платформе и четири портала која користе CKAN платформу. На преосталим порталима, примећено је да се сетовима података додељује и више категорија, у неким ситуацијама чак и све категорије које су доступне на порталу. Међутим, овакви случајеви су врло ретки. На слици 8 приказан је просечан број категорија које су додељене једном сету података у 2020. и 2021. години по порталима отворених података.



Слика 8 – Преглед просечног броја категорија које су додељене сетовима података и просечан број тагова којима су подаци описани на порталима отворених података у 2020. и 2021. години

3.8. Промене у броју тагова који се употребљавају за описивање једног сета података на порталу

У оквиру овог дела анализе, примећено је да се уобичајено један податак описује са више тагова. На анализираним порталима има података који су описани изузетно великим бројем тагова, али такве ситуације нису честе. Из тог разлога, просечан број тагова који се користи за описивање података на анализираним порталима био је 4,55 тага у 2020. години. Приликом анализе ових портала 2021. године просек је био 4,99 тага, али је ово повећање у великој мери последица великих промена које су настале на порталу Сједињених Америчких Држава, док су на осталим порталима бројеви остали слични. На слици 8 приказан је просечан број тагова којима су описани сетови података у 2020. и 2021. години.

Из приказане анализе може да се примети да број сетова података на порталима има тренд повећања, као и да на порталима мета-кључеви који представљају категорије и тагове често немају дефинисане вредности што драстично може да утиче на квалитет претраге података на порталима. Свега је пар портала у овом периоду имало напредак у смислу смањења процената некатегорисаних и нетагованих података, док је већина портала задржала сличне бројеве.

4. АНАЛИЗА ЗАВИСНОСТИ ИЗМЕЂУ ВРЕДНОСТИ МЕТА-КЉУЧЕВА ТАГОВА И КАТЕГОРИЈА

У оквиру ове докторске дисертације урађена је анализа употребе тагова у оквиру категорија на различитим порталима отворених података. Циљ ове анализе је да се установи колико често се тагови јављају у више категорија и у колико категорија се обично јављају, колико често се исти тагови понављају на порталима, односно колико често се јављају јединствене вредности тагова. Поред тога, у оквиру ове докторске дисертације је и анализа вредности које се додељују таговима, односно дужине текстуалних вредности како би се препознало да ли тагове обично чине појединачне речи или скупови од неколико речи.

За анализу је коришћено 40 портала отворених података са два пресека стања на овим порталима, јула 2020. године и децембра 2021. године. Основне карактеристике ових портала приказане су у претходном поглављу ове дисертације. У оквиру ове анализе над подацима са ових портала посматран је:

- однос укупног броја различитих тагова у односу на укупан број тагова на порталима,
- колико су честа појављивања истих тагова на порталима отворених података,
- процентуално у колико се категорија појављују тагови у односу на укупан број различитих тагова,
- колико тагова има дужину 1, процентуално по порталима отворених података,
- колика је медијана дужине тагова по порталима отворених података,
- колика је просечна дужина тагова по порталима отворених података,
- колике су дужине најдужих тагова на порталима отворених података.

Обзиром да тагови представљају кључне речи које ближе описују податке, различити подаци који припадају једној или више категорија могу да буду описани истим таговима. Из тог разлога, исте вредности се често јављају више пута. У претходном поглављу ове докторске дисертације приказана је расподела укупног броја различитих тагова на порталима отворених података који су анализирани. Међутим, те вредности су много мање у односно на укупан број тагова који је употребљен на

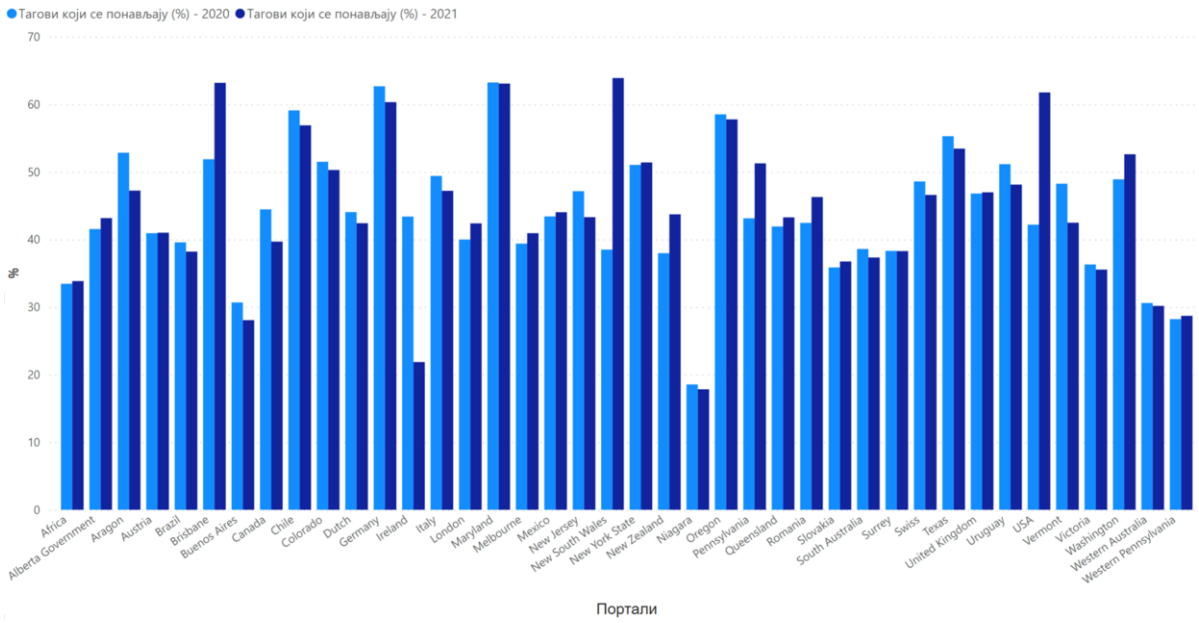
порталима за описивање података. Однос укупног броја тагова на свим анализираним порталима и укупног броја свих различитих тагова у 2020. и 2021. години приказан је у табели 2.

Табела 2 – Укупан број различитих и свих тагова у 2020. и 2021. години на анализираним порталима отворених података

Година	Укупан број различитих тагова	Укупан број тагова
2020	234 120	2 220 856
2021	527 080	8 028 406

Може да се примети да постоји велика разлика у укупним бројевима како различитих тако и свих тагова у 2020. и 2021. години. Неколико портала је забележило пораст у броју тагова у овом периоду, међутим портал Сједињених Америчких Држава је забележио значајне промене због којих се као последица и јавља ова велика разлика у укупним бројевима. На овом порталу 2020. године коришћено је 60.114 различитих тагова, док је тај број 2021. године био 304.160. Такође, може се приметити да је како у 2020. тако и у 2021. години, разлика између укупног броја употребљених тагова и укупног броја различитих тагова вишеструка. Овај резултат сугерише да се одређени број тагова понавља на порталима отворених података.

Из тог разлога, урађен је преглед на слици 9 у оквиру кога се може видети који проценат тагова се јавља барем два пута у описима података на сваком од



Слика 9 – Преглед процената тагова који се понављају на порталима отворених података у 2020. и 2021. години

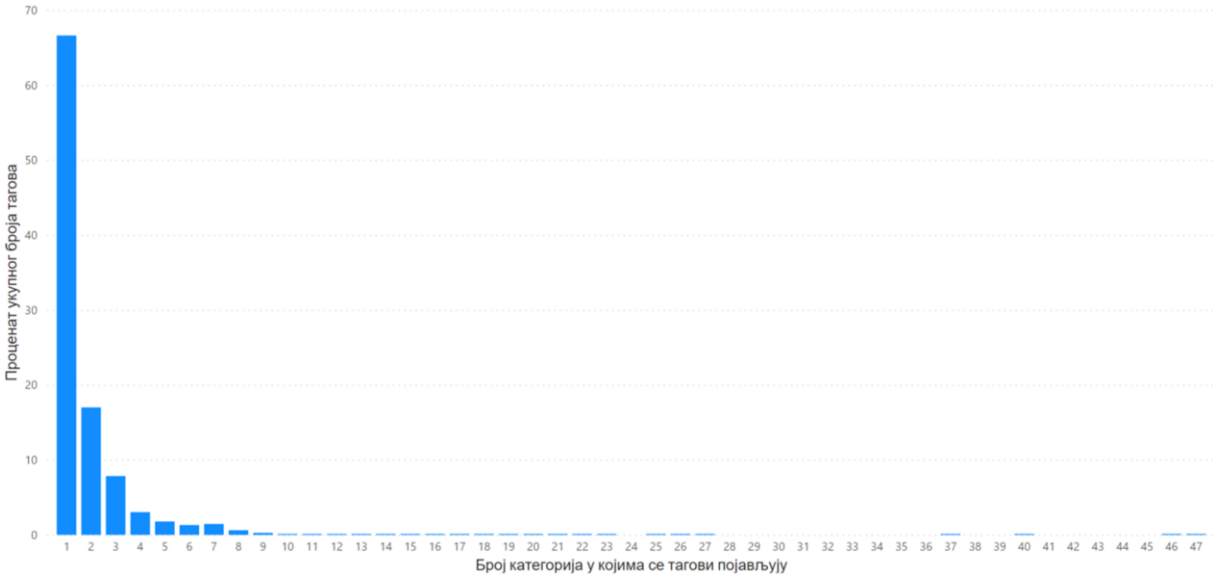
анализираних портала у 2020. и 2021. години. Из прегледа може да се закључи да се у просеку у 2020. години 43,99% различитих тагова понавља на анализираним порталима отворених података. Овај проценат је био мало већи у 2021. години и износио је 44,53%. Најмањи проценат понављања различитих тагова у обе анализиране године је на порталу Нијагаре и износи 18,54% и 17,83%. Портал Ирске је једини забележио значајно смањење овог процента у 2021. години у односу на претходну. На овом порталу је 43,39% свих различитих тагова употребљено барем два пута, док је 2021. године овај проценат износио 21,86%. Са друге стране, највећа повећања забележена су на порталима Сједињених Америчких Држава и Новог Јужног Веслa, на којима је разлика између процената свих различитих тагова употребљених барем два пута у 2020. и 2021. години износила 19,57% и 25,39%. На нивоу обе анализиране године, највећи проценат свих различитих тагова употребљених барем два пута, износио је 63,89% на порталу Новог Јужног Веслa у 2021. години.

Корисници уносе вредности тагова, па из тог разлога један таг може да буде опис већег броја података, који могу да припадају истој или различитим категоријама, као и да припадају већем броју категорија. Последично, тагови могу да се јаве у већем броју категорија, те је из тог разлога урађен сумарни преглед за свих 40 анализираних портала отворених података, приказан на слици 10, у оквиру којег може да се види процентуална расподела тагова у односу на број категорија у којима се јављају. Анализа је урађена над подацима из 2021. године применом следећих корака:

- за сваки портал је за све различите тагове на том порталу израчунато у колико се категорија појављују,
- након тога, урађено је сумирање резултата свих портала по броју категорија у којима се тагови јављају,
- проценти добијених резултата у односу на укупну суму бројева различитих тагова са свих портала приказан је на слици 10.

Из датог прегледа може да се закључи да се чак 66,6% тагова на порталима појављује само у оквиру једне категорије, у две категорије се појављује 17% тагова. Са повећањем броја категорија, смањује се проценат тагова, па се тако у три категорије појављује 7,8% тагова, а у четири 3%. У више од седам категорија се појављује укупно мање од 1% тагова, док када се посматрају тагови који се јављају у 16 и више категорија може да се закључи да се ради о појединачним случајевима.

Вредности тагова зависе искључиво од особе која их уноси и њихова дужина није унапред одређена, и може се разликовати. Из тог разлога, у оквиру ове докторске дисертације урађена је анализа дужина тагова, односно броја речи које се појављују у оквиру тагова.



Слика 10 – Преглед расподеле процената укупног броја тагова на свим порталима по броју категорија у којима се појављују у 2021. години

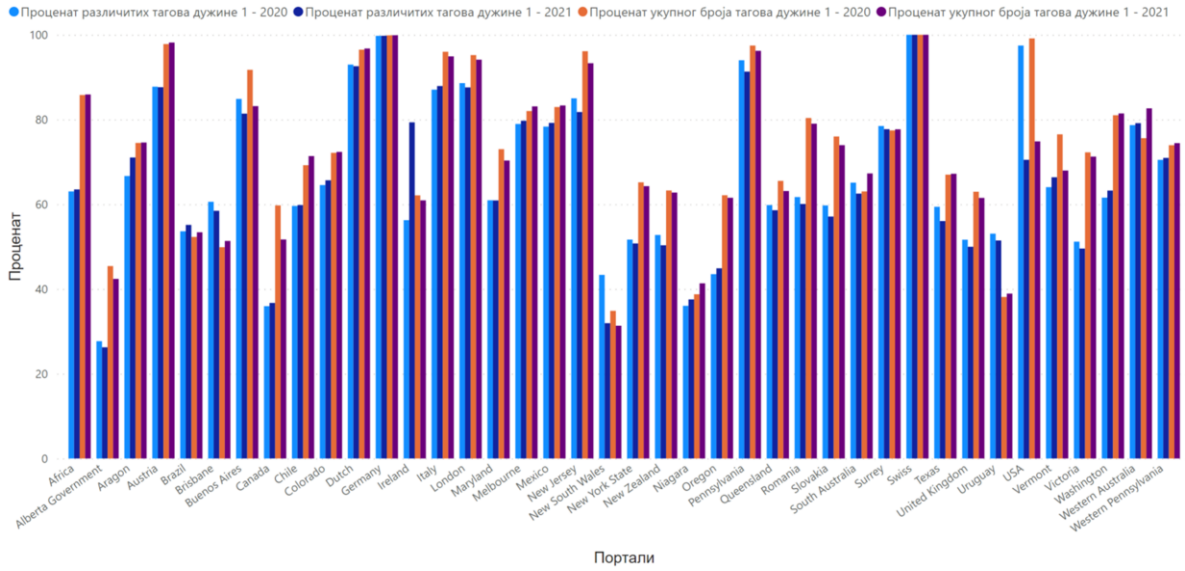
4.1. Проценат тагова дужине 1

Проценат тагова дужине 1 варира од портала до портала. На слици 11 приказан је проценат тагова дужине један у односу на укупан број различитих тагова као и укупан број свих тагова по анализираним порталима у 2020. и 2021. години. При креирању приказаног прегледа, подразумевано је да су речи у оквиру тагова раздвојене размацима.

Међутим, уочено је да су на неким порталима речи у оквиру једног тага раздвојене средњим цртама („-“), уместо размацима. Ово је посебно изражено на швајцарском порталу отворених података, на коме се користе искључиво средње црте за раздвајање речи у оквиру једног тага, као и немачком на коме се, осим у неколико изузетака, такође користе само средње црте.

На пример, на швајцарском порталу отворених података, у најупотребљаванијих 10 тагова у 2020. години, јављају се тагови као што су *complete-enumeration-survey, sample-surveys, rilevazione-totale, releve-exhaustif*, итд. Поред ових тагова, чести су и тагови попут *work-and-income, summary-statistics, statistica-di-elezioni-e-votazioni, statistics-of-*

elections-and-vote, education-and-science, crime-and-criminal-justice, economic-and-social-situation-of-the-population, statistical-basis-and-overviews итд.



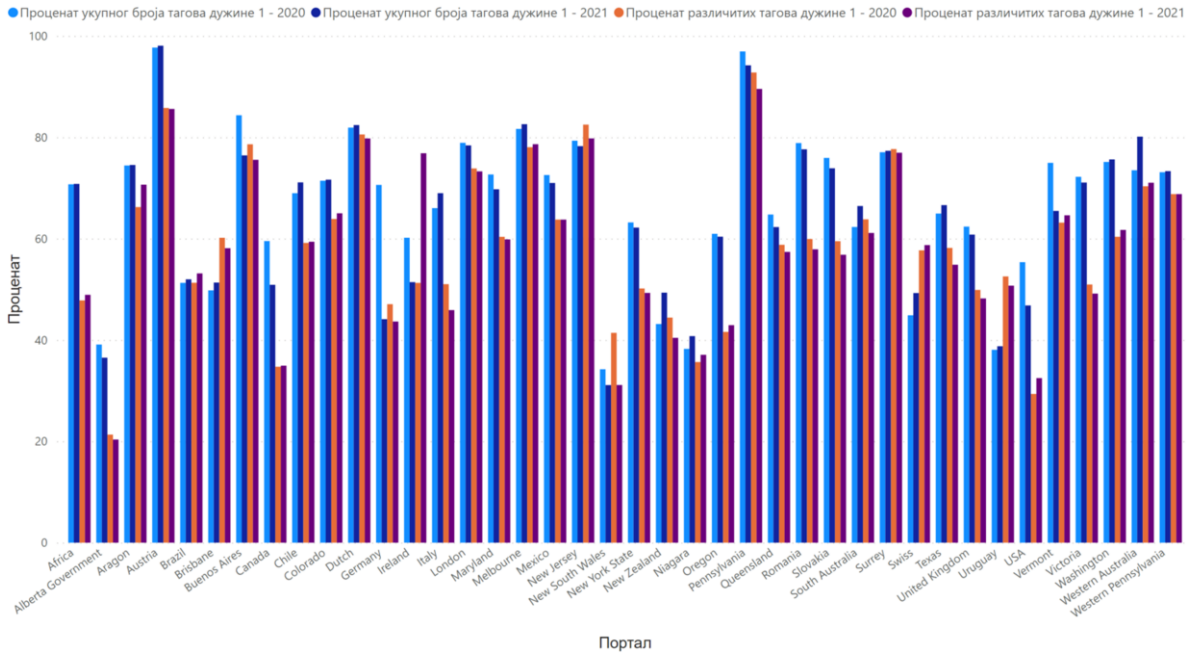
Слика 11 – Преглед односа броја тагова дужине 1 у односу на укупан број тагова за све различите тагове и за све тагове у 2020. и 2021. години по порталима отворених података

На немачком порталу отворених података у 2020. години, често су били коришћени тагови попут *menschen-mit-behinderung, umwelt-und-klima, offene-daten-bonn, open-data-düsseldorf, baugenehmigungen-und-bautätigkeit, verarbeitendes-gewerbe, verarbeitendes-gewerbe-sowie-bergbau-und-gewinnung-von-steinen-und-erden-in-schleswig-holstein* итд.

Слична ситуација је била и на порталу Сједињених Америчких Држава, на коме су у 2020. години међу најчешће употребљаваним таговима, вредности попут *national-geospatial-data-asset, earth-science, goddard-space-flight-center, glenn-research-center, spectral-engineering, drinking-water, ames-research-center, atmospheric-temperature, ocean-temperature, national-park-service* итд.

Да се овакви тагови не би водили као тагови са дужином 1, урађена је додатна анализа којом се осим размака за одређивање броја речи у тагу, користи и средња црта („-“) за одвајање речи. У оквиру ове анализе, размак је коришћен као примарни карактер за одвајање речи, док је средња црта коришћена за раздвајање речи у случају да таг након раздвајања речи размаком има дужину један. Оваквим механизмом раздвајања речи избегава се комбиновање карактера који се користе за раздвајање. Одлука да се бројање речи у тагу ради на овај начин донета је због претпоставке да уколико се у тагу јављају размаци, свака употреба средње црте у том тагу је намерна и није употребљена са циљем да раздвоји две речи.

Анализа на овај начин је урађена за све портале над подацима из 2020. и 2021. године. Резултати ове анализе су приказани на слици 12. Преглед резултата, као и на слици 11, дат је у процентима како за укупан број тагова, тако и на нивоу укупног броја различитих тагова. На овај начин, се може видети колико су чести тагови дужине 1, али и колико се често они понављају у односу на укупан број тагова.



Слика 12 – Преглед односа броја тагова дужине 1 у односу на укупан број тагова, за све различите тагове и за све тагове у 2020. и 2021. години по порталима отворених података, при комбиновању карактера за раздвајање речи

Упоређивањем добијених резултата на сликама 11 и 12, може се приметити да су највеће разлике у резултатима забележене на порталима отворених података Швајцарске, Немачке, Италије и Сједињених Америчких Држава. Поред ових портала још портали попут афричког, холандског, лондонског и мексичког имају разлику од преко 10% у резултатима на нивоу оба параметра за обе анализиране године. На осталим порталима, разлике у резултатима су знатно мање, тако да просечна разлика у резултатима са слика 11 и 12 износи 7,29% и 7,40% за разлику у процентима за укупан број тагова у 2020. и 2021. години респективно. На нивоу укупног броја различитих тагова, процентуално се резултати на сликама 11 и 12 разликују за 8,03% у 2020. години и 7,51% у 2021. години.

Из приказаних резултата са слике 12, може се приметити да је најмањи проценат тагова дужине 1 на нивоу свих различитих тагова у 2020. године био 21,32%, док је у

2021. години тај проценат износио 20,35%, а да су на нивоу укупног броја тагова ти проценти били 34,22% и 31,09% у 2020. и 2021. години респективно.

Такође може да се примети, да је у 2020. години четвртина анализираних портала имала проценат тагова дужине 1 на нивоу свих различитих тагова испод 50%, док је у 2021. години било 13 таквих портала. Са друге стране, у 2020. години, 21 портал отворених података је имао проценат тагова дужине 1 на нивоу свих различитих тагова између 50% и 70%, 7 између 70% и 85%, док је на 2 портала тај проценат износио више од 85%. У 2021. години 16 портала је имало проценат тагова дужине 1 на нивоу свих различитих тагова између 50% и 70%, 9 између 70% и 85%, док је на 2 портала тај проценат износио више од 85%.

У просеку на нивоу свих анализираних портала отворених података, проценат тагова дужине 1 на нивоу различитих тагова у 2020. Години, износио је 58,6%, док је у 2021. години износио 58,34%.

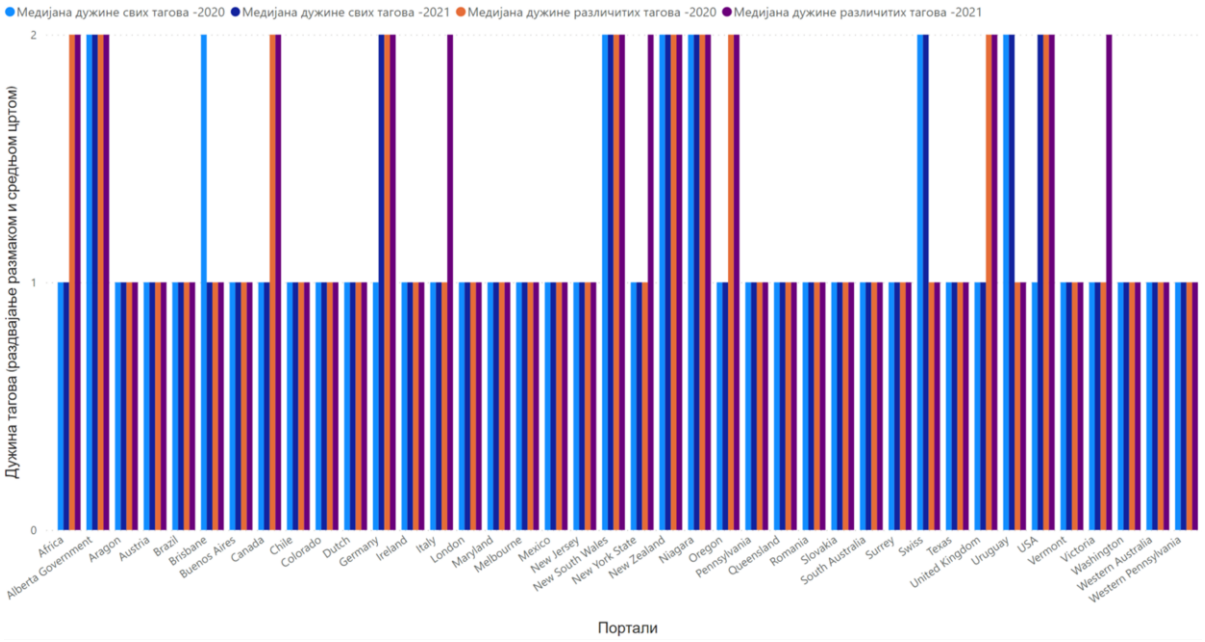
Са друге стране, на нивоу свих анализираних портала отворених података, проценат тагова дужине 1 на нивоу укупног броја тагова по порталу у 2020. години просечно је износио 66,52%, док је у 2021. години просечно износио 65,08%, што је више у односу на резултате добијене у односу на различите тагове по порталу. Може да се примети да је у 2020. години на само 9 портала забележен мањи проценат заступљености тагова дужине 1, на нивоу свих тагова на порталу у односу на број само различитих тагова на порталу, док је у 2021. години било 7 таквих портала. На свим осталим порталима забележени су већи проценти на нивоу свих тагова на порталу у односу на број само различитих тагова на порталу.

У 2020. години, 7 анализираних портала имало је проценат тагова дужине 1 на нивоу свих тагова испод 50%, 12 између 50% и 70%, 19 између 70% и 85%, док је на 2 портала тај проценат износио више од 85%. Слични резултати добијени су и за 2021. годину, када је 8 портала имало мање од 50% тагова дужине 1 на нивоу свих тагова, код 13 портала је тај проценат износио између 50% и 70%, код 17 између 70% и 85%, док је на 2 портала тај проценат износио више од 85%.

4.2. Медијана дужина тагова

Имајући у виду да су на неким порталима речи у таговима раздвојене средњом цртом, приликом одређивања медијане дужине тагова, односно медијане броја речи у таговима, за одређивање броја речи у једном тагу, коришћен је примарно размак, а

затим и средња црта, у случају да је број речи у тагу након раздвајања размаком био једнак јединици. Анализа је урађена над свим порталима за 2020. и 2021. годину и посматрана је медијана над две врсте скупа тагова за сваки портал: скуп само различитих тагова на порталу и скуп свих тагова на порталу. Резултати ове анализе су приказани на слици 13. Из приказаних резултата се може приметити да медијана на свим порталима, над оба анализирана скупа узима вредност из скупа {1, 2}. Додатно, може се приметити да 24 портала има искључиво вредност 1 за обе анализиране године, над оба скупа тагова, док само 4 портала узима вредност 2 за све анализиране вредности.



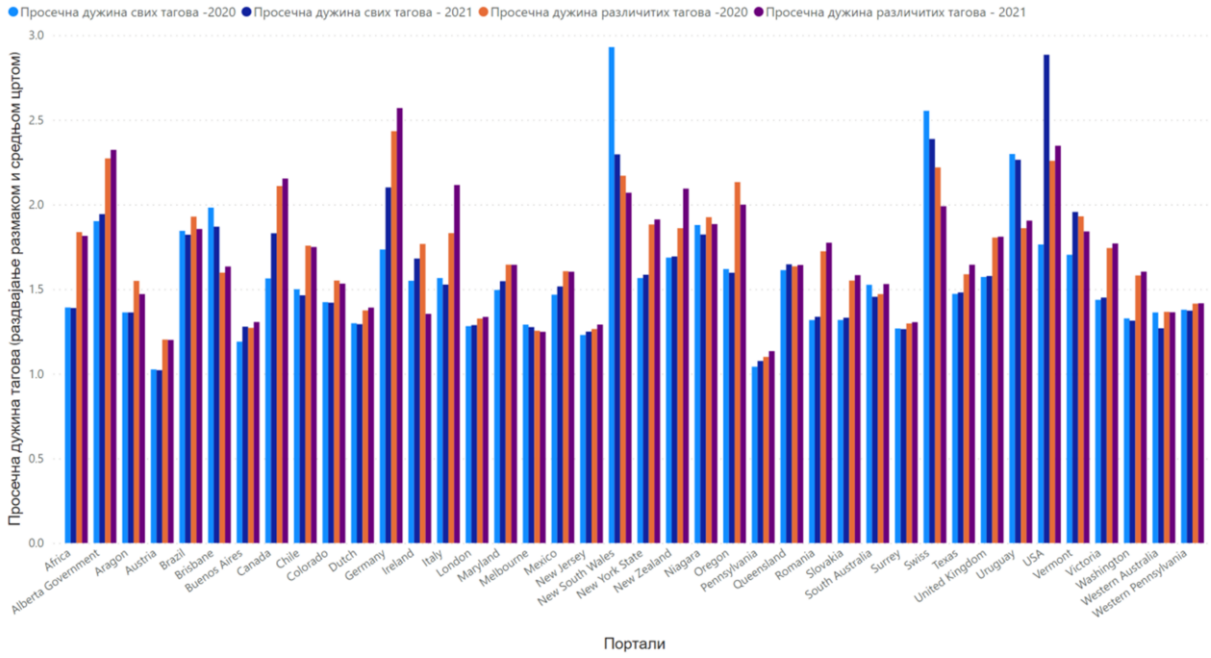
Слика 13 – Преглед медијане дужине тагова над скупом свих тагова и скупом различитих тагова у 2020. и 2021. години по порталима отворених података

4.3. Просечна дужина тагова

Поред анализе медијане дужине тагова, урађена је и анализа просечне дужине тагова, односно просечног броја речи које се користе у таговима. За потребе ове анализе је, као и при анализи медијане, коришћен примарно размак за раздвајање речи у оквиру једног тага, а затим и средња црта у случају да је број речи у тагу једнак јединици након раздвајања речи размаком. Анализа је урађена над две врсте скупа тагова за сваки портал: скуп свих тагова и скуп само различитих тагова, за 2020. и 2021. годину. Резултати анализе приказани су на слици 14.

Из приказаних резултата се може закључити да се просечан број речи у таговима по порталима отворених података креће у опсегу вредности од 1 до 3. Очекивано,

најмање просечне вредности имају портали Аустрије и Пенсилваније, као портали са изузетно високим процентом тагова дужине 1 над оба скупа података. Са друге стране, портали Немачке, Новог Јужног Велса, Швајцарске и Сједињених Америчких Држава једини имају по једну анализирану вредност која је већа од 2,5. Сви остали портали, за све анализиране вредности, имају просеке мање од 2,5. Просеци на нивоу свих портала су врло слични за 2020. и 2021. годину, и износе 1,56 над скупом свих тагова и 1,7 над скупом различитих тагова у 2020. години, док су резултати за 2021. годину 1,58 над скупом свих тагова и 1,7 над скупом различитих тагова.



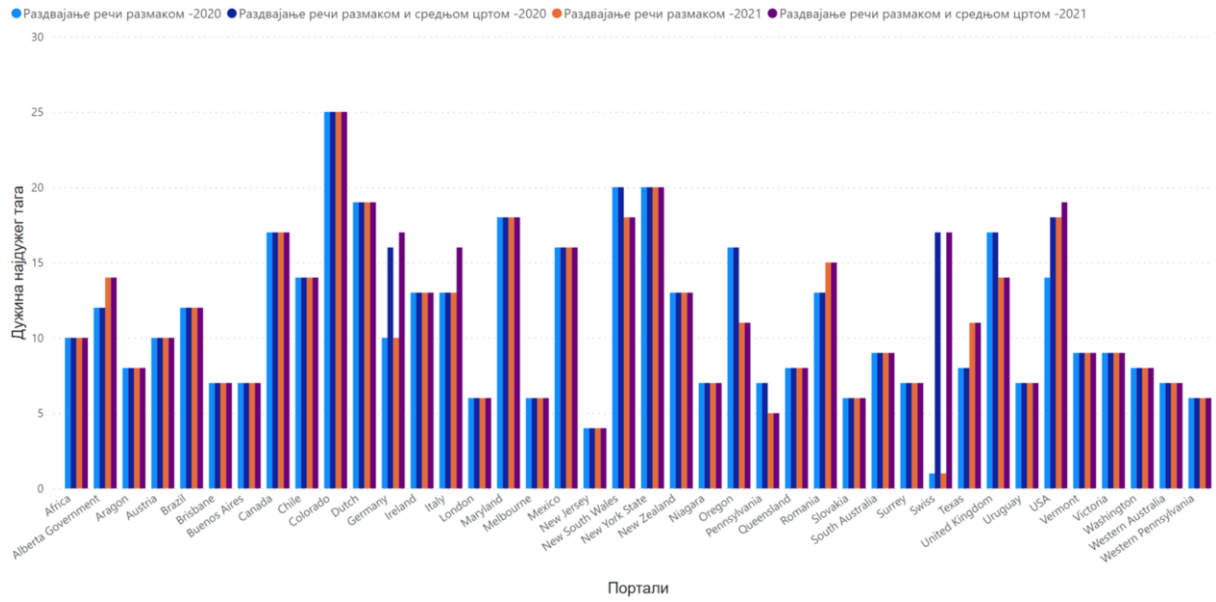
Слика 14 – Преглед просечне дужине свих тагова над скупом свих тагова и скупом различитих тагова у 2020. и 2021. години по порталима отворених података

4.4. Дужине најдужих тагова

У оквиру ове докторске дисертације урађена је и анализа колико речи садрже најдужи тагови на порталима отворених података у 2020. и 2021. години, по порталима отворених података. Анализа је рађена за два начина одређивања броја речи у таговима:

1) коришћењем искључиво размака за раздвајање речи у оквиру једног тага,
2) коришћењем примарно размака за раздвајање речи у оквиру једног тага, а затим и средње црте уколико таг након раздвајања речи размаком има дужину један.

Резултати анализе приказани су на слици 15.



Слика 15 – Преглед дужина најдужих тагова у 2020. и 2021. години по порталима отворених података

Из приказаних резултата се може приметити да се најдужи таг јавља на порталу отворених података Колорада, и износи 25 речи. Затим следе портали отворених података Државе Њујорк и Новог Јужног Велса, са таговима дужине 20 речи, а након тога портали Холандије и Сједињених Америчких Држава са 19 речи у оквиру барем једне од посматраних вредности.

Може се приметити да на порталима нема великих одступања између дужина најдужих тагова у 2020. и 2021. години за исти начин поделе тагова на речи. Када се посматрају резултати анализе у којој су речи раздвајане само размацима, на 32 портала најдужи тагови имају исти број речи у 2020. и 2021. години. Слични резултати су добијени и при раздвајању речи размацима и средњим цртама. Код ове анализе на 10 портала се разликују бројеви речи у најдужим таговима у 2020. и 2021. години.

Такође, упоређивањем резултата за исте године а различите начине раздвајања речи, може се приметити да се разлике у 2020. години јављају на 3 портала док се у 2021. години јављају на 4 портала. Највећа разлика је на швајцарском порталу отворених података, из разлога што овај портал користи искључиво средње црте за раздвајање речи у једном тагу. Последично, на овом порталу раздвајањем речи у тагу коришћењем искључиво размака вратиће увек дужину тага 1. Поред овог портала, остале разлике су такође у сетовима података који имају већу заступљеност коришћења средњих црта у оквиру тагова.

5. КАТЕГОРИЗАЦИЈА ПОДАТАКА НА ПОРТАЛИМА ОТВОРЕНИХ ПОДАТАКА

Аутори рада [26] су својом анализом из 2018. године приметили недостатак информација о категоријама у подацима доступним на порталима отворених података. Такође, и анализа приказана у трећем поглављу ове докторске дисертације, која прати стања на порталима у 2020. и 2021. години, указује да на већини портала отворених података постоји велики број података којима није додељена категорија. Додатно, из приказане анализе може да се примети да је проценат некатегорисаних података на већини портала повећан и да не постоји тренд додељивања категорије подацима који нису категорисани.

Имајући у виду да се портали отворених података користе за претрагу доступних података, као и преузимање, анализу и даљу употребну доступних докумената, недостатак информација о категоријама којима подаци припадају значајно утичу на ефикасност проналажења жељених сетова података. Из тог разлога, како би се унапредила доступност података, значајно је решавање овог проблема, и додељивање категорија некатегорисаним подацима. Услед велике количине објављених података на порталима, овом проблему може да се приступи развијањем методе за одређивање адекватних категорија којима некатегорисани подаци треба да припадају.

У овој докторској дисертација адресиран је овај проблем. Пошто се на порталима отворених података објављују различити подаци, у различитим форматима, у овој докторској дисертацији приступ за категоризацију података се базира на метаподацима који описују саме податке, односно на таговима којима су подаци описани. Додатно, приликом одабира категорија користи се хијерархија тагова која осликава начин употребе вредности тагова у оквиру појединачних категорија. Предлог методологије за категоризацију података на порталима отворених података приказан је у наставку.

5.1. Предлог методологије за категоризацију података

Методологија за категоризацију података дефинише кораке које је неопходно извршити како би се креирао поступак за категоризацију података на порталима отворених података, коришћењем вредности тагова који описују податке. Предложена методологија треба да одговори на следећа питања:

- Како преузети потребне информације са портала отворених података?
- Како креирати и анализирати хијерархијску структуру тагова?
- Како рачунати сличности између комбинација тагова?
- Како израчунате сличности употребити за категоризацију података?

Из тог разлога, предлог методологије за категоризацију података се састоји из следећих корака:

- Дефинисање начина за прикупљање података,
- Одређивање хијерархије тагова,
- Преглед и анализа добијених хијерархијских структура,
- Дефинисање поступка категоризације.

5.1.1. Дефинисање начина за прикупљање података

У оквиру ове дисертације, категоризација података заснива се на постојећим подацима на порталима отворених података и њиховим метаподацима. Стога, потребно је дефинисати начин преузимања информација о метаподацима свих података који су доступни на одређеном порталу отворених података. Као што је напоменуто у претходним поглављима ове дисертације, већина портала отворених података користи неку од постојећих платформи, попут CKAN и Socrata платформе, које имају своје програмске интерфејсе (АПИ). Ови програмски интерфејси обично омогућавају приступ различитим елементима доступних података као и лако преузимање жељених информација.

Из тог разлога, у овом кораку потребно је упознавање са програмским интерфејсом који користи портал отворених података за који се планира категоризација података, упознавање са специфичностима које портал има, препознавање потребних мета-кључева и дефинисање начина за приступ жељеним метаподацима.

На пример, за приступање метаподацима неког податка на основу јединственог идентификатора коришћењем програмског интерфејса CKAN платформе [27] може да се искористи следећи позив:

```
{URL}/api/3/action/package_show?id={DATASET_ID},
```

при чему URL представља URL адресу портала отворених података а DATASET_ID јединствени идентификатор тог податка. Резултат овог захтева садржаће све метаподатке траженог сета података у JSON формату. Један пример свих елемената

метаподатака који могу да се преузму на овај начин приказан је у Додатку Б ове докторске дисертације.

Након упознавања и анализе структуре метаподатака, неопходно је преузимање метаподатака свих података доступних на порталу отворених података и памћење ради даље употребе.

5.1.2. Одређивање хијерархије тагова

У оквиру овог корака, извршава се креирање структуре хијерархије која осликава употребу тагова приликом означавања докумената различитих категорија. Како би могло на нивоу једне категорије доступне на порталу да се анализира употреба тагова, потребно је извршити раздвајање свих података у групе, по категоријама којима припадају. Након тога, потребно је извршити и креирање хијерархије која осликава употребу вредности тагова у оквиру једне категорије и комбинација у којима се тагови појављују.

Детаљно објашњење овог корака методологије је приказано у поглављу *„Креирање хијерархије тагова“* у оквиру којег је преглед методе која се користи за креирање хијерархије као и начин примене методе над преузетим подацима отворених података.

5.1.3. Преглед и анализа добијених хијерархијских структура

Број тагова који се користи на порталима отворених података је велики. Такође, број комбинација у којима се неки тагови појављују, додатно повећава сложеност хијерархијске структуре употребе тагова. Из тог разлога, потребно је омогућити ефикасан начин за приказ и анализу хијерархије, како би могле да се препознају везе које се појављују у хијерархији и препознају елементи који су битни за категоризацију података. Алат који омогућава детаљну визуалну анализу и приказ хијерархија тагова је приказан у оквиру поглавља *„Алат за визуализацију и анализу хијерархијске организације тагова на порталима отворених података“*.

5.1.4. Дефинисање поступка категоризације

Дефинисање поступка категоризације подразумева давање одговора на последња два питања дефинисана у оквиру методологије.

Из тог разлога, у овом кораку потребно је дефинисати метод за рачунање сличности између два термина која се употребљавају за описивање података на

порталима отворених података. Овај метод треба да обезбеди начин препознавања колико су тагови у хијерархији слични са тагом податка који је потребно категорисати.

Поред тога, потребно је дефинисати параметре који ће бити искоришћени за категоризацију података, као и начин њиховог израчунавања на основу резултата сличности између тагова.

Детаљан поступак овог корака методологије је приказан у поглављу *„Поступак категоризације података“*.

6. КРЕИРАЊЕ ХИЈЕРАРХИЈЕ ТАГОВА

Из анализе приказане у четвртом поглављу ове докторске дисертације може се приметити да се подаци на порталима отворених података описују великим бројем различитих комбинација речи и израза које могу бити различите дужине. Такође, тагови могу да се понављају, као и да се јаве у више категорија појединачно или са другим таговима. Додатно, један таг у неким категоријама може се појављивати чешће, у другим ређе, а може да буде категорија у којима се не јавља. Последично, посматрано по начину употребе тагова за означавање података по категоријама, неки тагови могу да буду значајнији за одређену категорију од других, уколико се јављају у више комбинација или самостално.

Препознавање тагова који се могу сматрати да имају већу тежину приликом означавања података у одређеној категорији, је значајно за процес категоризације података. Стога, потребно је посматрати начин употребе тагова за означавање података у различитим категоријама.

Из тог разлога, у оквиру ове докторске дисертације искоришћена је Анализа формалних концепата (ФЦА), како би се омогућило креирање хијерархија тагова у оквиру једне категорије. У наставку овог поглавља докторске дисертације, дат је преглед анализе формалних концепата. Након тога, приказан је начин употребе ове методе за креирање хијерархије тагова. На крају, представљен је алат за визуализацију добијених хијерархија, у циљу анализе структуре хијерархија, препознавања веза између тагова и начина употребе тагова у оквиру једне категорије.

6.1. Анализа формалних концепата

Анализа формалних концепата (*Formal Concept Analysis – ФЦА*) је математички метод са све већом популарношћу у различитим областима истраживања и развоја. Rudolf Wille је увео Анализу формалних концепата почетком 1980-их година као приступ који се заснива на формализацији концепата [28]. Овај метод своје корене вуче у радовима Birkhoff-a [29], Barbut-a и Monjardet-a [30] и других и базиран је на математичкој теорији уређења, односно теорији комплетних мрежа (*Lattice Theory*), чију основу формалног описа чине дефиниције формалног контекста и формалног концепта. У овим дефиницијама придев „формални“ треба да нагласи да се ради о

математичким појмовима који одражавају само неке аспекте значења речи контекст и концепт у стандардном језику [31]:

Дефиниција 1: Формални контекст је тројка $K \coloneqq (G, M, I)$ која се састоји од скупа објеката *G*, скупа атрибута *M* и бинарне релације $I \subseteq G \times M$ при чему $(g, m) \in I$ указује да „објекат g има атрибут m“.

Дефиниција 2: За подскуп објеката $A \subseteq G$, и за подскуп атрибута $B \subseteq M$ дефинише се:

$$A^I \coloneqq \{m \in M | \forall g \in A : (g, m) \in I\}$$
$$B^I \coloneqq \{g \in G | \forall m \in B : (g, m) \in I\}.$$

Уколико важи да је $A \subseteq G, B \subseteq M, A^I = B, B^I = A$, пар (A, B) се назива формални концепт код кога се скуп A назива *extent* концепта а скуп B *intent* концепта.

Неформално дефинисано, подскуп B је подскуп свих атрибута који су заједнички за све објекте подскупа A, док се подскуп A састоји од свих објеката који имају све атрибуте подскупа B.

Дефиниција 3: Скуп S(C) свих концепата формалног контекста C са парцијалним уређењем $(A_1, B_1) \leq (A_2, B_2) : \leftrightarrow A_1 \subseteq A_2 \ \ (B_1 \supseteq B_2)$ је комплетна мрежа контекста C (*concept lattice*).

Описано на мање формалан начин, Анализа формалних концепата се сматра методом анализе података која користи однос између одређеног скупа објеката и одређеног скупа атрибута. Уобичајени начин да се визуализује формални контекст је визуализација односа између ова два скупа коришћењем матрице. Сваки ред унутар матрице представља објекат, док свака колона представља један атрибут. Вредности поља матрице које обично узимају вредности из скупа {да, не}, означавају однос између одређеног објекта и одређеног атрибута, односно да ли објекат садржи или не садржи тај атрибут. Пример једне визуализације формалног контекста приказан је на слици 16.

	Latin America	Europe	Canada	Asia Pacific	Middle East	Africa	Mexico	Caribbean	United States
Air Canada	X	X	X	X	X		X	X	X
Air New Zealand		X		X					X
All Nippon Airways		X		X					X
Ansett Australia				X					
The Austrian Airlines Group		X	X	X	X	X			X
British Midland		X							
Lufthansa	X	X	X	X	X	X	X		X
Mexicana	X		X				X	X	X
Scandinavian Airlines	X	X		X		X			X
Singapore Airlines		X	X	X	X	X			X
Thai Airways International	X	X		X				X	X
United Airlines	X	X	X	X			X	X	X
VARIG	X	X		X		X	X		X

Слика 16 – Пример формалног контекста [32]

У овом примеру, авио компаније групације Star Alliance представљају скуп објеката, док скуп атрибута садржи одредишта до којих те компаније имају летове [32]. Матрицом је представљен бинарни однос ова два скупа и описује које дестинације опслужује који члан Star Alliance групације.

Најзначајнији излаз анализе формалних концепата представља мрежа концепата. Имплементирањем ФЦА методе изводе се концепти из матрице и креира се колекција формалних концепата који су логички организовани у хијерархију концепата. Ови концепти су међусобно повезани коришћењем релација подконцепт-суперконцепт и креирају мрежу концепата [32]. Дакле, мрежа концепата одражава генерализацију и специјализацију између формалних концепата унутар једног формалног контекста [33]. Добра страна употребе мреже концепата јесте могућност да се из мреже уради реконструкција оригиналних података, па се самим тим оваква мрежа може сматрати верном репрезентацијом оригиналних података.

Како наводи Belohlavek у раду [34], Анализа формалних концепата омогућава откривање и резоновање концепата у подацима, откривање и резоновање зависности у подацима и визуализацију података, концепата и њихових зависности. У последњих 20 година овај метод је нашао примену у различитим областима попут откривања знања, софтверског инжењерства и проналажења информација. У наставку поглавља дато је неколико примера употребе ове методе у различитим областима истраживања.

Истраживања су показала да се приступи засновани на ФЦА могу користити као одличне методе за класификацију. Аутори радова [35] и [36] дали су преглед метода за класификацију заснованих на анализи формалних концепата. Група аутора коју је

предводио Fu је у раду [37] урадила поређење алгоритама за класификацију заснованих на ФЦА и стандардних алгоритама за класификацију попут C4.5, Naïve Bayes и IB1. Системи попут GRAND [38], RULEARNER [39], GALOIS [40], NAVIGALA [41] и CITREC [42] користе мрежу концепата као простор тражења у коме је једноставан прелазак са једног на други ниво у мрежи, чиме се може анализирати структура концепта.

Како се ФЦА показао као добар метод за класификацију и организацију података, Chekol и Napoli су у раду [43] користили ФЦА како би изградили формални контекст од скупа одговора на SPARQL упите, и препознали и визуализовали скривене односе унутар њих. Alam, Buzmakov, Codocedo и Napoli у раду [44] користе ФЦА како би организовали RDF податке у мрежу концепата како би открили импликације и информација које недостају. Пример употребе ФЦА за налажење концепата који потенцијално недостају у биомедицинској терминологији је приказан у раду [45]. Аутори овог рада су представили лексички приступ заснован на ФЦА, који на основу лексичких карактеристика постојећих термина креира мрежу концепата на основу које открива нове концепте који потенцијално недостају у терминологији.

Аутори Boutari, Carpineto и Nicolussi у свом раду [46] користе ФЦА за развој приступа за проширење термина, како би превазишли проблем оскудне репрезентације и недостатак заједничког контекста међу кратким текстовима доступним на Вебу. Њихов приступ се заснива на коришћењу односа између појмова присутних у мрежи концепата која је повезана са корпусом документа. У раду Dufour-Lussier-a, Lieber-a, Nauer-a, и Toussaint-a [47] ФЦА се користи као механизам за адаптацију рецепата за кување. Они креирају мрежу концепата на основу рецепата који садрже састојак којим треба заменити постојећи састојак у рецепту. Користе је како би пронашли секвенцу кулинарских радњи која се најближе подудара са оном која се примењује на састојак који ће бити замењен у рецепту који треба адаптирати.

Аутори рада [48] користе анализу формалних концепата као алат за класификацију Веб сервиса на основу сличности њихових функционалности, са циљем да се омогући проналажење адекватног резервног сервиса у случају прекида рада сервиса у употреби, у композитним Веб апликацијама. На основу мере сличности између функционалности сервиса Carpineto, Michini и Nicolussi су у раду [49] приказали приступ за класификацију текста коришћењем методе потпорних вектора (*Support Vector Machine*) у комбинацији са Анализом формалних концепата, при чему мрежу концепата

употребљавају за одређивање веза између термина у документу. Пример употребе Анализе формалних концепата као класификационог алата за избор Веб сервиса приказан је у раду [50]. Аутори овог рада класификују постојеће сервисе и креирају стабло одлуке коришћењем ФЦА методе како би омогућили кориснику избор адекватног сервиса у зависности од спецификације. Креирано стабло, касније могу да употребе за избор еквивалентног сервиса уколико постојећи треба да се замени у току рада.

Употреба анализе формалних концепата на семантичком Вебу је разматрана у раду [51]. Аутори су у овом раду приказали приступ како се ФЦА алгоритми могу употребити за креирање формалних концепата из података семантичког Веба, и детаљно су анализирали утицај хетерогености оваквих података на скалабилност и перформансе ФЦА алгоритама за креирање концепата. Формални и полуаутоматски приступ за развој онтологија заснован на анализи формалних концепата, аутор Gaihua Fu приказао је у раду [52]. Циљ овог рада је да се организацијама обезбеди механизам за интеграцију података који показују имплицитне и двосмислене информације. Предложени механизам се у експериментима аутора на неколико индустријских сетова података показао ефикасним у организацији и спајању хетерогених података, и креирању знања које задовољава потребе пословања. Додатни примери употребе техника Анализе формалних концепата за креирање онтологија дати су у радовима [53] и [54]. Такође, ФЦА може да се користи и за мапирање онтологија. Пример ове употребе дат је у раду [55] у коме аутори израчунавају мере између ентитета различитих онтологија и изводе мапирање поткласа. Примери употребе Анализе формалних концепата за мапирање биомедицинских онтологија су приказани у радовима [56] и [57].

Бројни други примери употребе Анализе формалних концепата у различитим областима дати су у радовима [58] и [59]. Аутори Poelmans, Kuznetsov, Ignatov и Dedene су у ова два рада анализирали преко 1000 радова објављених у периоду од 2003. до 2011. године, који описују различите примене ФЦА. Они су искористили ФЦА како би представили и визуализовали главне истраживачке теме у ФЦА заједници. Они су у раду [58] дали опширан преглед примене ФЦА у онтологијама и откривању знања у различитим доменима, док су у раду [59] дали преглед метода заснованих на ФЦА за обраду знања и различита ФЦА проширења. Singh, Cherukuri и Gani су у раду [60] анализирали више од 350 радове објављених у периоду од 2011. до 2016. године на

тему Анализа формалних концепата. Њихово истраживање је потврдило да је Анализа формалних концепата нашла примену у онтологијама, откривању знања, резоновању као и да се проширује другим оквирима за представљање знања.

6.2. Примена Анализе формалних концепата за креирање хијерархије тагова

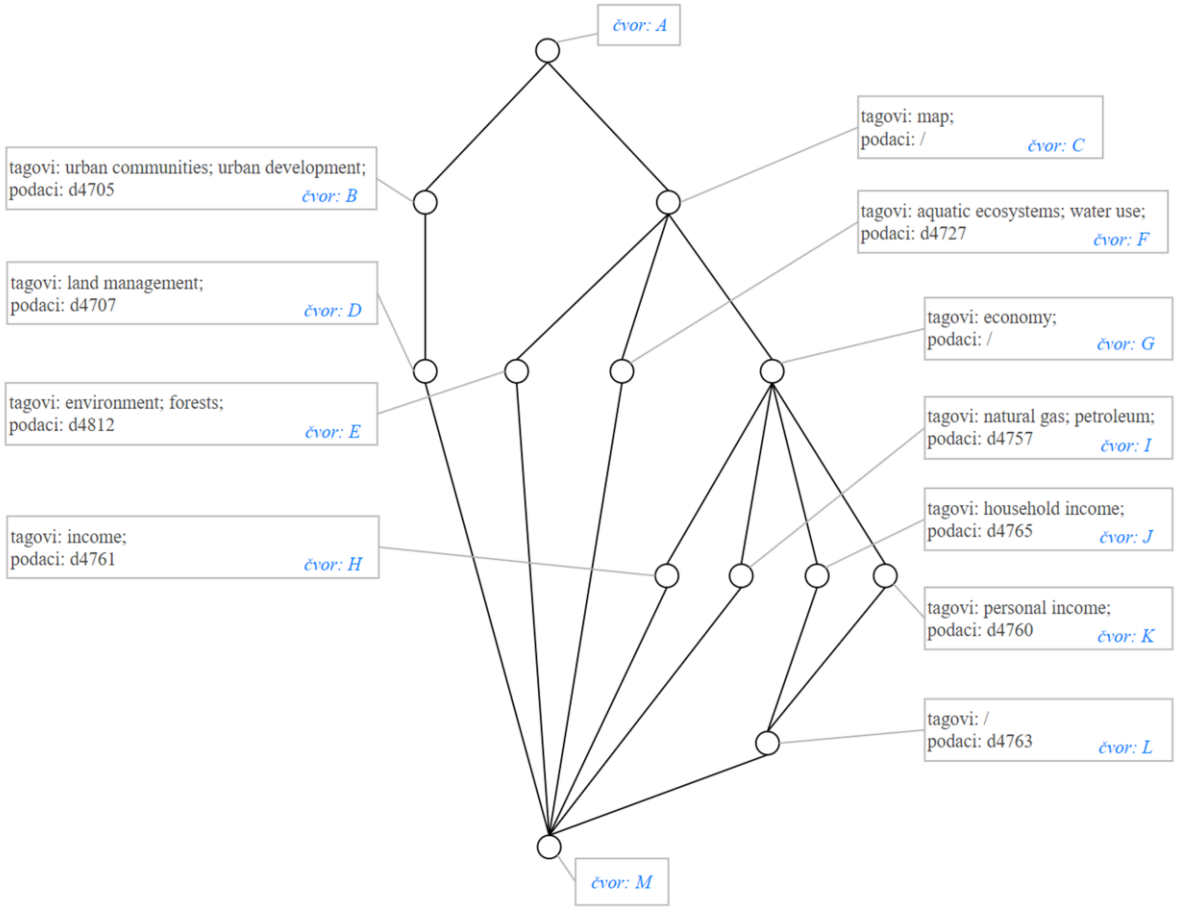
Анализа формалних концепата може да се употреби за креирање хијерархије тагова који се јављају у оквиру једне категорије на порталима отворених података на основу начина њихове употребе [26]. Ово се може постићи уколико се, на основу свих сетова података који припадају једној категорији креира формални контекст. За креирање једног оваквог формалног контекста користе се информације доступне у оквиру метаподатака сваког податка. Уколико се посматра једна категорија на неком порталу отворених података, сваки податак који припада тој категорији представљаће један објекат у формалном контексту, док ће сви тагови који се појављују у подацима те категорије представљати атрибуте формалног контекста. Пример једног формалног контекста за креирање хијерархије тагова на нивоу једне категорије на порталу отворених података приказан је у табели 3.

Табела 3 – Пример дела формалног контекста за категорију economics and industry са портала отворених података Канаде

Dataset	economy	map	natural gas	petroleum	personal income	income	environment	forests	aquatic ecosystems	water use	urban communities	urban development	household income	land management
d4757	x	x	x	x										
d4760	x	x			x									
d4761	x	x				x								
d4812		x					x	x						
d4727		x							x	x				
d4705											x	x		
d4763	x	x			x								x	
d4765	x	x											x	
d4707											x	x		x

Приказани пример креиран је на основу података доступних на порталу отворених података Канаде. Издвојено је девет сетова података (у наставку текста *податак*) који

припадају категорији *economics and industry* којима су додељене нове јединствене ознаке. У скуп атрибута формалног контекста, додати су само они тагови који се јављају у описима објеката формалног контекста. У приказаном примеру, свака ознака „x“ представља везу између податка и тага формалног контекста, и означава да одређени податак у својој листи тагова садржи одређени таг. На пример, податак *d4757* описан је таговима *economy*, *map*, *natural gas* и *petroleum*, док је податак d4707 описан таговима *urban communities*, *urban development* и *land management*.



Слика 17 – Мрежа концепата креирана на основу формалног контекста из табеле 3

На основу приказаног формалног контекста извршавањем Анализе формалних концепата се формира мрежа концепата приказана на слици 17. Сваки чвор у приказаној мрежи представља једну комбинацију тагова. За сваки податак у формалном контексту постоји чвор у мрежи који одговара комбинацији тагова која га описује. Међутим, не мора сваки чвор да представља комбинацију тагова неког података. Неки чворови могу да представљају комбинацију тагова која чини подскуп тагова који се јавља заједно у неким подацима, али не и комплетан скуп описа неког податка формалног контекста. Линије у мрежи представљају хијерархијске везе између чворова.

Веза између два чвора означава да је први чвор (чвор који је виши у мрежи) генералнији од другог чвора (чвор који се налази испод њега). Последично, то значи да доњи чвор садржи све тагове обухваћене горњим чвором и најмање још једну додатну вредност.

Генерално посматрано, чворови који се налазе при врху мреже концепата су генералнији, а самим тим и виши у хијерархији у односу на чворове при дну мреже. Сваки чвор у мрежи обухвата све вредности које се јављају у чворовима са којима је повезан ка врху хијерархије. Сходно томе, чвор на врху садржи ознаке које се појављују унутар свих скупова података за дату категорију и представља најгенералнији чвор. Кретањем низ мрежу концепата, сваки чвор има најмање једну додатну вредност и специфичнији је од чворова који се налазе изнад њега. Последично, последњи чвор у мрежи садржи све атрибуте који се јављају у формалном контексту.

На пример, у мрежи концепата са слике 17, чвор *A* је најгенералнији чвор и налази се на врху хијерархије. Обзиром да се у оквиру формалног контекста не појављује ниједан таг који је заједнички за све објекте, чвор *A* не садржи тагове. Испод њега се у хијерархији налазе чворови *B* и *C*, док је чвор *M* на дну хијерархије и обухвата све тагове који се јављају у формалном контексту. Чворови *C* и *G* представљају пример чворова који не садрже комплетан скуп тагова ниједног објекта формалног контекста.

Посматрањем приказане мреже концепата може да се закључи да је на пример, чвор *G* генералнији чвор од чворова *H, I, J* и *K* и да се налази изнад њих у хијерархији. Такође, овај чвор подржава све тагове који се налазе у чворовима који су хијерархијски изнад њега (чворови *C* и *A*), па је његова комбинација тагова скуп {map, economy}. Чворови *H, I, J* и *K* нису међусобно хијерархијски повезани пошто ниједан од њих не садржи све ознаке другог. Чвор *L* је пример чвора који самостално нема додатне тагове, али представља комбинацију чворова *J* и *K*, те његова комбинација тагова одговара скупу {map, economy, household income, personal income}.

Мрежа концепата креирана на основу употребе тагова у описивању података на порталима отворених података може да дâ увид у хијерархијску структуру тагова. Ова структура може да буде од великог значаја у процесу категоризације података на основу тагова којима су подаци описани. Оваква хијерархија омогућава да се препознају, како тагови, тако и комбинације тагова које су значајније у оквиру неке категорије, али и да се препознају тагови који су при дну хијерархије и мање су значајни у описивању података који припадају одређеној категорији.

Лоша страна овог приступа јесте величина формалног контекста када су портали отворених података у питању. Такође, величина и сложеност мреже концепата може драстично да се повећава повећањем броја атрибута у формалном контексту и разноликости комбинација атрибута која се појављује у објектима. Из тог разлога, алати за визуализацију опште намене нису адекватни за визуализацију толико великих мрежа концепата. Последично, да би се омогућила ефикасна анализа употребе тагова у оквиру категорија на порталима отворених података развијен је посебан алат са механизмима који омогућавају да се превазиђе проблем величине мреже концепата и омогући ефикасна визуална анализа.

6.3. Алат за визуализацију и анализу хијерархијске организације тагова на порталима отворених података

Истраживање и визуализација података су средства за извлачење знања и давање смисла подацима [61][62]. Њихов циљ је да подрже тумачење и манипулацију информацијама. Визуална анализа треба да омогући корисницима увид у везе које могу бити од значаја, корелације између објеката и правила која се јављају у подацима.

Већина традиционалних система за визуализацију достигла је своје границе када се визуализује веома велика количина података. У радовима [63], [64], [65], [66] и [67], аутори су говорили да је развој алата и система који могу успешно да раде са великим подацима постао изазов. У раду [68], аутори су нагласили да „величина“ великих скупова података и њихова сложеност у смислу хетерогености података, доприносе сложености репрезентације података, као и да је ефективно представљање свих информација у исто време веома изазовно. Стога је истраживање и визуализација великих скупова података постало велики истраживачки изазов. Један од главних захтева који треба да испуне савремени системи за визуализацију је визуална презентација и интеракција која ће обезбедити јасан преглед, лако истраживање и руковање великим бројем објеката, као и спречити преоптерећење приказа информацијама. Из тог разлога, прибегава се систему који функционише по принципу „прво преглед, зумирање и филтрирање, а затим детаљи на захтев“ [69].

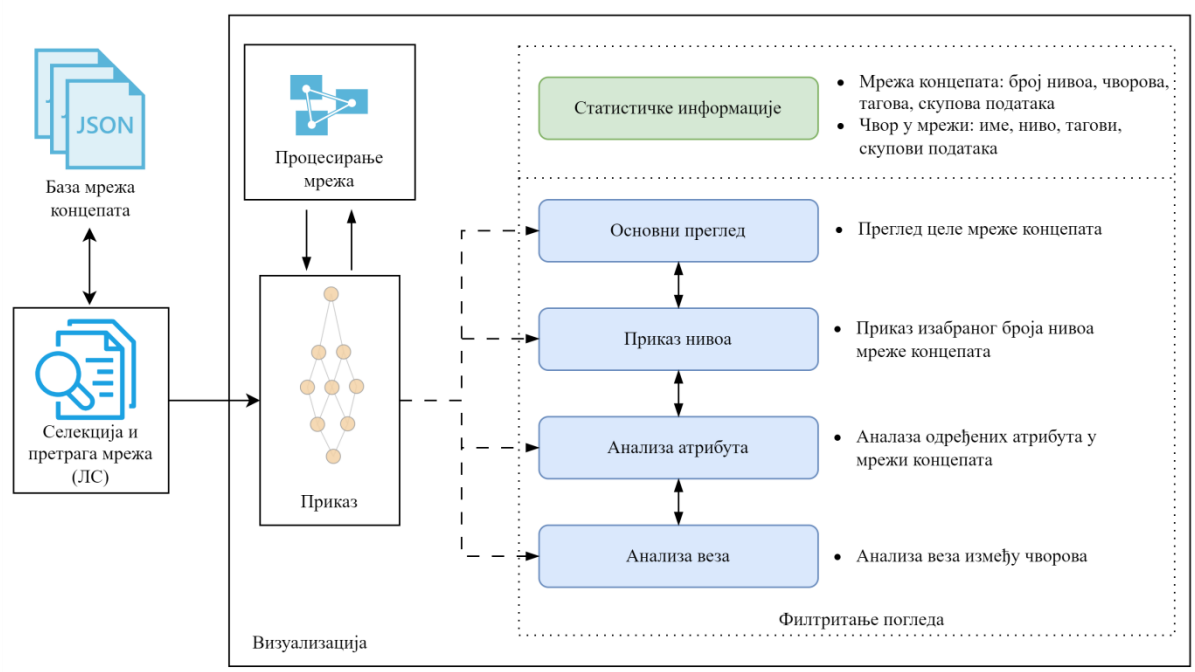
Последњих година уложен је значајан напор у решавање проблема визуалног истраживања и анализе [70], [71], [72]. Као резултат тога, развијен је велики број алата и система међу којима су и системи визуализације који се заснивају на графу. Већина система заснованих на графу имплементира приступ чворова и потега [73], где се

чворови користе за представљање ентитета, а потези (везе) постоје између чворова ако постоји релација између представљених ентитета. Визуализација заснована на графовима може бити веома драгоцена за анализу података из различитих области истраживања, јер открива структуру и односе између анализираних података и може се користити као база знања у процесима доношења одлука. Аутори рада [74] су извршили веома добар преглед и поређење система визуализације заснованих на графовима анализирајући њихове функционалности. У свом раду аутори су издвојили карактеристике попут претрага по кључним речима, филтрирање података, технике узорковања, технике агрегације и других, као карактеристике које су од значаја за употребљивост система за визуализацију засновану на графовима.

Визуализација мреже концепата је изузетно важна за анализу када се користи Анализа формалних концепата. Као што је објављено у раду [33], визуализација хијерархијске структуре мреже концепата је значајна за практичну примену овог алгоритма. Међутим, главни проблем за визуализацију може бити величина мрежа концепата. Мреже концепата са великим бројем формалних концепата могу постати веома сложене и непрактичне за анализу. Додатно, нема много алата који могу да превазиђу овај изазов и успешно визуализују велике мреже концепата. Из тог разлога, у оквиру ове докторске дисертације креиран је алат који примењује приступ заснован на графу са интерактивним механизмима како би се превазишао проблем визуализације мреже концепата. Циљ алата је да омогући јасан и информативан преглед хијерархије података који се може користити за процесе доношења одлука као што је категоризација података.

Алат који је приказан у овој докторској дисертацији је алат опште намене који може да се употребљава за визуализацију и анализу података са структуром чворова и потега, али је његова главна намена да подржи визуализацију мрежа концепата [75]. Алат је реализован као Веб апликација која користи d3.js библиотеку [76] за визуални приказ података. Визуализација мреже у овом алату је заснована на графу, па се последично чворови користе за представљање одређених комбинација атрибута, а везе између чворова представљају хијерархијске односе између чворова. Додатно, принцип гравитације доступан у d3.js библиотеци се користи за обезбеђивање адекватног позиционирања чворова и креирања изгледа хијерархије у којој су чворови који су виши у хијерархији приказани изнад чворова који су испод њих у хијерархији.

Како би проблем комплексности мреже концепата био превазиђен, алат за визуализацију мора бити интерактиван, као и да нуди могућност прегледа, како целе хијерархије, тако и детаљних односа између атрибута у мрежи. Стога, алат за визуализацију који је представљен у овој докторској дисертацији подржава неколико кључних функционалности које пружају јасан преглед и анализу мреже концепата. На слици 18, приказане су структура и функционалности алата за визуализацију.



Слика 18 – Алат за визуализацију

При покретању апликације алат препознаје све мреже концепата које су доступне у библиотеци апликације (База мрежа концепата на слици 18). Мреже концепата се памте у JSON формату, при чему се за сваки чвор чува информација о идентификатору чвора, таговима који су обухваћени тим чвором, сетовима података који су описани том комбинацијом и нивоу тог чвора у мрежи концепата. Корисник на располагању има две могућности на почетку рада (ЛС компонената на слици 18):

- да одабере мрежу концепата за приказ и анализу,
- да сузи избор мрежа које би биле од интереса за корисника, претрагом над свим доступним мрежама.

Претрага доступна у овом кораку везана је за атрибуте формалног контекста, у овом случају за тагове којима се подаци описују. Корисник уноси листу тагова на основу којих жели да анализира мреже и алат за сваку од мрежа наводи који се од наведених тагова појављује у мрежи а који не. На овај начин, корисник може лакше да

препозна које од понуђених мрежа су потенцијално од интереса за одређену анализу, посебно ако се ради о анализи из неког домена.

Након избора мреже, кориснику се приказује интерактивни визуални приказ одабране мреже који подржава следеће опције:

- приказ комплетне мреже концепата
- приказ одређеног броја нивоа у мрежи концепата
- преглед позиције одређеног атрибута у мрежи концепата
- претрагу свих веза за одређени чвор у мрежи
- приказ статистичких информација

Приказ комплетне мреже концепата

Приказ комплетне мреже концепата омогућава преглед свих чворова у мрежи концепата, као и приказ свих веза између њих. Овај преглед је погодан за стицање општег утиска о целој структури и сложености хијерархије.

Приказ одређеног броја нивоа у мрежи концепата

Приказ одређеног броја нивоа омогућава преглед једног дела мреже концепата. У оквиру овог прегледа су приказани само чворови који припадају изабраном броју нивоа у графу. Ова опција је корисна када је мрежа концепата сложена јер даје кориснику могућност да анализира само горњи део мреже. Овај део мреже садржи атрибуте који су виши у хијерархији, а самим тим, имају и већи значај за одређени скуп података.

Преглед позиције одређеног атрибута у мрежи концепата

У оквиру овог приказа кориснику се нуди могућност претраге мреже концепата на основу атрибута формалног контекста (у овом случају тагова). Корисник уноси жељене тагове, након чега се мрежа претражује. Резултати претраге се приказују означавањем значајних чворова различитим бојама. Боја којом ће чвор бити означен зависи од подскупа атрибута које је корисник унео а који су обухваћени тим чвором, као и комбинације атрибута које тај чвор подржава. Алат не истиче све чворове који садрже неке од тражених атрибута, већ само најзначајније. Значај чвора се одређује на основу броја тражених атрибута обухваћених одређеним чвором и броја додатних атрибута. Под додатним атрибутима се подразумевају они атрибути који су обухваћени чвором, а које корисник није унео у поље за претрагу.

На основу значаја, чворови су обојени једном од следећих боја: црвеном, љубичастом, зеленом и плавом. Црвена боја се користи за означавање чворова који имају 100% подударање са листом тражених атрибута. Овом бојом се означава чвор који садржи све тражене атрибуте и нема ниједан додатни атрибут. За означавање чворова који садрже један део тражених атрибута, а притом немају ниједан додатни атрибут користи се љубичаста боја. Зелена боја се користи за означавање чворова који садрже све тражене атрибуте и минимални могући број додатних атрибута. За означавање чворова који садрже део тражених атрибута и минимални могући број додатних атрибута користи се плава боја. Овај начин визуализације је посебно користан за брзо откривање и позиционирање одређених атрибута.

Претрагу свих веза за одређени чвор у мрежи

Претрага веза омогућава кориснику да изабере један чвор у мрежи, након чега се чворови који су изнад и испод одабраног чвора у хијерархији означавају, као и везе између њих. Везе ка чворовима хијерархијски изнад одабраног чвора, као и сами чворови, означавају се зеленом бојом. Плавом бојом означавају се чворови који су хијерархијски испод одабраног чвора укључујући и везе ка њима. Овакав приказ омогућава корисницима да лако и јасно, анализирају везе између одређеног подскупа чворова.

Приказ статистичких информација

Поред различитих опција за приказ, кориснику се даје могућност прегледа сумарних информација о самој мрежи концепата и појединачним чворовима.

У оквиру сумарних информација о мрежи концепата приказују се информације о укупном броју чворова у мрежи, укупном броју скупова података у мрежи, укупном броју атрибута у мрежи, као и број нивоа на којима су чворови распоређени.

Избором одређеног чвора, могу да се погледају све информације у вези са изабраним чвором: назив чвора, ниво на коме је чвор позициониран, број атрибута које садржи, као и листу самих атрибута, листу објеката обухваћених изабраним чвором и број објеката који су представљени обухваћеним атрибутима.

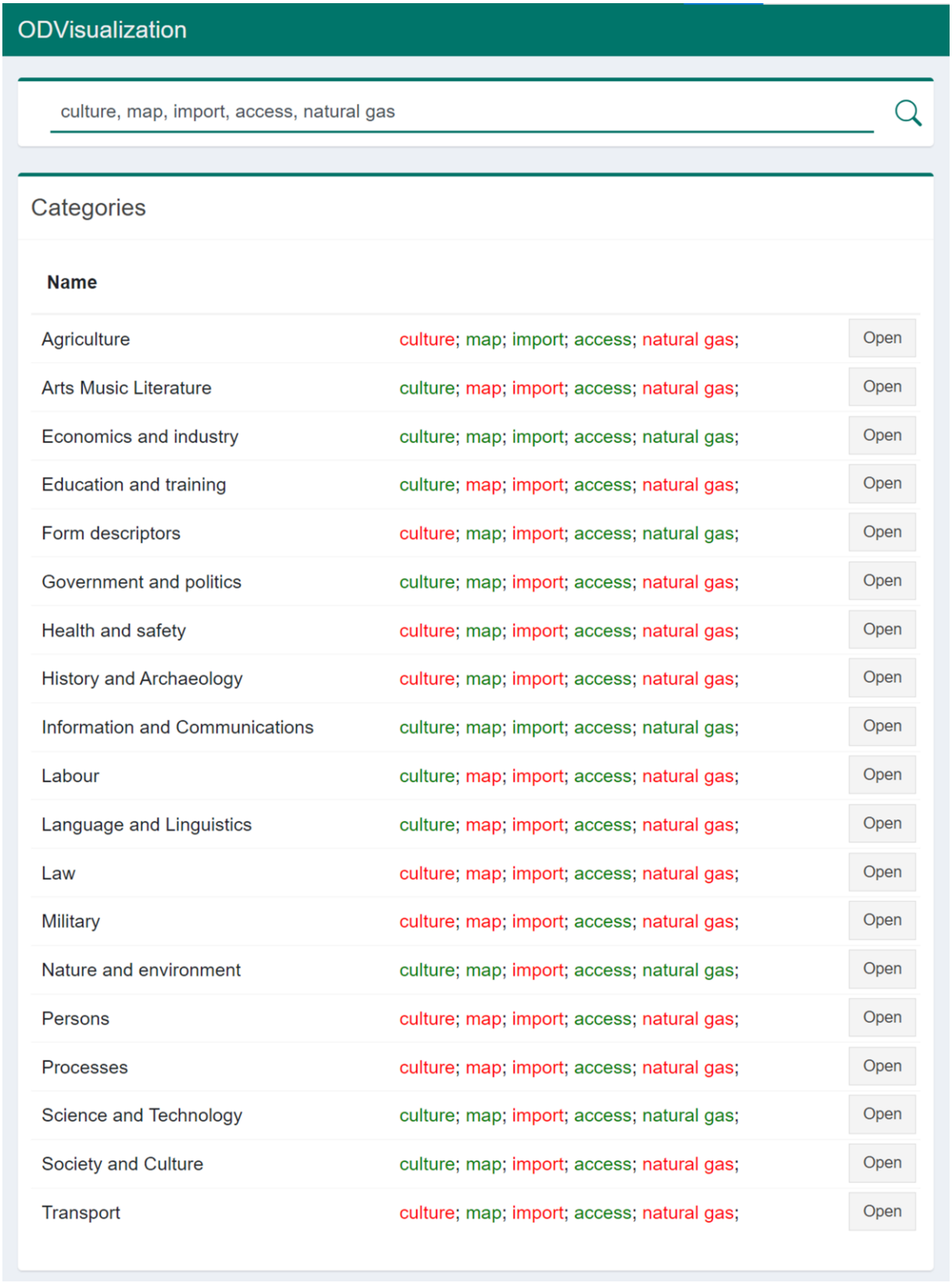
Наведене опције за приказ могу се комбиновати што додатно олакшава анализу позиције неког атрибута, а самим тим и употребу и значај тог атрибута у једној мрежи концепата. Комбиновањем опција за приказ, приказ једне мреже може прилично да се

упрости и да се извуку само значајни делови мреже, што је посебно важно за мреже са великим бројем чворова и веза.

У наставку овог поглавља докторске дисертације приказан је пример употребе овог алата за анализу употребе тагова на једном порталу отворених података. За потребе овог примера коришћени су подаци са канадског портала отворених података [87] из 2020. године. У том периоду на порталу је било преко 80.000 сетова података организованих у 19 категорија при чему су неке категорије садржале значајан број сетова података и тагова који су коришћени за описивање тих података.

Како би се омогућила анализа тагова на целом порталу отворених података, за сваку доступну категорију на порталу, креирана је по једна мрежа концепата на основу свих сетова података који припадају истој категорији и тагова који описују те податке. За креирање мрежа концепата на основу датих формалних контекста коришћен је *NextClosure* алгоритам [77], при чему је излаз алгоритма адаптиран да памти све потребне информације о концептима описане у претходном делу овог поглавља. Покретањем апликације кориснику се приказују све мреже концепата доступне у библиотеци апликације, у овом случају по једна мрежа концепата за сваку од категорија (слика 19).

У примеру који је приказан на слици 19, демонстрирана је и функционалност претраге тагова на нивоу свих мрежа. У поље за претрагу, које се налази на врху слике 19, унети су тагови *culture, map, import, access, natural gas.* Након тога, поред сваке од мрежа концепата приказана је заступљеност сваког од унетих тагова. Црвеном бојом означени су тагови који се не налазе у мрежи а зеленом они који се налазе. Може се приметити да се у категоријама *Economics and industry* и *Information and Communications* појављују сви тражени тагови. Визуализација мреже се отвара кликом на дугме *Open*.

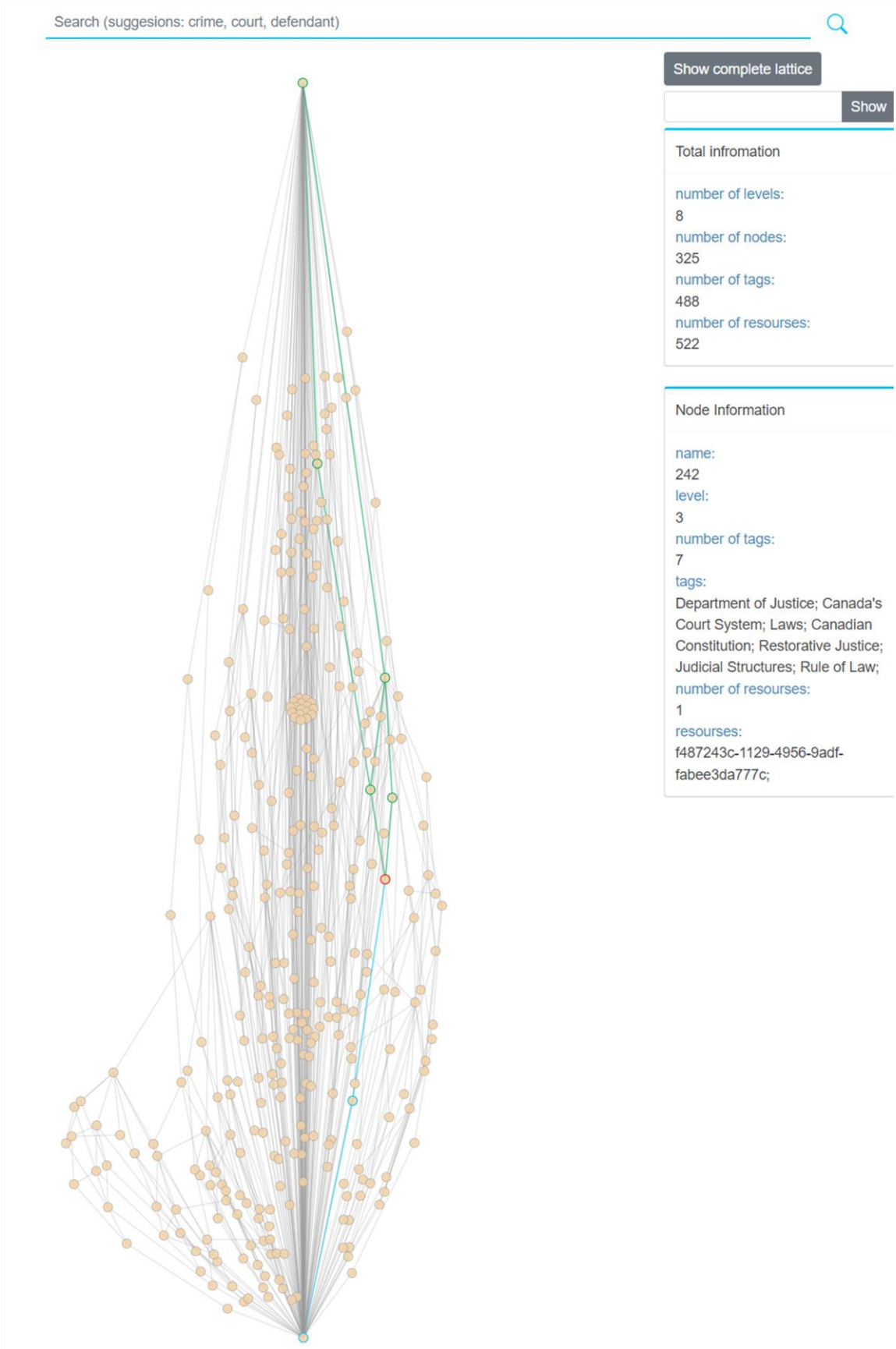


Слика 19 – Алат за визуализацију - избор мреже за анализу

У оквиру ове докторске дисертације, за потребе демонстрације визуалне анализе користи се категорија *Law* са канадског портала отворених података. Ова категорија је једна од мањих категорија на овом порталу, како по броју скупова података, тако и по броју различитих тагова који се користе за описивање скупова података ове категорије. Поред тога што је једна од мањих категорија, мрежа концепата генерисана за ову категорију је прилично сложена и има 8 нивоа у оквиру којих се налази 325 чворова. Преглед сумарних информација за целу мрежу концепата за категорију *Law*, као и пример сумарних информација за појединачни изабрани чвор у мрежи приказан је на слици 20.

У првом делу информација се приказују информације о целој мрежи, а након тога приказане су информације о једном чвору. Селекција чвора чије информације корисник жели да види се постиже кликом на тај чвор, чиме се тај чвор у мрежи обележава црвеном ивицом и покреће се приказ свих веза са којима је изабрани чвор у мрежи повезан, како уз тако и низ хијерархију.

За изабрани чвор на слици се може видети да се налази на трећем нивоу у мрежи, да подржава укупно седам тагова чија је листа приказана и да та комбинација одговара једном сету података чија је ознака такође приказана. Везе ка чворовима који су изнад њега у хијерархији, као и ивице тих чворова, означене су зеленом бојом, док су плавом означене везе ка чворовима који су испод њега у хијерархији, укључујући и ивице тих чворова.



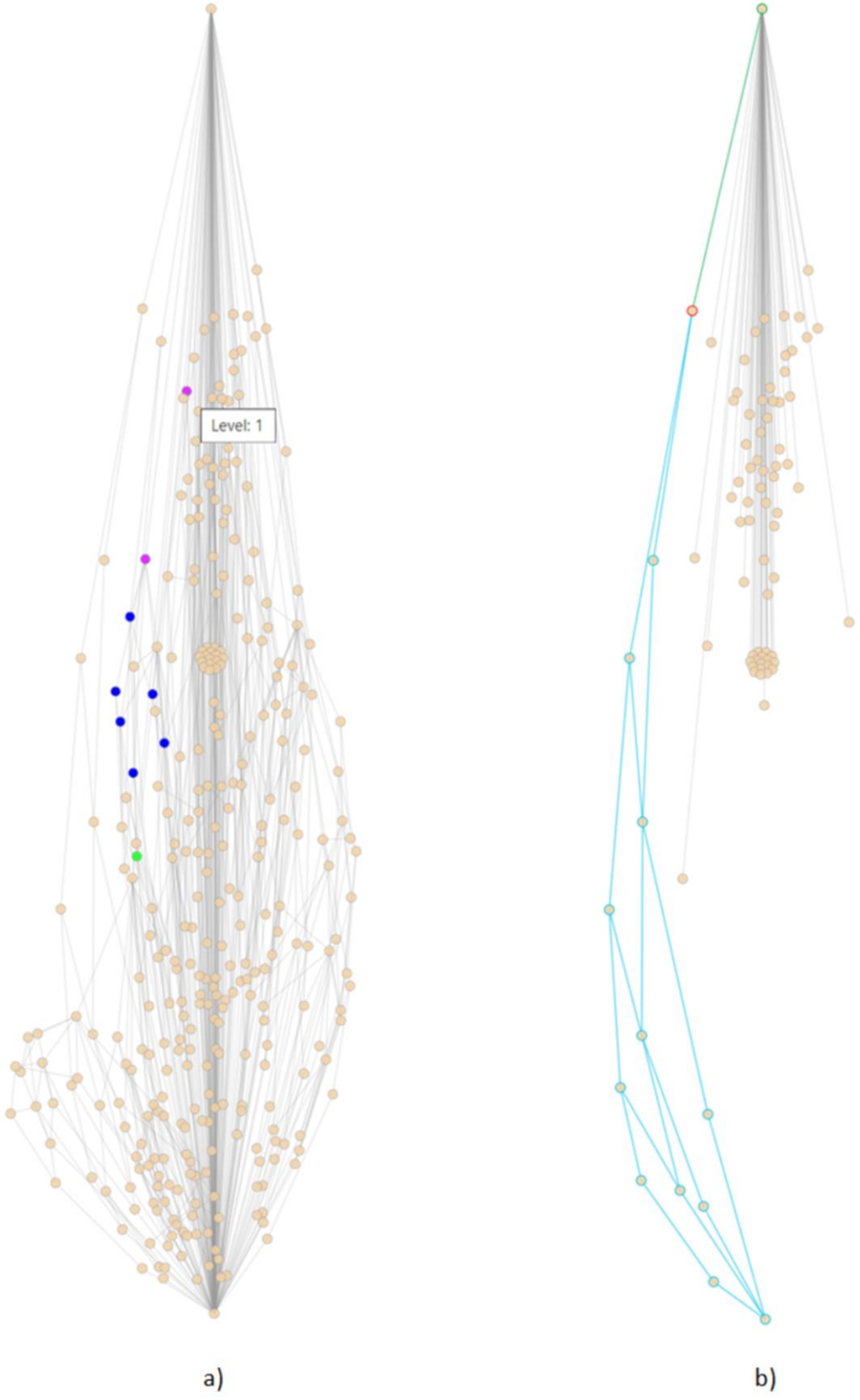
Слика 20 – Приказ збирних информација за целу мрежу концепата и приказ информација о изабраном чвору

Визуални приказ целе мреже концепата, посебно комплексне као што је ова, може се користити као први корак у анализи, као средство креирања опште слике о мрежи и позицији неких од чворова. Приказ комплетне мреже концепата категорије *Law* приказан је на слици 21а. У оквиру овог примера, приказ целе мреже је комбинован са претрагом чворова на основу тагова које подржавају. У приказаном примеру у поље за претрагу су унете вредности *Research* и *Annual reports.* Из приказаних резултата може се закључити да мрежа не подржава чвор који одговара унетој комбинацији тагова, а да притом нема додатне вредности. Међутим, у мрежи су означена два љубичаста чвора, један садржи само таг *Research*, док други садржи само таг *Annual reports.* Постављањем курсора на ова два чвора се може приметити да су оба чвора на првом нивоу у мрежи концепата. Поред ових чворова, означено је више чворова који садрже део тражених тагова и исти минимални број додатних тагова - у овом примеру по један додатан таг. За тражене тагове, препознат је и један зелени чвор који подржава оба тражена тага и три додатна.

За дубљу анализу, попут анализе веза између чворова, велики број чворова и веза између њих може изазвати проблеме у сагледавању приказаних информација. Из тог разлога, како би се омогућио јаснији преглед, алат нуди могућност уклањања неких чворова из приказа кроз опцију одабира само неколико нивоа који ће бити приказани. На овај начин, приказ се растерећује од чворова који нису од интереса.

На слици 21б је приказан један пример анализе са неким од чворова који су сакривени, при чему је комбинован приказ изабраног броја нивоа у мрежи концепата са приказом за претрагу свих веза за одабрани чвор. У овом примеру, значајно је смањен број приказаних чворова у мрежи постављањем број нивоа мреже за приказ на један. Последично, омогућено је да се над растерећеним приказом, селектовањем чвора од интереса уради лакша анализа свих његових веза.

Као што се може из примера видети, овакав преглед је много јаснији у поређењу са оним који се је приказан на слици 21а. Овај тип прегледа се може користити за детаљну анализу односа између подскупова међусобно повезаних тагова унутар једне категорије. Приказани алат пружа неколико режима за преглед и анализу мреже концепата који се могу комбиновати како би се омогућила лака и прегледна анализа у зависности од потреба корисника.



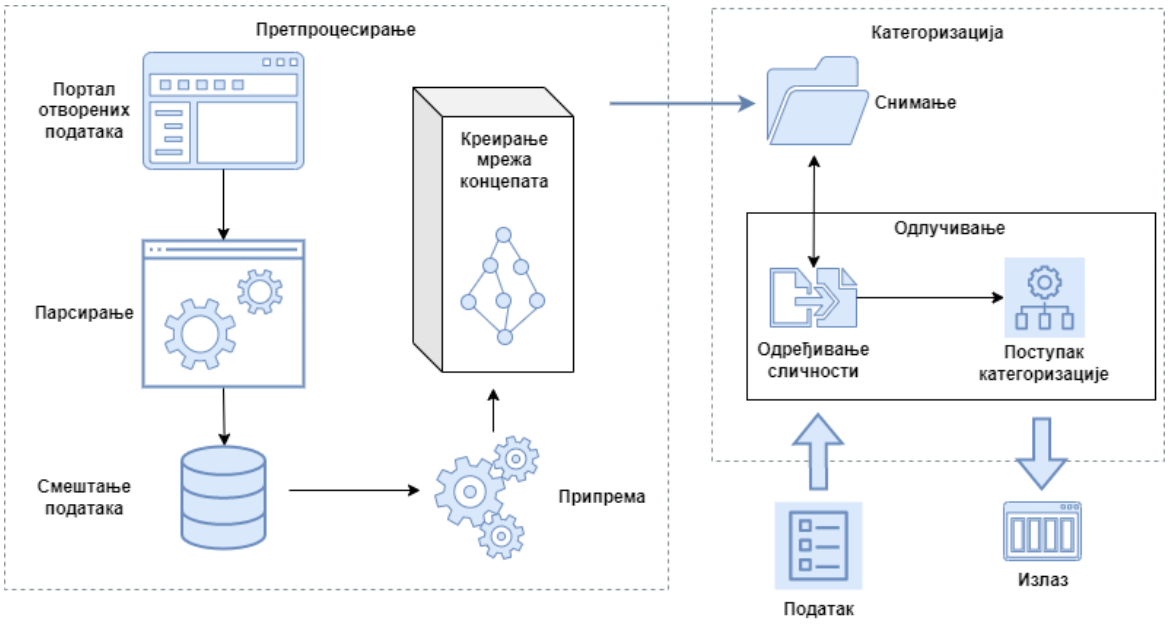
Слика 21 – а) Комплетна мрежа концепата категорије *Law* у комбинације са претрагом; б) Комбинација различитих опција за приказ категорије *Law*

7. ПОСТУПАК КАТЕГОРИЗАЦИЈЕ ПОДАТАКА

Као што је приказано у првом делу ове докторске дисертације, на порталима отворених података није реткост да неким подацима није додељена категорија. Како би могла да се повећа видљивост ових података и омогући њихово лакше проналажење на порталима а са тим олакша њихова даља употреба, потребно је оваквим подацима доделити категорију која им недостаје.

Одређивање адекватне категорије може да се уради на основу постојећих вредности мета-кључева који су део метаподатака сетова података. У оквиру ове дисетације разматрана је употреба вредности мета-кључа који представља тагове како би се одредила једна или више категорија којима податак може да припада. На тај начин за један податак који је потребно категорисати уноси се комбинација тагова која га описује а алгоритам враћа листу категорија којима податак може да припада.

Алгоритам за категоризацију предложен у оквиру ове дисертације предвиђен је да извршава категоризацију података у оквиру једног портала. За потребе категоризације потребно је да се од постојећих, већ категорисаних података, креира база знања која се употребљава као основа за одређивање категорија некатегорисаних података. Након тога, може се приступити категоризацији података уношењем тагова којима се податак описује. Из тога разлога, цео процес може се поделити у два дела приказана на слици 22– претпроцесирање и категоризација података.



Слика 22 – Процес категоризације података

7.1. Претпроцесирање

Први део процеса категоризације података је претпроцесирање које представља припрему постојећих података и креирање базе знања која се употребљава приликом категоризације. Овај корак се састоји од неколико међукорака при чему први представља прикупљање свих постојећих података на порталу.

Различити портали могу да користе различите мета-кључеве за памћење информација о подацима. Ово се односи и на памћење информација које су тема ове докторске дисертације а то су информације о категоријама којима податак припада и таговима којима је податак описан. Из тог разлога, на нивоу портала за који се припрема категоризација, потребно је дефинисати који се мета-кључеви користе за памћење ових информација. На основу одабраних мета-кључева, за сваки податак извршава се филтрирање метаподатака и издвајање само оних информација које су значајне за овај процес. Имајући у виду да један податак може да буде категорисан у више категорија врши се разврставање сређених података по категоријама. За сваку категорију која је доступна на порталу креира се по један сет података који садржи само оне податке који припадају датој категорији. Подаци по категоријама памте се у JSON формату као листа података описана:

- јединственим идентификатором податка и
- листом тагова који описују податак.

На основу овако припремљених података прелази се на креирање базе знања на нивоу категорија. Имајући у виду да један таг може да се јави у више категорија и у различитим комбинацијама са другим таговима, у оквиру ове докторске дисертације база знања базираће се на хијерархијској организацији тагова који се јављају на порталу. На основу овакве хијерархије може се утврдити начин употребе тагова, важност одређеног тага али и комбинације тагова као и начин појављивања у подацима описаних једном категоријом. Из тог разлога, за сваку од категорија доступних на порталу креира се по једна хијерархијска организација тагова. Одређивање једне овакве хијерархије обављено је употребом Анализе формалних концепата приказане у претходном делу ове дисертације. За потребе ове докторске дисертације искоришћен је *NextCloseur* ФЦА алгоритам који на основу скупа података креира мрежу концепата формалног контекста креираног над тим сетом података. Мрежа концепата памти се у JSON формату и садржи информације о концептима и везама између концепата, при чему се сваки концепт у мрежи дефинише:

- јединственим идентификатором који се дефинише при креирању мреже,
- комбинацијом атрибута концепта,
- листом сетова података који су у формалном контексту дефинисани задатом комбинацијом атрибута (листа података).

На крају овог корака, за сваку од категорија доступних на порталу памти се по једна мрежа концепата, при чему је сваки концепт у мрежи дефинисан једном комбинацијом тагова која ја описује. Све мреже концепата заједно чине базу знања на основу које се ради категоризација података уношењем комбинације тагова који описују задати податак.

7.2. Категоризација

Након дела претпроцесирања и креирања базе знања може се приступити процесу категоризације података. Категоризација као улаз прима скуп тагова који описују податак који је потребно категорисати $T = \{t_1, t_2, \dots, t_k\}$, где је k број тагова који описују податак. За унету комбинацију рачуна се сличност са сваком категоријом Cat_i, т.ј. мрежом концепата L_i креиране за сваку од категорија која је део генерисане базе знања. На основу резултата сличности по категоријама предлажу се категорије којима податак може да припада. Скуп акција које се извршавају у оквиру алгоритма категоризације приказан је у наставку:

1. Креирати листу укупних резултата $R = [R_1, R_2, \dots, R_n]$, где је n број категорија на порталу
2. За сваку категорију Cat_i на порталу:
   2.1. Креирати листу резултата SR_i за категорију Cat_i
   2.2. Учитати мрежу концепата L_i за категорију Cat_i
   2.3. За сваки чвор $node_j$ у мрежи концепата L_i
      2.3.1. Израчунати сличност са таговима из скупа T
      2.3.2. Уписати резултате сличности у листу резултата SR_i
   2.4. На основу резултата уписаних у SR_i, израчунати сличност за целу категорију R_i
   2.5. Уписати резултат R_i у листу укупних резултата R
3. На основу резултата за све категорије R применити процес одређивања категорија (поступак категоризације)
4. Вратити предложену листу категорија.

У приказаном алгоритму могу се издвојити три дела:

- одређивање сличности између две комбинације тагова – одређује сличност између унете комбинације тагова и комбинације тагова које чине један концепт у мрежи концепата,
- одређивање сличности комбинације унетих тагова и категорије – на основу резултата сличности унете комбинације тагова са свим концептима у мрежи концепата рачуна се сличност за једну категорију,
- алгоритам категоризације – избор категорија на основу резултата сличности по категоријама.

Сваки од ових делова детаљно је објашњен у наставку.

7.3. Одређивање сличности између две комбинације тагова

Одређивање сличности између две комбинације тагова користи се како би се одредила сличност комбинације тагова која описује податак који је потребно категорисати са таговима који чине један концепт у мрежи концепата. Рачунање ове сличности се састоји из два дела:

- одређивање сличности између два тага и
- одређивање сличности између две листе тагова.

Рачунање сличности између два тага се заснива на упоређивању значења унетих комбинација речи које су саставни део тагова. Из тог разлога, у првом кораку за рачунање семантичке сличности термина употребљени су глобални вектори за представљање речи (GloVe) како би се речи које је потребно упоредити представиле векторима.

Глобални вектори за представљање речи (*GloVe - Global Vectors for Word Representation*) [78] је метод за креирање векторских репрезентација речи који су креирали Pennington, Socher и Manning на Стендфорд универзитету 2014. године. *GloVe* је регресиони модел који комбинује предности две главне групе модела: факторизација глобалних матрица, попут латентне семантичке анализе (*LSA - Latent Semantic Analysis*) [79] и метода заснованих на локалним прозорима контекста (*Local Context Window Methods*) попут *skip-gram* модела [80]. Модел производи векторски простор са смисленом подструктуром и даје боље резултате од сличних модела у рачунању сличности [78]. Унапред тренирани *GloVe* модели су јавно доступни и дали су боље

резултате од сличних модела и метода (FastText [81], ngram2vect метода [82], dict2vect метода [83]) над *WordSim-500* сету података [84].

У оквиру ове докторске дисертације коришћен је *GloVe* модел који је унапред трениран над *Common Crawl* [85] подацима и садржи 840 милијарди генерисаних токена и речником који садржи 2,2 милиона уноса. Овај модел је изабран због разноврсности података које садржи као и чињенице да вектори представљају појединачне речи чиме се омогућава поређење сличности на ниво речи.

Одређивање граничне вредности у случајевима када се сличност одређује коришћењем *word embedding* модела је изазован задатак. Због својих карактеристика, *word embedding* модели могу да доведу до незнатно другачијих дистрибуција речи чак и у случајевима када се током процеса тренирања користе исти параметри и исти подаци [86]. За потребе ове докторске дисертације, праћене су препоруке и резултати приказани у раду [86]. Подешавања експеримента и модела приказаних у овом истраживању извештавају о ефектима димензионалности *word embedding* модела на вредност прага сличности спровођењем експеримената у различитим векторским димензијама. Њихово истраживање сугерише да је одговарајућа гранична вредност 0,7 за моделе са 300 димензија, као што је модел који се користи у оквиру ове докторске дисертације.

Како би се одредила сличност тагова треба да се узме у обзир да се сваки таг састоји од једне или више кључних речи - термина које ближе описују податке. Поред тога, анализом тагова који се употребљавају за описивање података на порталима отворених података, примећено је да се у таговима јављају чланови *a, an* и *the,* као и да то није правило које сви примењују, па се ови чланови често и изостављају. Као последица неконзистентне употребе чланова нису ретке ситуације у којима се неке речи јављају у неким таговима са, а у другим без чланова. У наставку је дато неколико примера таквих тагова на канадском порталу отворених података:

- *„The presence“ и „presence“*
- *„The minimum wage“ и „minimum wage“*
- *„North“ и „The north“*
- *„First“ и „The first“*
- *„Regulations“ и „The regulations“*
- *„Council“ и „An council“*
- *„Error“ и „An error“*

- *„Electronic cigarette“ и „An electronic cigarette“*
- *„A mask“ и „mask“*
- *„A pan-Canadian“ и „pan-Canadian“*
- *„Tick“ и „A tick“*
- *„National“ и „A national“*
- *„Vaccine“, „vaccine“ и „a vaccine“*
- *„Resource“ и „A resource“*

Додатно је примећено да неки тагови у себи садрже више независних термина раздвојених карактером „;“ попут тагова:

- *„fish; fishery; purchases; port;“*
- *„oil sands; monitoring; biodiversity; contaminants; amphibians“*
- *„water level; flow; discharge; hydrometric; hydrology; water quantity; sediment; oil sands; monitoring“*
- *„human rights; treaties; United Nations; legislation; justice; country profile; demographic; data; non-discrimination“*
- *„general public; report; taxpayers; income statistics; statistics; Canada Revenue Agency; personal income taxes“*

Из тог разлога, пре рачунања семантичке сличности комбинација тагова ради се претпроцесирање тагова у смислу уклањања свих појављивања чланова *a*, *an* и *the*, док су за раздвајање речи употребљени карактери размак („ “), „„“ и „;“. Након уклањања чланова, тагови се представљају њиховим векторским репрезентацијама коришћењем *GloVe* модела а сличност између два тага A и B одређује се дефиницијама 1 и 2:

Дефиниција 1: За таг A представљен речима $Ra_i, i \in \{1,2,\dots n\}$, при чему је n број речи у тагу A, и таг B представљен речима $Rb_j, j \in \{1,2,\dots m\}$, при чему је m број речи у тагу B:

- Сличност (S) је 1 уколико су тагови A и B исти,
- Уколико нису исти сличност се одређује као

$$S = avg\big(similarity(Ra_i, B)\big), \text{ за } i \in \{1,2,\dots n\}.$$

где $similarity(Ra_i, B)$ представља сличност између речи Ra_i и тага B.

Дефиниција 2: Сличност $similarity(X, Y)$ између речи X и тага Y представљен речима $Y_j, j \in \{1,2,\dots m\}$, где је m број речи у Y, се одређује као

$$similarity(X, Y) = max\left(sim\big(X, Y_j\big)\right), \text{ за } j \in \{1,2,\dots m\},$$

где $sim(X, Y_j)$ представља косинусну сличност између речи X и речи Y_j.

На основу ових дефиниција сличност између два тага A и B који се састоје од више речи се одређује као просек сличности сваке речи тага A са тагом B, при чему се сличност једне речи тага A са тагом B одређује као највећа сличност те речи са сваком од речи тага B.

Резултати сличности два тага се користе као основа за рачунање сличности између две комбинације тагова TL_A и TL_B, при чему се TL_A састоји од T_{Ai} тагова, где је $i \in \{1,2, \dots n\}$, и n представља број тагова у комбинацији тагова TL_A, док се TL_B састоји од T_{Bj} тагова, где је $j \in \{1,2, \dots m\}$, и m представља број тагова у комбинацији тагова TL_B.

За овако дефинисане комбинације тагова TL_A и TL_B сличност се одређује са неколико параметра:

- *NumTagSim1* – параметар који чува информацију о броју тагова комбинације тагова TL_A који имају сличност 1 са најмање једним тагом комбинације тагова TL_B,
- *FM* – параметар који чува информацију да ли постоји потпуно поклапање између комбинација тагова TL_A и TL_B. Потпупо поклапање значи да сваки таг комбинације тагова TL_A има таг у комбинацији тагова TL_B са којим има сличност 1 при чему у комбинацији тагова TL_B нема додатних тагова са којима није одређена сличност 1 са неким од тагова из TL_A. Овај параметар има вредност 1 уколико је потпуно поклапање пронађено, односно 0 уколико није,
- *amSim* – параметар који чува просек највећих сличности комбинације тагова TL_A са комбинацији тагова TL_B.
- *maSim* – параметар који чува највећу просечну сличност комбинације тагова TL_A са комбинацијом тагова TL_B.

Вредност параметра *amSim* за комбинације тагова TL_A и TL_B одређује се коришћењем следећих дефиниција:

Дефиниција 3: Просек највећих сличности за комбинацију тагова TL_A са комбинацијом тагова TL_B рачуна се као

$$amSim(TL_A, TL_B) = avg(msim(T_{Ai}, TL_B)), \text{ за } i \in \{1,2, \dots n\},$$

где $msim(T_{Ai}, TL_B)$ представља највећу сличност сваког тага из комбинације тагова TL_A са комбинацијом тагова TL_B.

Дефиниција 4: Највећа сличност једног тага T са комбинацијом тагова TL, се рачуна као

$$msim(T, TL) = max(S(T, TL_k)), \text{ за } k \in \{1,2, \dots l\},$$

где TL_k представља један таг у комбинацији тагова TL која садржи укупно k тагова.

Вредност параметра *maSim* за комбинације тагова TL_A и TL_B одређује се коришћењем следећих дефиниција:

Дефиниција 5: Највећа просечна сличности за комбинацију тагова TL_A са комбинацијом тагова TL_B рачуна се као

$$maSim(TL_A, TL_B) = max(asim(T_{Ai}, TL_B)), \text{ за } i \in \{1,2, \dots n\},$$

где $asim(T_{Ai}, TL_B)$ представља просечну сличност сваког тага из комбинације тагова TL_A са комбинацијом тагова TL_B.

Дефиниција 6: Просечна сличност једног тага T са комбинацијом тагова TL, се рачуна као

$$asim(T, TL) = avg(S(T, TL_k)), \text{ за } k \in \{1,2, \dots l\},$$

где TL_k представља један таг у комбинацији тагова TL која садржи укупно k тагова.

7.4. Одређивање сличности комбинације унетих тагова са категоријом

Одређивање сличности једне комбинације тагова са једном категоријом представља тражење укупне сличности те комбинације тагова са свим концептима који се налазе у мрежи концепата која престаљa ту категорију. Рачунање сличности унетих тагова са једним концептом може да се посматра као рачунање сличности између две комбинације тагова, при чему је прва комбинација листа унетих тагова податка које је потребно категорисати а друга комбинација престаљa тагове који чине један концепт мреже концепата.

Ова сличност се одређује на основу вредности параметара *NumTagSim1, FM, amSim* и *maSim* који се добијају на основу рачунања сличности унете комбинације тагова са сваким чвором у мрежи концепата појединачно и представљена је помоћу пет параметара:

- *FMK* – параметар који има вредност једнаку укупном броју чворова код којих параметар *FM* има вредност 1 приликом рачунања сличности са унетом комбинацијом тагова. Уколико такав чвор не постоји у мрежи концепата овај параметар има вредност 0,

- *NumTags* – параметар који памти информацију о томе колико је највише тагова унете комбинације тагова имало сличност 1 са чворовима у мрежи концепата.
- *maxAverageM* – параметар који одређује највећи просек највећих сличности унете комбинације тагова са свим чворовима у мрежи концепата,
- *avgAverageM* – параметар који одређује просек свих просека највећих сличности унете комбинације тагова са свим чворовима у мрежи концепата,
- *maxMaximumA* – параметар који одређује највећу вредност највећих просечних сличности унете комбинације тагова са свим чворовима у мрежи концепата.

Формално одређивање вредности параметра *NumTags* може да се дефинише следећом дефиницијом:

Дефиниција 7: Највећи број тагова комбинације тагова TL који има сличност 1 у једном чвору за целу мрежу концепата L категорије Cat рачуна се као

$$NumTags\ (TL, L) = max(NumTagSim1(TL, Node_i)),\ i \in \{1,2, \dots n\},$$

где $Node_i$ представља један чвор у мрежи концепата L а n укупан број чворова мреже концепата.

Одређивање вредности параметра *maxAverageM* формално може да се дефинише следећом дефиницијом:

Дефиниција 8: Највећи просек највећих сличности комбинације тагова TL са мрежом концепата L категорије Cat рачуна се као

$$maxAverageM(TL, L) = max(amSim(TL, Node_i)),\ i \in \{1,2, \dots n\},$$

где $Node_i$ представља један чвор у мрежи концепата L а n укупан број чворова мреже концепата.

Додатно, одређивање вредности параметра *avgAverageM* формално може да се дефинише следећом дефиницијом:

Дефиниција 9: Просек свих просека највећих сличности комбинације тагова TL са мрежом концепата L категорије Cat рачуна се као

$$avgAverageM(TL, L) = avg(amSim(TL, Node_i)),\ i \in \{1,2, \dots n\},$$

где $Node_i$ представља један чвор у мрежи концепата L а n укупан број чворова мреже концепата.

Одређивање вредности параметра *maxMaximumA* формално може да се дефинише следећом дефиницијом:

Дефиниција 10: Највећа вредност највећих просечних сличности комбинације тагова TL са мрежом концепата L категорије Cat рачуна се као

$$maxMaximumA\ (TL, L) = max(maSim(TL, Node_i)),\ i \in \{1,2, \dots n\},$$

где $Node_i$ представља један чвор у мрежи концепата L а n укупан број чворова мреже концепата.

7.5. Алгоритам категоризације

Алгоритам за категоризацију се ослања на параметре сличности израчунате за унету комбинацију тагова и сваку од доступних категорија на порталу отворених података. Ови резултати су улаз у алгоритам и на основу њихових вредности за сваку од категорија, алгоритам предлаже једну или више категорија за унету комбинацију тагова. Дијаграм алгоритма приказан је на слици 23.

Процес одлучивања се заснива на неколико критеријума. Први критеријум приликом одабира категорије за једну комбинацију тагова, јесте да ли у некој од категорија параметар *FMK* има вредност већу или једнаку 1. Вредност параметра *FMK* за једну категорију је једнака броју чворова одговарајуће мреже концепата који имају потпуно поклапање са унетом комбинацијом тагова. То значи да та комбинација тагова постоји самостално у хијерархији тагова и тиме има већи значај у целокупној хијерархији. Категорије у којима су пронађени такви чворови имају предност у односу на друге категорије у којима нема потпуних поклапања и оне се издвајају као кандидати за одабир категорија.

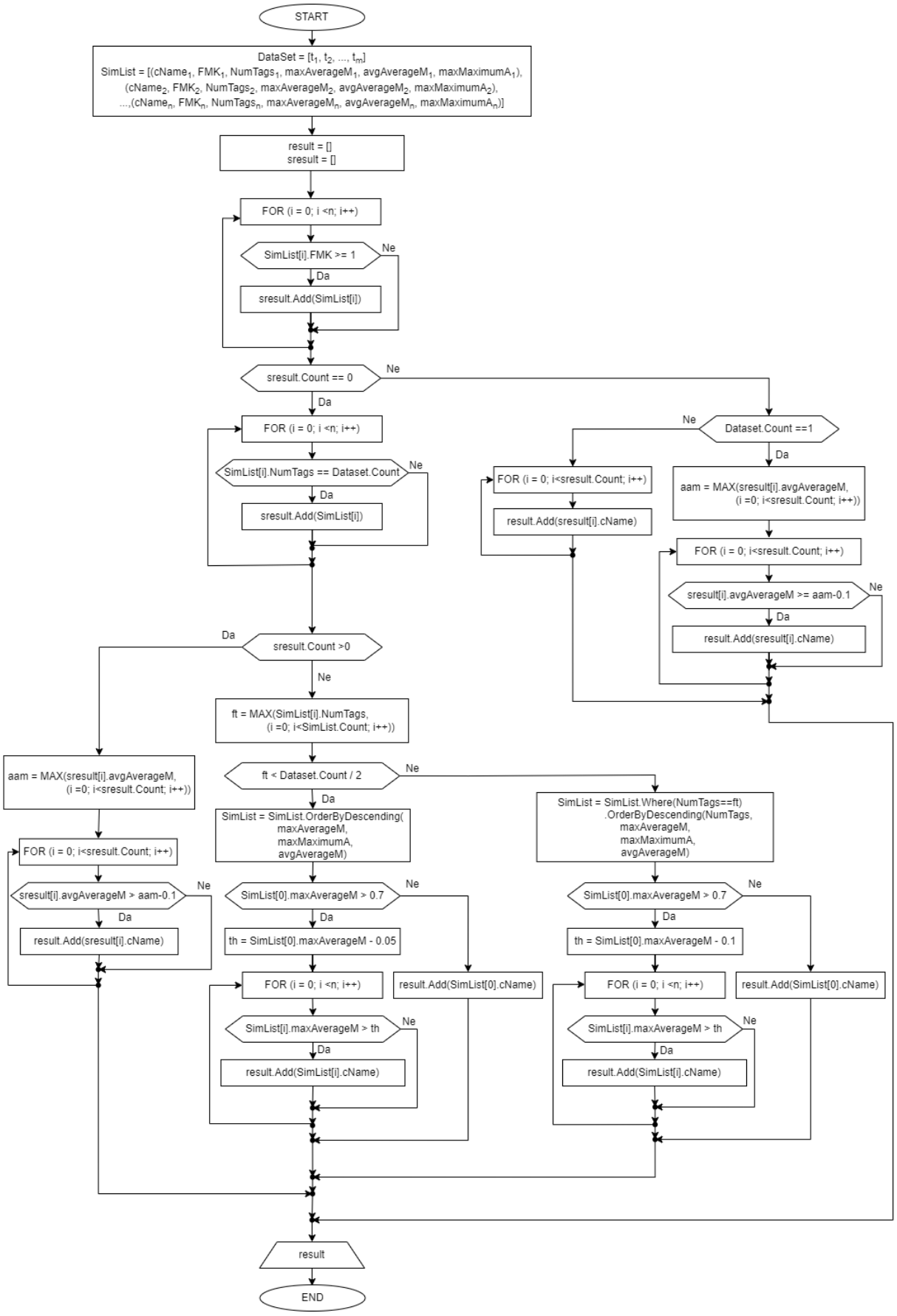
Уколико има издвојених категорија кандидата, потребно је размотрити колико има тагова у унетој комбинацији тагова. Експерименталним путем, примећено је да за категоризацију података описаних само једним тагом, некада има много категорија са потпуним поклапањем. Број оваквих категорија зависи од вредности тага, и што је вредност тага генералнија то је већа вероватноћа да се јавља у више категорија. Међутим, иако се таг појављује у већем број у категорија, његово појављивање не мора бити често и у већем броју комбинација, те је из тога разлога потребно урадити додатно филтрирање категорија у тим ситуацијама. Стога, уколико је категоризацију потребно урадити на основу више унетих тагова, кориснику се предлажу све категорије које се налазе у листи категорија кандидата. Међутим, уколико је категоризацију потребно урадити на основу само једног тага, ради се додатно филтрирање категорија кандидата на основу вредности параметра *avgAverageM* који указује на заступљеност унете

комбинације у целој мрежи концепата. Тражи се највећа вредност овог параметра међу издвојеним категоријама и креира се праг који износи *avgAverageM – 0,1*. Све категорије међу кандидатима које имају вредност параметра *avgAverageM* већу или једнаку од дефинисаног прага предлажу се кориснику.

Уколико нема категорија у којима се јавља потпуно поклапање, издвајају се категорије у којима вредност параметра *NumTags* има исту вредност као и број тагова на основу којих се ради категоризација. Чворови у мрежи концепата који задовољавају овај критеријум представљају чворове у којима су нађене сличности 1 за све унете тагове али који садрже и неке друге атрибуте који немају сличност 1 са унетим таговима. Пошто има додатних атрибута у тим чворовима, овакви чворови не могу да се издвоје као потпуна поклапања, али упућују да се унета комбинација тагова јавља заједно у хијерархији тагова, те да је допуњена додатним атрибутима, па се овакви чворови могу сматрати потпуним поклапањима са додатком.

Све категорије које задовољавају овај критеријум постају категорије кандидати. Уколико има нађених категорија кандидата ради се њихово даље филтрирање на основу параметра *avgAverageM.* Тражи се највећа вредност овог параметра међу издвојеним категоријама и креира се праг који износи *avgAverageM – 0,1*. Све категорије међу кандидатима које имају вредност прага *avgAverageM* већу од дефинисаног прага предлажу се кориснику.

У ситуацији када није пронађена ниједна категорија у којој вредност параметра *NumTags* има исту дужину као и број тагова на основу којих се ради категоризација, ради се одабир категорија на основу преосталих параметара. Проверава се за сваку категорију, колико највише унетих тагова има сличност 1 у једном чвору. Уколико је та вредност мања од 50% свих унетих тагова, за све категорије, ради се избор категорија који није заснован на вредности параметра *NumTags.*



Слика 23 – Алгоритам за поступак категоризације

У овом случају, прво се ради сортирање свих категорија у опадајући редослед према вредности параметра *maxAverageM,* Додатно, ако више категорија има исту вредност овог параметра, ради се сортирање у опадајући редослед на основу параметра *maxMaximumA,* а затим по параметру *avgAverageM,* такође, у опадајући редослед. Након тога, проверава се да ли прва категорија у сортираној листи категорија, има вредност параметра *maxAverageM* већу од 0,7. Уколико има, из листе категорија кориснику се предлажу оне категорије код којих је вредност параметра *maxAverageM* већа од вредности параметра *maxAverageM* првог елемента у листи умањене за 0.05. Међутим, у случају да је вредност параметра *maxAverageM* првог елемента у листи мања од 0,7, сматра се да унета комбинација има слабу сличност са категоријама и кориснику се предлаже прва категорија из сортиране листе.

У супротном, ако је било категорија у којима је вредност параметра *NumTags* била већа или једнака половини свих унетих тагова избор категорије се разликује. На почетку се издвајају све категорије који имају највећу вредност *NumTags* и сортирају у опадајући редослед по вредности параметра *NumTags.* Додатно, ако више категорија има исту вредност овог параметра ради се сортирање на основу параметра *maxAverageM* у опадајући редослед, а затим по параметру *maxMaximumA* и на крају параметру *avgAverageM,* такође, у опадајући редослед. Након тога, проверава се да ли прва категорија у сортираној листи категорија има вредност параметра *maxAverageM* већу од 0,7. Уколико има, из листе категорија кориснику се предлажу оне категорије код којих је вредност параметра *maxAverageM* већа од вредности параметра *maxAverageM* првог елемента у листи умањене за 0.1. Додатно, ако је вредност параметра *maxAverageM* првог елемента у листи мања од 0,7, сматра се да унета комбинација има слабу сличност са категоријама и кориснику се предлаже прва категорија из сортиране листе. Овако дефинисан алгоритам на излазу предложиће једну или више категорија које испуњавају дефинисане критеријуме.

7.6. Имплементација поступка категоризације

Имплементација поступака категоризације, приказаног у претходном поглављу ове докторске дисертације, урађена је са циљем да се на реалним примерима валидира рад предложеног поступка. Комплетан поступак је имплементиран у два дела:

- Веб АПИ *CategorizationAPI* и
- Веб апликација *ODCategorization.*

CategorizationAPI је реализован као REST API у оквиру којег је имплементиран описани поступак категоризације. Овом АПИ-ју се прослеђује листа тагова који описују неки сет података, а као одговор се добија резултат категоризације састављен од листе предложених категорија, као и сви параметри на основу којих се ради категоризација, за сваку од доступних категорија. *CategorizationAPI* се ослања на већ постојећу и припремљену базу знања за неки портал отворених података.

АПИ имплеметира GET методу којој се прослеђује листа текстуалних вредности (листа тагова). Curl позив ове методе је следећи:

```
curl --location
'{API_URL}/api/Categorization?tagsList={TAG_1}&tagsList={TAG_2}..&
tagsList={TAG_N}'
```

при чему API_URL представља адресу на којој се АПИ налази, а TAG_1, TAG_2 до TAG_N представљају листу тагова која се прослеђује.

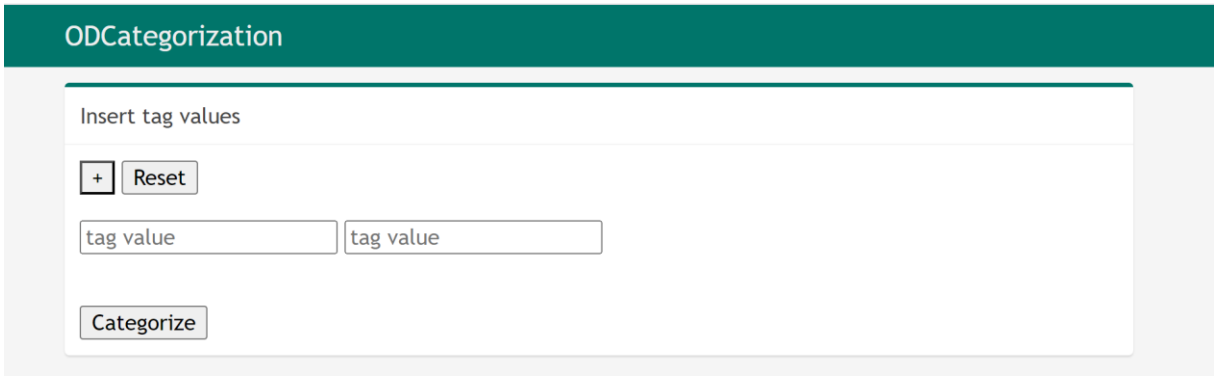
АПИ као одговор враћа објекат у JSON формату. Генерални пример одговора приказан је у наставку:

```
{
    "chosen_categories": [
           "CATEGORY_1",
           ...,
           "CATEGORY_N"
     ],
    "all_similarities": [
        {
            "categoryName": "CATEGORY_1",
            "FMK": C1_FMK,
            "NumTags": C1_NumTags,
            "avgAverageM": C1_avgAverageM,
            "maxAverageM": C1_maxAverageM,
```

```
            "maxMaximumA": C1_maxMaximumA,
        },
        ...,
        {
            "categoryName": "CATEGORY_M",
            "FMK": CM_FMK,
            "NumTags": CM_NumTags,
            "avgAverageM": CM_avgAverageM,
            "maxAverageM": CM_maxAverageM,
            "maxMaximumA": CM_maxMaximumA,
        }
    ]
}
```

Веб апликација *ODCategorization* реализована је у оквиру овог дела имплементације са циљем да омогући клијентима да из Веб претраживача испробају креирани АПИ. Клијенти уносе комбинацију тагова и као резултат добијају листу предложених категорија, као и вредности свих параметара на основу којих је категоризација извршена за сваку од доступних категорија. На овај начин, добијају се предложени резултати, али и увид у све параметре ради лакше верификације резултата.

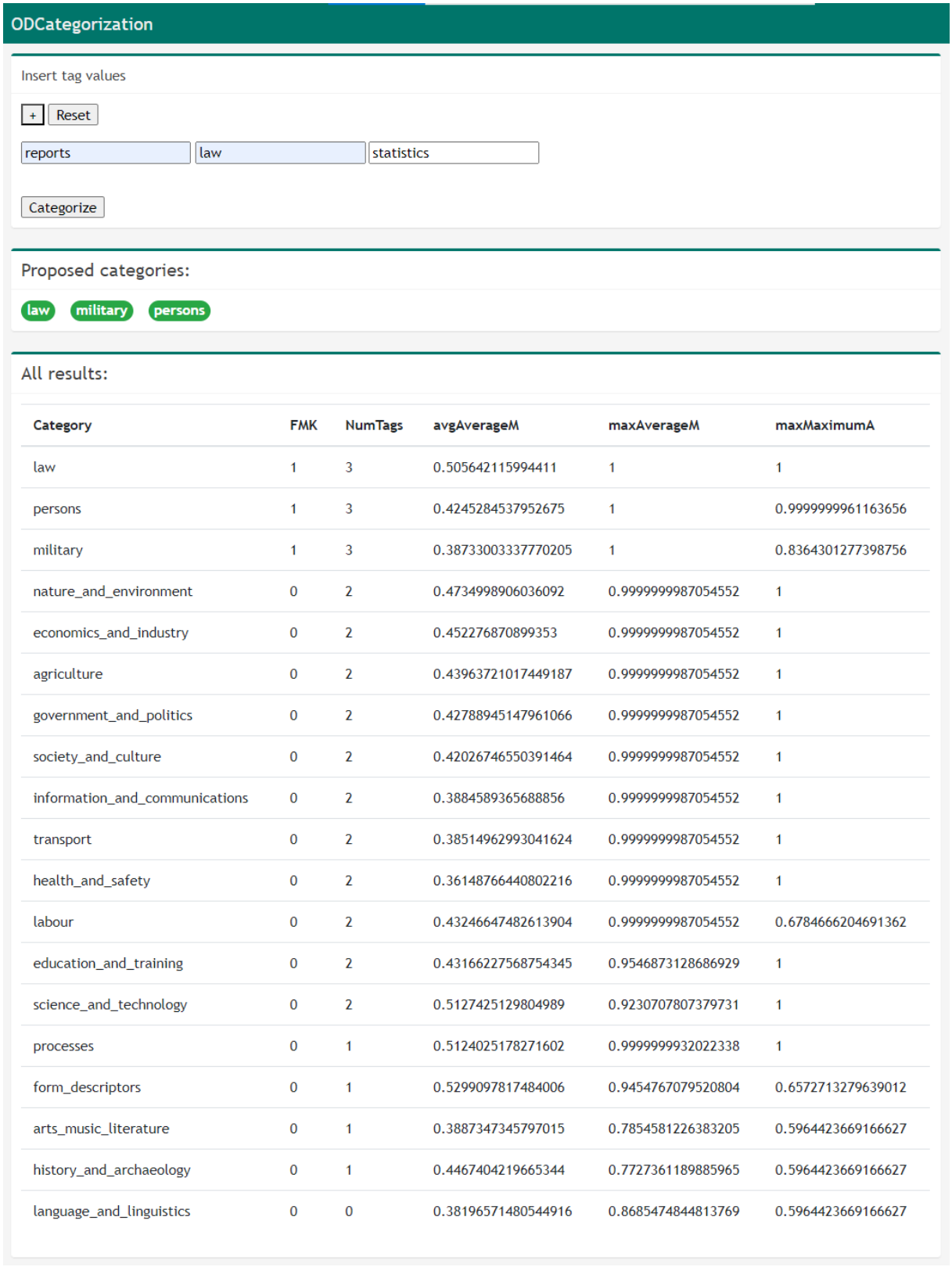
Основни поглед *ODCategorization* приказан је на слици 24.



Слика 24 – Основни поглед *ODCategorization* апликације

У овом кораку корисник уноси тагове коришћењем форме. Кликом на дугме „+“ корисник има могућност да дода нова поља за уношење вредности тагова. Кликом на дугме „Reset“ има могућност да обрише све унете вредности и крене поново са уносом. Након тога, притиском на дугме „Categorize“, на основу унетих тагова генерише се позив који се прослеђује АПИ-ју на обраду.

Резултати позива приказују се кориснику кроз 2 дела. У првом делу приказују се само називи категорија које генерише алгоритам категоризације а у другом се приказују резултати параметара за сваку од доступних категорија (Слика 25).



Category	FMK	NumTags	avgAverageM	maxAverageM	maxMaximumA
law	1	3	0.505642115994411	1	1
persons	1	3	0.4245284537952675	1	0.9999999961163656
military	1	3	0.38733003337770205	1	0.8364301277398756
nature_and_environment	0	2	0.4734998906036092	0.9999999987054552	1
economics_and_industry	0	2	0.452276870899353	0.9999999987054552	1
agriculture	0	2	0.43963721017449187	0.9999999987054552	1
government_and_politics	0	2	0.42788945147961066	0.9999999987054552	1
society_and_culture	0	2	0.42026746550391464	0.9999999987054552	1
information_and_communications	0	2	0.3884589365688856	0.9999999987054552	1
transport	0	2	0.38514962993041624	0.9999999987054552	1
health_and_safety	0	2	0.36148766440802216	0.9999999987054552	1
labour	0	2	0.43246647482613904	0.9999999987054552	0.6784666204691362
education_and_training	0	2	0.43166227568754345	0.9546873128686929	1
science_and_technology	0	2	0.5127425129804989	0.9230707807379731	1
processes	0	1	0.5124025178271602	0.9999999932022338	1
form_descriptors	0	1	0.5299097817484006	0.9454767079520804	0.6572713279639012
arts_music_literature	0	1	0.3887347345797015	0.7854581226383205	0.5964423669166627
history_and_archaeology	0	1	0.4467404219665344	0.7727361189885965	0.5964423669166627
language_and_linguistics	0	0	0.38196571480544916	0.8685474844813769	0.5964423669166627

Слика 25 – Пример приказа резултата

8. ЕВАЛУАЦИЈА РАЗВИЈЕНЕ МЕТОДОЛОГИЈЕ

У оквиру ове докторске дисертације урађена је евалуација предложеног поступка категоризације некатегорисаних сетова података. На основу анализе урађене у оквиру ове докторске дисертације, закључено је да је канадски портал отворених података [87] пример добро структуираног портала на коме је свим сетовима података била додељена најмање једна категорија као и да су сви сетови података били описани најмање једним тагом. Додатно, овај портал спада у један од већих портала по броју података у односу на портале анализиране у оквиру ове докторске дисертације. Из тог разлога, одлучено је да се подаци са канадског портала отворених података искористе за евалуацију методе за категоризацију. За потребе ове анализе, преузето је ново стање свих података са канадског портала јуна 2022. године.

Анализом преузетих података 2022. године утврђено је да квалитет података у смислу процента података којима је додељена категорија и процента података који су описани таговима није промењен у односу на стање портала из 2021. године које је детаљно приказано у овој докторској дисертацији. У овом пресеку стања са портала је преузето преко 33000 податка, који су организовани у 19 категорија. Свим подацима је додељена најмање једна категорија и један таг. На порталу је било укупно 30496 различитих тагова, док је њихов укупан број износио 243607.

Евалуација је извршена кроз 10 циклуса тестирања. Преузети подаци подељени су у 10 дисјунктних скупова података S_i, $i \in \{1,2, \dots ,10\}$, који садрже једнак број сетова података. Сваки сет S_i се користи у оквиру једног циклуса тестирања као контејнер сетова података које треба категорисати и за које се проверава тачност предложеног поступка категоризације. У сваком циклусу, подацима који се налазе у скупу који се користи за тестирање поступка, уклања се постојећа информација о категорији и они пролазе кроз поступак категоризације. За сваки циклус креира се посебна база знања која се употребљава за категоризацију на основу преосталих сетова података који се не налазе у скупу S_i. Структура базе знања, тј. мрежа концепата за све категорије за све циклусе приказана је у табелама 4 и 5 и садржи информације о укупном броју чворова за сваку од категорија (БЧ), укупном броју различитих тагова за сваку од категорија (БТ) и колико нивоа има у мрежи концепата за сваку од категорија (Ниво).

Табела 4 – Структура мрежа концепата за циклусе 1 - 5

Циклус	1			2			3			4			5		
Категорија	**БЧ**	**БТ**	**Ниво**	**БЧ**	**БТ**	**Ниво**	**БЧ**	**БТ**	**Ниво**	**БЧ**	**БТ**	**Ниво**	**БЧ**	**БТ**	**Ниво**
agriculture	590	879	9	691	978	9	699	1.036	9	764	1.070	9	731	1.007	9
arts_music_literature	48	213	4	94	306	6	96	312	6	98	319	6	88	286	6
economics_and_industry	3.969	4.085	13	4.241	4.455	13	4.303	4.631	13	4.461	4.696	13	4.214	4.258	13
education_and_training	382	710	8	394	669	8	442	807	8	461	815	8	452	775	8
form_descriptors	16.925	4.080	44	2.776	3.408	15	17.600	4.774	44	17.997	4.946	44	17.875	4.889	44
government_and_politics	2.932	3.923	13	3.233	4.217	13	3.209	4.344	13	3.117	4.333	13	2.713	3.765	13
health_and_safety	11.505	10.962	16	11.798	11.398	16	11.764	11.394	16	11.782	11.440	16	9.132	10.114	14
history_and_archaeology	13	84	2	13	85	2	14	89	2	14	89	2	11	79	2
information_and_communications	974	1.383	9	899	1.202	9	1.017	1.476	9	1.035	1.482	9	933	1.352	8
labour	688	838	12	742	866	13	819	972	13	796	966	13	815	898	13
language_and_linguistics	129	279	8	98	218	8	133	289	8	128	278	8	123	253	8
law	614	755	11	638	808	11	650	822	11	608	809	11	499	677	10
military	138	344	8	83	243	7	155	361	8	163	379	8	158	370	8
nature_and_environment	23.701	9.510	44	10.552	9.151	20	25.630	10.783	44	26.197	10.938	44	25.333	10.480	44
persons	1.066	1.014	14	1.072	967	14	1.148	1.058	14	927	1.027	13	1.000	868	13
processes	203	484	7	226	525	8	233	535	8	233	545	8	162	448	7
science_and_technology	18.586	5.701	44	4.631	4.984	15	19.540	6.593	44	20.034	6.748	44	19.676	6.475	44
society_and_culture	2.181	2.841	13	2.612	2.986	14	2.521	3.478	15	2.837	3.629	15	2.374	3.198	15
transport	497	1.078	9	716	1.249	14	662	1.286	16	771	1.382	16	722	1.238	16

Табела 5 – Структура мрежа концепата за циклусе 6 - 10

Циклус	6			7			8			9			10		
Категорија	**БЧ**	**БТ**	**Ниво**	**БЧ**	**БТ**	**Ниво**	**БЧ**	**БТ**	**Ниво**	**БЧ**	**БТ**	**Ниво**	**БЧ**	**БТ**	**Ниво**
agriculture	743	1.058	9	698	1.017	9	746	999	9	594	847	9	779	1.054	9
arts_music_literature	98	319	6	88	296	6	87	272	6	85	291	6	96	311	6
economics_and_industry	4.100	4.757	13	4.347	4.726	13	4.059	4.140	13	4.144	4.195	13	4.571	4.690	13
education_and_training	390	791	6	441	815	8	465	768	8	428	723	8	479	840	8
form_descriptors	18.182	5.063	44	18.036	5.015	44	17.740	4.528	44	18.122	5.007	44	18.166	5.056	44
government_and_politics	3.376	4.520	11	3.117	4.340	12	2.783	3.788	13	3.171	4.185	13	3.446	4.525	13
health_and_safety	8.948	10.760	12	7.719	8.604	16	9.188	9.266	16	9.171	9.878	16	11.863	11.443	16
history_and_archaeology	14	89	2	13	73	2	9	40	2	13	84	2	14	89	2
information_and_communications	1.044	1.529	9	1.031	1.478	9	742	1.248	9	1.004	1.438	9	1.082	1.510	9
labour	604	945	9	724	937	12	796	888	13	778	867	13	835	974	13
language_and_linguistics	92	274	7	120	270	8	127	240	8	123	283	8	134	290	8
law	649	825	11	514	744	9	550	676	11	605	762	11	661	834	11
military	164	382	8	159	356	8	132	279	8	161	379	8	164	382	8
nature_and_environment	26.592	11.156	44	25.214	10.563	44	24.671	9.839	44	24.469	9.946	44	26.708	11.175	44
persons	907	1.008	14	1.046	1.018	14	1.091	913	14	1.109	1.003	14	1.155	1.044	14
processes	235	545	8	226	519	8	156	398	7	207	500	8	236	553	8
science_and_technology	20.057	6.826	44	19.817	6.686	44	19.085	5.993	44	19.833	6.625	44	20.184	6.850	44
society_and_culture	2.609	3.564	15	2.783	3.566	15	2.662	3.334	15	2.526	3.331	15	2.903	3.654	15
transport	751	1.350	16	759	1.353	16	680	1.180	16	748	1.291	16	786	1.361	16

Из датог приказа, може се приметити да сложеност мрежа концепата није иста за све категорије, већ да је у неким категоријама знатно већи број чворова и тагова у односу на преостале. Тако, у највећем броју циклуса, категорије *nature_and_environment, science_and_technology* и *form_descriptors* имају много већи број чворова као и број нивоа у односи на преостале категорије. Изузетак је циклус два, у коме по броју чворова доминирају категорије *nature_and_environment* и *health_and_safety*. Такође, мреже концепата се разликују и по укупном броју различитих тагова. У највећем броју циклуса по овом параметру издвајају се категорије *health_and_safety* и *nature_and_environment.* Додатно, може да се примети да су мреже концепата категорија попут *history_and_archaeology* и *arts_music_literature* много једноставније у односу на остале категорије. Овакве разлике у структурама мрежа концепата су последица разлика у броју сетова података као и броју различитих комбинација које се користе за описивање података који припадају овим категоријама.

Последично, средња вредност укупног броја нивоа мрежа концепата за све категорије у оквире једног циклуса је од 11 до 13, просечан број различитих тагова за све мреже концепата по циклусима је у опсегу од 2564 до 2981, док је просечан број чворова у мрежама концепата за све категорије у оквиру једног циклуса у опсегу од 2395 до 4961.

На основу ове базе знања извршена је евалуација предложеног поступка категоризације а опис комплетног просеца евалуације приказан је у наставку:

За сваки циклус $i, i \in \{1,2,\dots,10\}$ важи:

1. креирати скуп података за тестирање S_t, $S_t = S_i$ – скуп који садржи податке које треба категорисати
2. сваком податку у скупу S_t уклонити информацију о категорији
3. креирати скуп SU_t као $SU_t = \bigcup_{k=1}^{10} S_k,\ i \neq k$ - скуп који садржи све податке који се не налазе у тест скупу података, представља основу за базу знања на основу које се ради категоризација
4. креирати празан скуп мрежа концепата $CL_i = \{\}$
5. за сваку категорију на порталу Cat_j:
   5.1. креирати скуп података DS_j као подскуп SU_t који садржи само податке који припадају категорији Cat_j
   5.2. креирати формални контекст FC_j на основу метаподатака скупа DS_j
   5.3. за формални контекст FC_j креирати мрежу концепата L_j

5.4. додати L_j у скуп CL_i

6. сваки податак из скупа S_t категорисати коришћењем базе знања CL_i

У оквиру последњег корака (тачка 6) ради се категоризација сваког податка коришћењем поступка представљеног у овој докторској дисертацији. Након категоризације свих сетова података урађена је провера тачности добијених резултата по циклусима. За сваки податак из скупа S_i посматран је однос категорија којима податак заиста припада и предложених категорија. Резултати евалуације по циклусима приказани су у табели 6.

У приказаним резултатима за сваки циклус дат је преглед:

- броја података за које је рађена категоризација (колона Број података),
- процента потпуно категорисаних података (колона Потпуно (%)) – проценат података којима су додељене све категорије којима податак заиста припада и није додељена ни једна додатна категорија,
- процента потпуно категорисаних података са додатним категоријама (колона Потпуно + (%)) – проценат података којима су додељене све категорије којима податак заиста припада али и најмање једна додатна категорија за коју није дефинисано да податак припада,
- процента делимично категорисаних података (колона Делимично (%)) – проценат података којима је додељен део категорија којима податак заиста припада и ни једна категорија којој податак не припада,
- процента делимично категорисаних података са додатним категоријама (колона Делимично + (%)) – проценат података којима је додељен део категорија којима податак заиста припада и најмање једна додатна категорија којој податак не припада,
- проценат лоше категорисаних података (колона Погрешно (%)) – проценат података којима није додељена ни једна категорија којој податак припада.

Табела 6 – Преглед резултата евалуације

Циклус	Број података	Потпуно (%)	Потпуно+ (%)	Делимично (%)	Делимично+ (%)	Погрешно (%)
1	3320	51,17	20,60	8,64	2,86	16,72
2	3320	28,31	11,87	47,65	2,80	9,37
3	3320	50,84	32,17	4,01	3,73	9,25
4	3320	74,67	16,81	2,20	0,90	5,42
5	3320	70,75	13,34	6,45	2,29	7,17
6	3320	74,04	16,45	4,58	2,71	2,23
7	3320	72,92	15,36	5,57	1,57	4,58
8	3320	59,91	16,14	10,24	2,89	10,81
9	3320	76,42	12,23	3,98	1,69	5,69
10	3320	98,16	0,87	0,06	0,09	0,81

Сваки од циклуса евалуације садржао је 3320 података над којима је урађена категоризација. Из приказаних резултата може да се примети да проценат потпуно категорисаних података варира од 28,31% у другом циклусу до 98,16% у последњем циклусу, те је средња вредност потпуно категорисаних података 71,84%. Додатно, проценат делимично категорисаних података има вредност у опсегу од 0,06% у последњем циклусу до 47,65% у другом циклусу тестирања. Уколико се проценти потпуно категорисаних података посматрају заједно са процентима делимично категорисаних података, може да се примети да је овај збир у свим циклусима већи од 50%. Најнижу вредност има у циклусима један и три у којима збир има вредност 59,82% и 54,85%. У осталим циклусима овај збир има вредност већу од 70%, а највишу вредност има у циклусу десет 98,22%. Последично, средња вредност овог параметра је 77,03% а просечна 75,06%.

Са друге стране, проценат података којима није додељена ниједна исправна категорија износи просечно 7,2% за све циклусе, док њена средња вредност износи 6,43%. Овај параметар има највећу вредност у првом циклусу 16,72%, након тога у осмом циклусу где износи 10,81%, док је у осталим циклусима испод 10%. Најмање погрешно категорисаних података има у последњем циклусу 0,81%.

Међутим, из приказаних резултата може да се примети да је проценат података којима је поред свих тачних категорија додељена још барем једна додатна у свим циклусима, осим трећег и десетог између 11% и 21%. У десетом циклусу услед изузетно доброг процента потпуно категорисаних података овај проценат износи

0,87%, док је у трећем циклусу он највећи и износи 32,17%. Последично, просечна вредност овог параметра за све циклусе износи 15,58%. Проценти делимично категорисаних података којима је додата још барем једна додатна категорија у свим циклусима, имају вредност испод 4%. Додатно, просечна вредност овог параметра за све циклусе је 2,15%, а средња вредност износи 2.5%.

Такође, може се приметити и да су осим у циклусу два, проценти потпуно категорисаних података и потпуно категорисаних података са додатним категоријама, значајно већи у односу на проценте делимично категорисаних података као и делимично категорисаних података са додатним категоријама. Средња вредност збира процената потпуно категорисаних података и потпуно категорисаних података са додатним категоријама за све циклусе износи 86,19%. Гледано по циклусима, само у три циклуса – првом, другом и осмом овај параметар има вредност мању од 83%. У три циклуса овај проценат има вредност преко 90%, при чему у десетом циклусу има вредност 99.04%.

Уколико се циклуси посматрају појединачно може да се примети да је циклус са најбољим резултатима евалуације циклус десет, код кога је 98,16% података потпуно категорисано а проценат лоше категорисаних износи 0,81%. Са друге стране, циклуси један и три имају најслабије резултате. Циклус један, има највећи проценат погрешно категорисаних података и укупно 59,82% података који су потпуно и делимично категорисани без додатних категорија. Циклус три, има укупно 54,85% података који су потпуно и делимично категорисани без додатних категорија, што га чини циклусом са најнижим процентом тачно додељених категорија којима није придружена ни једна додатна категорија. Међутим, у овом циклусу забележен је највећи проценат података којима су додељене све потребне категорије али и нека додатна – 32,17%.

Циклус два, има прилично низак проценат комплетно категорисаних података али и 47,65% делимично категорисаних чиме овај циклус има укупно 75,96% података са добро додељеним категоријама без додатних категорија. Последично, овај циклус има сличне укупне проценте потпуно и делимично категорисаних података без додатних категорија са осталим циклусима.

Додатно су анализиране категорије које су додаване потпуно категорисаним подацима са додатним категоријама и делимично категорисаним подацима са додатним категоријама. За ова два случаја категоризације, у табели 7 приказан је однос између категорија на канадском порталу отворених података и колико често су оне додељиване

сетовима података као категорије а да притом подаци не припадају тим категоријама. Резултати приказани у табели представљају проценте у односу на укупан број свих додатих категорија којима податак не припада а које су додељене подацима. У табеларном приказу колона *Потпуно +* приказује информације за потпуно категорисане тест податке са додатним категоријама, док колона *Делимично +* приказује информације за делимично категорисане тест податке са додатним категоријама.

Табела 7 – Преглед додавања додатних категорија сетовима података

Категорија	**Потпуно + Придружене категорије (%)**	**Делимично + Придружене категорије (%)**
agriculture	2.72	1.52
arts_music_literature	0.20	0.16
economics_and_industry	15.20	15.24
education_and_training	2.30	2.95
form_descriptors	11.06	6.86
government_and_politics	9.67	8.38
health_and_safety	6.20	10.38
history_and_archaeology	0.10	0.00
information_and_communications	2.83	3.99
labour	2.41	8.62
language_and_linguistics	0.29	0.40
law	1.02	2.00
military	0.35	0.56
nature_and_environment	11.38	11.33
persons	3.65	3.67
processes	1.14	2.39
science_and_technology	11.67	6.78
society_and_culture	12.98	11.81
transport	4.83	2.95

Из приказаног прегледа може да се примети је да су неке од категорија чешће придруживане сетовима података. Уколико се посматрају подаци који су потпуно категорисани али им је додата још најмање једна додатна категорија најчешће је придруживана категорија *economics_and_industry* са 15.2%. Након ове категорије следе категорије *society_and_culture*, *science_and_technology*, *form_descriptors* па

nature_and_environment код којих је проценат већи од 10%. Остале категорије су ређе придруживане, при чему је у 12 категорија тај проценат мањи од 5%, док је у четири категорије проценат мањи од 1%.

Слични резултати добијени су и за сетове података који су делимично категорисани а којима је додељена додатна категорија. И овим подацима је најчешће додавана категорија *economics_and_industry,* након ње следе категорије *society_and_culture, nature_and_environment* и *health_and_safety* чији проценти прелазе 10%. Код осталих категорија проценти су испод 10%, при чему 11 категорија има проценат испод 5% а четири категорије испод 1%.

Додатно су анализирани сетови података који су потпуно категорисани али им је додата још најмање једна категорија због процента података који припада овој групи. Међу овим подацима примећено је да се јављају и сетови података описани само једним тагом. Посматрано процентуално по циклусима, проценат ових података се креће од 0,18% у шестом циклусу до 21,68% у четвртом циклусу, те је просечан број ових података за све циклусе 7,14%. Када се посматрају тагови који се користе за описивање ових података, најчешће су се јављали тагови „geographical maps“, „census of population“, „Oceans“, „Topography“, „Fees Report“, „child support“, „Health“, „Parks“, „HESA“, „Annual report“, „Audit“ и „infographics“. Поред ових вредност појављују се и друге вредност попут „Regions“, „Map“, „plan“, „Report“, „Annual report“, „service“, „Statistics“ и „survey“ које се јављају у већем броју категорија у различитим комбинацијама. Ове вредности се могу сматрати генералнијим вредностима тагова пошто се јављају у већем броју категорија, те се појављују у мрежама концепата већег броја категорија.

Поред анализе података који су описани само једном категоријом, примећено је да има случајева када се јављају и подаци описани комбинацијом тагова која се појављује у подацима који су међусобно различито категорисани.

На пример, податак означен таговима „ferries“ и „road transport“ који припада категорији *government_and_politics*, алгоритмом је придружен категоријама *government_and_politics, form_descriptors, nature_and_environment* и *transport* пошто постоји податак описан овом комбинацијом тагова а која је заиста означена овим категоријама. Такође је и податак описан таговима „inland waters“ и „lakes“ који припада категорији *government_and_politics,* алгоритмом категоризован у категорије *government_and_politics, form_descriptors, nature_and_environment* и *transport* пошто

заиста постоји податак описан том комбинацијом тагова а који припада свим овим категоријама.

Примећено је да има ситуација да се нека комбинација тагова јавља и у више од две различите комбинације категорија. Тако податак, описан таговима „Department of Justice“, „Access to Information“, „Canada's System of Justice“, „Justice Canada Publications“, „Charter of Rights and Freedoms“, „Charter Statements“, „42nd Parliament: 1st Session“ за који се очекивало да буде додељен категоријама *government_and_politics, law, persons* и *society_and_culture*, додељен је категоријама *government_and_politics, society_and_culture, persons, law* и *processes* зато што постоје подаци са овом комбинацијом тагова који су описани категоријама *government_and_politics, law, persons, processes* и *government_and_politics*, али и комбинацијом категорија *law, persons, processes* и *society_and_culture.* У табели која се налази у наставку (Табела 8), дато је још неколико примера комбинација тагова које су таговане различитим комбинацијама категорија.

Табела 8 – Примери комбинација тагова који припадају различитим комбинацијама категорија

Комбинације тагова	**Комбинације категорија**
"Prince Edward Island data", "Government information"	- society_and_culture - nature_and_environment - transport
"PSC", "Public Service Commission", "Advertisements", "Applications", "Applicants", "Appointments"	- government_and_politics - government_and_politics, persons, processes
"Nature and Biodiversity - Habitat", "Species", "Protect Species Well-Being", "Protect and Restore Species", "National (CA)", "Habitats"	- form_descriptors, nature_and_environment, science_and_technology - nature_and_environment, science_and_technology'
"OGIP", "Directive on Open Government"	- agriculture, economics_and_industry, information_and_communications - information_and_communications - government_and_politics, information_and_communications
"Plan", "Evaluation", "Fiscal"	- economics_and_industry, government_and_politics,

	science_and_technology - economics_and_industry, government_and_politics
"Mobility", "Wireless"	- information_and_communications, science_and_technology - education_and_training, information_and_communications, science_and_technology
"Transition", "Proactive Publication", "C-58", "Access to Information", "ATI"	- government_and_politics - government_and_politics, information_and_communications

У последњем делу евалуације резултата категоризације, анализирани су подаци који су погрешно категорисани, односно подаци којима није додељена ниједна адекватна категорија. Када се посматрају ови резултати може да се примети да је просечно 6,62% категорисаних података у свим циклусима евалуације било описано само једним тагом, док је у осталим случајевима коришћено више тагова за описивање података. Примери неких од тагова који су самостално описивали погрешно класификоване податке су следећи: „open government“, „statistics“, „age“, „Nova Scotia Open Data“, „Web Access“, „Elections“, „Survey“, „work“, „Privacy“, „fees report“, „Annual Report“, „Cloud Security“, „boundaries“, „infographics“, „manufacturing industry“, „railway networks“, „photos“, „Soil“, „Prince Edward Island data“, „Activity“, „Library“, „video“ и други.

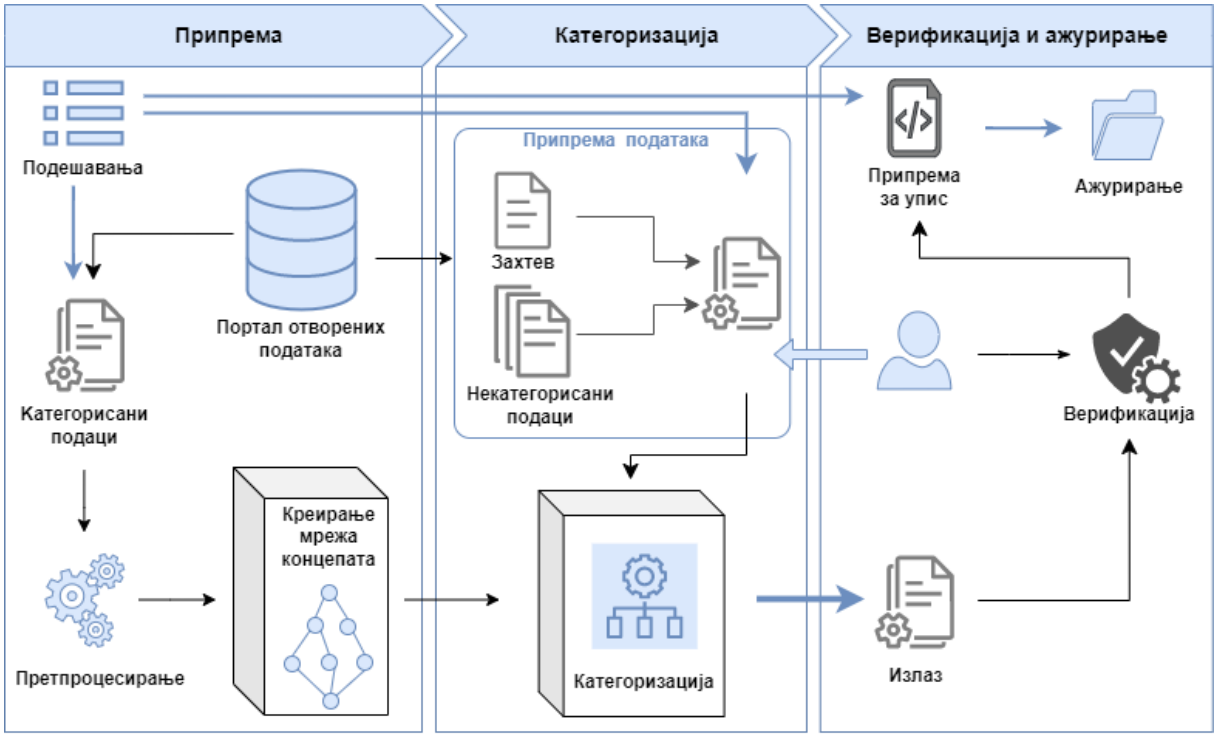
Када се посматра којим категоријама су припадали сетови података који су погрешно категорисани, може да се примети да су најчешће изостављане категорије *society_and_culture* и *government_and_politics* које су изостављене у 15,96% и 14,67% случајева. Код осталих категорија овај проценат је испод 10%, а најмање износи код категорија *processes, language_and_linguistics* и *history_and_archaeology* код којих је проценат испод 1%. Детаљан преглед, процентуално у колико случајева је која категорија изостављена код података који су погрешно категорисани, приказан је у табели 9.

Табела 9 – Процентуални преглед недодељених категорија

Категорија	**%**
agriculture	4.47
arts_music_literature	1.32
economics_and_industry	9.83
education_and_training	3.61
form_descriptors	4.11
government_and_politics	14.67
health_and_safety	6.46
history_and_archaeology	0.03
information_and_communications	6.66
labour	2.15
language_and_linguistics	0.40
law	1.92
military	1.99
nature_and_environment	7.12
persons	4.30
processes	0.96
science_and_technology	7.72
society_and_culture	15.96
transport	6.32

9. ПРЕДЛОГ МОДЕЛА ЗА ДОПУНУ МЕТАПОДАТАКА ДОКУМЕНАТА НА ПОРТАЛИМА ОТВОРЕНИХ ПОДАТАКА

Квалитет метаподатака игра кључну улогу за успех отворених података [4]. Бројна истраживања која су објављена сугеришу да је квалитет и потпуност метаподатака јако битна. Из тог разлога, у оквиру овог поглавља биће предложен модел за допуну метаподатака, са акцентом на допуни информација о категоријама којима некатегорисани подаци треба да припадају. Модел који је представљен у овом поглављу приказује како методологија приказана у седмом поглављу ове докторске дисертације може да се искористи за допуну метаподатака докумената на порталима отворених података. Архитектура модела приказана је на слици 26.



Слика 26 – Модел за допуну метаподатака на порталима отворених података

Дефинисан модел може да се прилагоди сваком порталу отворених података и састоји се из три целине:

- Припрема система за категоризацију
- Примена категоризације
- Верификација и ажурирање промена

У оквиру модела предвиђено је да човек користи систем на два начина:

- Да класификује одређени податак на захтев, прослеђивањем јединственог идентификатора податка
- Да периодично класификације скуп некатегорисаних података који се налазе на порталу.

9.1. Припрема система за категоризацију

Сваки портал отворених података може да поседује своју шему која дефинише структуру метаподатака који се памте за сваки сет података. Из тог разлога, на почетку је потребно дефинисати подешавања за модел за категоризацију, који ће садржати информације о шеми метаподатака и дефинисане мета-кључеве који представљају тагове и категорије у метаподацима. Поред тога, иако није често, неки портали отворених података подржавају само једну категорију којој податак може да припада. Из тог разлога, информација о броју категорија коју је могуће доделити једном податку треба да буде део подешавања, како би се у тим ситуацијама предлагала само једна, најадекватнија категорија.

Поред дефинисања подешавања потребно је урадити припрему базе знања на коју се ослања категоризација. Предложени модел извршава категоризацију података на основу информација које се налазе на том порталу, те захтева припрему података који ће бити употребљени за креирање базе знања. Из тог разлога, издвајају се сви сетови података којима су додељене категорије и који су описани таговима. На овај начин извршава се филтрирање сетова података чији метаподаци нису довољно дефинисани да би могли да буду део припреме система за категоризацију.

На основу дефинисаних подешавања система за категоризацију, издвојени подаци се пребацују у облик погодан за креирање базе знања. Од потребних информација о сетовима података памте се информације о јединственом идентификатору, таговима и категоријама.

Након издвајања релевантних сетова података, приступа се претпроцесирању које подразумева обраду тагова којима су сетови података описани у следећем редоследу:

- уклањање чланова и знакова интерпункције уколико су они део тагова.
- за сваки сет података извлачи се скуп јединствених вредности тагова како би се избегла дуплирања тагова која су могла да се јаве као резултат њихове обраде.

Сетови података који се добију као излаз ове обраде постају улаз за креирање мрежа концепата за сваку од дефинисаних категорија на порталу отворених података. Креирањем мрежа концепата за сваку од категорија постављена је база знања на којој се заснива категоризација.

Успех категоризације зависи од формалног контекста који постоји за сваку од категорија, што значи да зависи од постојећих тагова и њихових комбинација. Тагове дефинише корисник система, и као што је објашњено у претходном делу докторске дисертације, различити корисници могу да податке означавају различитим вредностима које међусобно могу да буду сличне, али и не морају. Како би се омогућио што бољи рад овако дефинисаног модела за допуну метаподатака докумената, потребно је периодично ажурирање базе знања у зависности од промена које се дешавају на порталу отворених података. Као што се може приметити из анализе употребе тагова на порталима која је приказана у претходном делу ове дисертације, на неким порталима и даље има великих промена које утичу на скуп тагова који се јавља на порталима. Такође, може се приметити да има и портала на којима нема великих промена, већ да се само додају нове вредности.

Из тог разлога, ажурирање мрежа концепата за категорије није потребно извршавати након сваког појединачног новог уписа или категоризације података, већ у одређеним временским интервалима. На порталима са великом количином података, на којима нема великих промена, попут додавања велике количине нових података, ово ажурирање може ређе да се извршава. Међутим, на малим порталима, порталима са малим бројем података, ажурирање би требало чешће извршавати. У малим мрежама концепата, формираним над малим формалним контекстом, већа је вероватноћа да ће проширивање формалног контекста довести до додавања нових термина који имају малу сличност са постојећим терминима или променити изглед постојеће хијерархије. Самим тим, овакво проширење повећава шансу да категоризација за ту категорију буде прецизнија.

Додатно, након додавања велике количине података, попут додавања података из новог извора података, креирање нових мрежа концепата за све категорије може да повећа прецизност категоризације. Додавањем података из новог извора на портал, значи и додавање тагова дефинисаних од стране корисника који су нови за портал отворених података а који податке можда означавају на другачији начин и другачијим комбинацијама вредности. Из тог разлога, након таквих акција пожељно је ново

креирање мрежа концепата. Креирање нове базе знања треба урадити и у ситуацијама реорганизације постојећих информација на порталу или додавања нових категорија.

9.2. Примена категоризације

Део модула за категоризацију се састоји од две целине: дела за припрему података које треба категоризовати и дела за категоризацију података.

У зависности од начина употребе система, односно да ли корисник категоризује један одређени податак или све податке који нису додељени категоријама врши се припрема података које је потребно категоризовати. У случају појединачне категоризације, на основу јединственог идентификатора прибавља се тражени податак, извршава се претпроцесирање његових тагова и креира се објекат који се прослеђује модулу за категоризацију а који садржи јединствени идентификатор и припремљену комбинацију тагова која описује жељени податак. У случају категоризације свих некатегоризованих података, преузимају се подаци које је потребно категоризовати, за сваки од података извршава се претпроцесирање тагова и креира се листа улазних објеката у којој се сваки податак представља његовим јединственим идентификатором и скупом припремљених тагова. У оба случаја претпроцесирање тагова подразумева, као и код креирања базе знања, уклањање чланова и знакова интерпункције а затим креирање листе јединствених вредности за улазне тагове.

Овако припремљени подаци прослеђују се моделу за категоризацију који коришћењем алгоритма за категоризацију представљеног у овој докторској дисертацији, за сваки од прослеђених података генерише објекат који садржи идентификатор податка, листу тагова и листу категорија које је алгоритам предложио. У случају да портал отворених података дозвољава да податак припада само једној категорији, потребно је прилагодити модул за категоризацију. Прилагођење подразумева да се у том случају омогући да модул предложи само једну категорију са најбољим резултатима уместо све категорије које задовољавају критеријуме сличности. Излаз из овог модула прослеђује се делу за верификацију и ажурирање.

9.3. Верификација и ажурирање промена

У последњем кораку овог модела извршава се верификација и уписивање резултата на портал отворених података. Након креираних резултата категоризације кориснику се приказују добијени резултати. Корисник за сваки од података врши верификацију и прослеђује модулу за *Припрему за упис*.

Модул за *Припрему за упис* је модул који је потребно прилагодити сваком порталу посебно, и његова имплементација зависи од платформе коју портал отворених података користи, тачне инсталације која је постављена, као и програмског интерфејса који има на располагању. У зависности од доступних опција, креирају се наредбе за ажурирање постојећих података на порталу. На основу јединственог идентификатора сваког податка креира се наредба за ажурирање метаподатака, односно вредности за мета-кључ који представља категорије у који се уписују предложене вредности. Након извршене припреме, врши се упис информација о категоријама на портал отворених података.

10. ДИСКУСИЈА И ЗАКЉУЧАК

Објављивање отворених података је глобални тренд усклађен са стратегијама и акционим плановима које усвајају владе широм света, као средство за већу транспарентност и ефикасност јавних управа. Међутим, услед великог пораста броја података, све је теже претраживање података и добијање тражених информација. Права вредност објављених података лежи у њиховом даљем коришћењу и могућности да се из њих генеришу нова знања. Из тог разлога, како би олакшали доступност података, портали отворених података су увели различите механизме претраге података, као што су по категорији, типу, таговима, организацији и слично. Међутим, непотпуност описа података може да утиче на квалитет резултата претраге доступних података и да доведе до смањења видљивости података.

У оквиру ове докторске дисертације адресиран је проблем непотпуних информација о категорији којој подаци припадају на порталима отворених података и понуђена је методологија за категоризацију података на порталима отворених података. Приступ приказан у овој докторској дисертацији се заснива на ФЦА алгоритму, при чему се формални контекст креира на основу постојећих података и вредности мета-кључева који представљају тагове у метаподацима података. На основу формалног контекста се креирају мреже концепата, које осликавају начин употребе тагова по категоријама, и које су основа за приказани поступак категоризације.

У оквиру докторске дисертације приказан је и интерактивни алат за визуализацију и анализу креираних мрежа концепата, а који решава познате проблеме визуализације сложених мрежа концепата и даје јасан преглед карактеристика мреже и самих концепата.

Приказани поступак категоризације за унету комбинацију тагова рачуна сличност са свим мрежама концепата, и приказани поступак предлаже једну или више категорија којој податак описан неком комбинацијом тагова треба да припада.

У оквиру ове докторске дисертације, за поступак категоризације урађена је евалуација коришћењем података са канадског портала отворених података и добијени резултати имају добар проценат успешности. На основу ових резултата, предложен је модел за допуну метаподатака сетова података на порталима отворених података.

Имплементирани поступак категоризације евалуиран је коришћењем података са канадског портала отворених података кроз 10 циклуса. Анализом резултата евалуације

може се приметити да приказани поступак има средњу вредност успешности категорисања од 77,03% (потпуно и делимично категорисани подаци), док је средња вредност погрешно категорисаних података 6,43%. Ови проценти указују да приказани поступак категоризације даје добре резултате категоризације, посебно ако се узме у обзир и проценат потпуно категорисаних података али са додељеним и додатним категоријама. У том случају, средња вредност за све циклусе потпуно категорисаних података и потпуно категорисаних података са додатним категоријама износи 86,19%.

Оно што се може додатно приметити из резултата евалуације, јесте да су најчешће додатно придруживане категорије оне које су уједно и највеће категорије доступне на порталу. Ови резултати нису изненађујући, пошто се ради о великим категоријама којима припада највише различитих тагова и комбинација тагова. Последично, те категорије садрже и много генералних тагова, као и тагова који се јављају у више категорија.

Анализом резултата је примећено и да постоји неконзистентност у таговању података на порталима отворених података, односно да се неки подаци који припадају различитим скуповима категорија означавају истим вредностима тагова. Последично, представљени поступак за такве комбинације тагова може да предложи шири списак категорија од оног који је очекиван.

Међутим, алгоритам је показао непрецизности када се ради о комбинацијама тагова у којима се јављају често употребљаване вредности, као и у неким ситуацијама када се за описивање податка користе генералније вредности.

Када се посматрају сетови података којима није додељена ниједна адекватна категорија, алгоритам је најчешће изостављао категорије *society_and_culture* и *government_and_politics*, и то у 30.63% случајева, док су све остале категорије ређе изостављане. Додатно, категорија *history_and_archaeology* се јавила само једанпут међу подацима којима није додељена ниједан адекватна категорија, а категорија *language_and_linguistics* 12 пута. Последично, ове категорије су најмање изостављане.

Код предложеног поступка категоризације, треба узети у обзир да се она базира на бази знања која представља хијерархију употребе тагова, те квалитет саме категоризације зависи од квалитета мрежа концепата. Додатно, портали отворених података се константно мењају и додају се нови подаци. Самим тим, мењају се и скупови комбинација тагова који се користе за описивање података у оквиру категорија.

Последично, потребно је извршити ажурирање базе знања у временским интервалима који прате промене на порталима, како би се очувала прецизност категоризације. Међутим, због карактеристика алгоритма, ово ажурирање није потребно радити након сваког новог унетог податка. Представљени алгоритам рачуна сличности између тагова и комбинација тагова, па уношење нових података сличних описа неће значајно утицати на прецизност. За нове сетове података, који су садржајно другачији од података који већ постоје на порталима, може се очекивати да су описани потпуно другачијим таговима и комбинацијама тагова. За ове тагове се може претпоставити да немају велику сличност са постојећим таговима у мрежама концепата. Из тог разлога, у ситуацији када се јави унос већег броја оваквих података, као што је додавање података из неких нових извора података, потребно је извршити креирање нових мрежа концепата за сваку од доступних категорија. Такође, у случају додавања нових категорија, или поделе постојећих категорија у више, потребно је урадити креирање нових мрежа концепата.

10.1. Правци даљег истраживања

Приказани поступак категоризације пружа могућности за даља унапређења, посебно у области креирања хијерархија тагова. Тренутно се за креирање хијерархије тагова користи итеративни *NextClosure* алгоритам. При раду са великом количином података и комбинацијама тагова, извршавање овог алгоритама може да буде временски захтевно, због карактеристика алгоритма. Из тог разлога, алгоритам би могао да буде замењен неким оптималнијим ФЦА алгоритмом, који би био временски ефикаснији.

Додатно, даље истраживање може бити усмерено у правцу претпроцесирања података. Тренутно се у оквиру методологије, пре креирања хијерархија, не ради претпроцесирање података, већ се хијерархије креирају на основу свих комбинација тагова које се користе у свом основном облику. У будућности, могло би се искористити претпроцесирање тагова које би могло да има два правца.

Један правац претпроцесирања тагова подразумевао би креирање методе која би детектовала тагова који нису значајни за категоризацију. Након тога, ова метода би извршила њихово уклањање из података, као и раздвајање делова тагова на више тагова у случају за тим има потребе, као у случају тагова чија вредност садржи набрајање независних термина.

Други правац би био претпроцесирање тагова у циљу смањења комплексности мрежа концепата. Како би се мреже концепата поједноставиле, требало би урадити редукцију тагова која се може постићи заменом сличних тагова једном вредношћу. На овај начин, мањи број вредности тагова би постојао у мрежи, што би последично довело до смањења њене комплексности, а самим тим и времена за њено креирање.

На крају, у оквиру методе категоризације може да се проба увођење нових параметара на основу којих се ради одабир категорија како би се побољшао квалитет предложеног алгоритма.

ЛИТЕРАТУРА

[1] Judie Attard, Fabrizio Orlandi, Simon Scerri, and Sören Auer. A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4), 399-418, 2015. doi: 10.1016/j.giq.2015.07.006

[2] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Automated quality assessment of metadata across open data portals. *Journal of Data and Information quality*, vol. 8, no.1, pp. 2:1-2:29, 2016.

[3] Sander van der Waal, Krzysztof Węcel, Ivan Ermilov, Valentina Janev, Uroš Milošević, and Mark Wainwright. Lifting open data portals to the data web. *In Linked Open Data--Creating Knowledge Out of Interlinked Data*, Lecture Notes in Computer Science, vol 8661 Springer, Cham, pp. 175-195, 2014.

[4] Sylvain Kubler, Jérérmy Robert, Sebastian Neumaier, Jürgen Umbrich, and Yves Le Traon. Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, vol. 35, no.1, pp.13-29, 2018.

[5] The Open Definition. Доступно на: http://opendefinition.org/ Приступљено: марта 2022.

[6] Open Knowledge Foundation: Open data handbook. Приступљено: марта 2022. Доступно на: http://opendatahandbook.org/guide/en/what-is-open-data/

[7] The Annotated 8 Principles of Open Government Data, Приступљено: новембра 2021. Доступно на: https://opengovdata.org/

[8] Закон о електронској управи Републике Србије, Службени гласник РС, https://www.ite.gov.rs/extfile/sr/2983/Zakon_o_elektronskoj_upravi_c.pdf#page=3

[9] Jan Kučera, Dušan Chlapek, and Martin Nečaský. Open Government Data Catalogs: Current Approaches and Quality Perspective. *In: Technology-Enabled Innovation for Democracy, Government and Governance. EGOVIS/EDEM 2013. Lecture Notes in Computer Science*, vol. 8061. Springer, Berlin, Heidelberg, 2013.

[10] Jorn Berends, Wendy Carrara, Cosmina Radu. Analytical Report 9: The Economic Benefits of Open Data, EU publications, 2017.

[11] European Commission, Directorate-General for the Information Society and Media, Wendy Carrara, Sander Fischer, Eva van Steenbergen, Wea San Chan. Creating

Value through Open Data: Study on the Impact of Re-use of Public Data Resources. EU Publications, 2015; doi: 10.2759/328101

[12] European Data Market Study, SMART 2013/0063, Final Report, IDC, Open Evidence, 2017

[13] European Commission, Directorate-General for Communications Networks, Content and Technology, Innesti, Barbero M, Bartz K, Linz F et al. Study to support the review of directive 2003/98/EC on the re-use of public sector information. 2018. doi:10.2759/ 373622.

[14] Ahmad Assaf, Raphaël Troncy, and Aline Senart. HDL-Towards a Harmonized Dataset Model for Open Data Portals. *In Usewod-profiles@ ESWC*, pp. 62-74, 2015.

[15] The official portal for European data. Приступљено: фебруара 2022. Доступно на: https://data.europa.eu/en

[16] The official portal for European data: broj setova podataka po zemljama, Приступљено: фебруара 2022. Доступно на: https://data.europa.eu/catalogue-statistics/Evolution?locale=en

[17] Petar Milić, Nataša Veljković, Leonid Stoimenov. Comparative analysis of metadata models on e-government open data platforms. *IEEE Transactions on Emerging Topics in Computing*, 2018. doi: 10.1109/TETC.2018.2815591

[18] DCAT - Data Catalog Vocabulary. Доступно на: https://www.w3.org/TR/vocab-dcat/ Присупљено: 2022.

[19] European Commission, Directorate-General of Communications Networks, Content and Technology, Daphne van Hesteren, Laura van Knippenberg. Open Data Maturity Report 2021. Last update 02-12-2021.

[20] Katrin Braunschweig, Julian Eberius, Maik Thiele and Wolfgang Lehner. The State of Open Data Limits of Current Open Data Platforms, 2012.

[21] Zuiderwijk, Anneke, Cécile Volten, Maarten Kroesen, and Mark Gill. Motivation perspectives on opening up municipality data: Does municipality size matter? *Information* 9(11): 267, 2018. doi: 10.3390/info9110267.

[22] Konrad Johannes Reiche and Edzard Höfig. Implementation of metadata quality metrics and application on public government data. In *2013 IEEE 37th Annual Computer Software and Applications Conference Workshops*, Japan, pp. 236-241, 2013. doi: 10.1109/COMPSACW.2013.32.

[23] Jürgen Umbrich, Sebastian Neumaier and Axel Polleres. Quality assessment & evolution of open data portals. In *Proceedings IEEE International Conference on Open and Big Data*, IEEE, Rome, pp. 1-8, 2015.

[24] Milena Frtunić Gligorijević, Miloš Bogdanović, and Leonid Stoimenov. Tracking metadata changes in the government open data portals. In *Zdravković, M., Trajanović, M., Konjović, Z. (Eds.) ICIST 2022 Proceedings*, pp.180-184, 2022. ISBN 978-86-85525-24-7

[25] Ross Thompson. Growing by Shrinking - Consolidating Data on the Open Government Portal. Доступно на: https://open.canada.ca/en/blog/growing-shrinking-consolidating-data-open-government-portal, Приступљено: фебруара 2022.

[26] Milena Frtunić Gligorijević, Miloš Bogdanović, Nataša Veljković, and Leonid Stoimenov. Open data categorization based on formal concept analysis. In *IEEE Transactions on Emerging Topics in Computing*, vol. 9, no. 2, pp. 571-581, 1 April-June 2021, doi: 10.1109/TETC.2019.2919330

[27] CKAN API. Доступно на: https://docs.ckan.org/en/latest/api/index.html

[28] Rudolf Wille. Restructuring lattice theory: An approach based on hierarchies of concepts. *Ordered Sets*, Springer, Dordrecht, pp. 445–470, 1982.

[29] Garrett Birkhoff, Lattice theory, American Mathematical Society Coll. Publ. 25, Providence, RI, 1973.

[30] Marc Barbut, Bernard Monjardet. Ordre et classification, algèbre et combinatoire. Paris, Hachette, 1970.

[31] Bernhard Ganter and Rudolf Wille. Formal Concept Analysis: Mathematical Foundations. Springer Berlin, 1999. doi: 10.1007/978-3-642-59830-2

[32] Bernhard Ganter and Gerd Stumme. Formal concept analysis: Methods and applications in computer science. Technical Report Otto – von – Guericke – Universitat Magdeburg. 2003

[33] Prem Kumar Singh, Cherukuri Aswani Kumar and Abdullah Gani. A comprehensive survey on formal concept analysis, its research trends and applications. *International Journal of Applied Mathematics and Computer Science*, vol. 26, no. 2, pp. 495–516, 2016. doi: 10.1515/amcs-2016-0035

[34] Radim Belohlavek. Introduction to formal concept analysis. Palacký University Olomouc. 2008, Доступно на: https://phoenix.inf.upol.cz/esf/ucebni/formal.pdf. Приступљено: септембра 2022.

[35] Olga Prokasheva, Alina Onishchenko and Sergey Gurov. Classification methods based on formal concept analysis. *FCAIR 2012 – Formal Concept Analysis Meets Information Retrieval*, pp. 95-104, 2013.

[36] Hayfa Azibi, Nida Meddouri and Mondher Maddouri. Survey on Formal Concept Analysis Based Supervised Classification Techniques. *Machine Learning and Artificial Intelligence*, IOS Press, 2020. doi:10.3233/FAIA200762

[37] Huaiyu Fu, Huaiguo Fu, Patrik Njiwoua and Engelbert Mephu Nguifo. A Comparative Study of FCA-Based Supervised Classification Algorithms. In: Eklund, P. (eds) Concept Lattices. ICFCA 2004. Lecture Notes in Computer Science(), vol. 2961. Springer, Berlin, Heidelberg. 2004. doi: 10.1007/978-3-540-24651-0_26

[38] Nida Meddouri and Mondher Meddouri. Classification Methods based on Formal Concept Analysis. *CLA 2008 (Posters)*, pp. 9–16, Palacky University, Olomouc, 2008.

[39] Mehran Sahami. Learning classification Rules Using Lattices. In *N. Lavrac and S. Wrobel (eds.) Machine Learning: ECML-95*. ECML 1995, pp. 343–346, Heraclion, Crete, Greece, 1995.

[40] Claudio Caprineto and Giovanni Romano. GALOIS: An order-theoretic approach to conceptual clustering. In *Proceedings of ICML93*, pp. 33–40, Amherst, USA, 1993.

[41] Stéphanie Guillas, Karell Bertet and Jean-Marc Ogier. Extension of Bordat's algorithm for attributes. In *Proceedings of the Fifth International Conference on Concept Lattices* and *Their Applications*, CLA 2007, Montpellier, France, October 24-26, 2007

[42] Brahim Douar, Chiraz Latiri, and Yahya Slimani. Approche hybride de classification supervisée à base de treillis de Galois: application à la reconnaissance de visages. *Extraction et Gestion des Connaissances (EGC08)*, 309–320, Nice, France, 2008

[43] Melisachew Wudage Chekol and Amedeo Napoli. An FCA framework for knowledge discovery in SPARQL query answers. In *Proceedings of the 12th*

International Semantic Web Conference (Posters & Demonstrations Track) - Volume 1035 (ISWC-PD '13), pp. 197–200, 2013.

[44] Mehwish Alam, Aleksey Buzmakov, Victor Codocedo and Amedeo Napoli. Mining definitions from RDF annotations using formal concept analysis. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015).* 2015.

[45] Fengbo Zheng and Licong Cui. A Lexical-based Formal Concept Analysis Method to Identify Missing Concepts in the NCI Thesaurus. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Seoul, Korea (South), 2020, pp. 1757-1760, doi: 10.1109/BIBM49941.2020.9313186.

[46] Alessandro Marco Boutari, Claudio Carpineto, and Raffaele Nicolussi. Evaluating term concept association measures for short text expansion: two case studies of classification and clustering. In *Proceedings of the 7th International Conference on Concept Lattices and Their Applications (CLA 2010)*, pp. 163–174, 2010.

[47] Valmi Dufour-Lussier, Jean Lieber, Emmanuel Nauer, and Yannick Toussaint. In: Bichindaritz, I., Montani, S. (eds) Case-Based Reasoning. Research and Development. ICCBR 2010. Lecture Notes in Computer Science, vol 6176, pp. 96–110. Springer, Berlin, Heidelberg. 2010.

[48] Zeina Azmeh, Fady Hamoui, Marianne Huchard, Nizar Messai, Chouki Tibermacine, Christelle Urtado, and Sylvain Vauttier. Backing Composite Web Services Using Formal Concept Analysis. In: *Valtchev, P., Jäschke, R. (eds) Formal Concept Analysis. ICFCA 2011. Lecture Notes in Computer Science*, vol 6628, pp. 26–41. Springer, Berlin, Heidelberg. 2011. doi: 10.1007/978-3-642-20514-9_4

[49] Claudio Carpineto, Carla Michini, and Raffaele Nicolussi. A concept lattice-based kernel for SVM text classification. In *Ferré, S., Rudolph, S. (eds) Formal Concept Analysis. ICFCA 2009. Lecture Notes in Computer Science*, vol 5548. Springer, Berlin, Heidelberg, pp. 237–250, 2009.

[50] Stephanie Chollet, Vincent Lestideau, Philippe Lalanda, Yoann Maurel, Pierre Colomb, and Olivier Raynaud. Building FCA-based Decision Trees for the Selection of Heterogeneous Services. In *SCC '11: Proceedings of the 2011 IEEE International Conference on Services Computing*, pp. 616–623, Washington, DC, USA, 2011.

[51] Markus Kirchberg, Erwin Leonardi, Yu Shyang Tan, Sebastian Link, Ryan K. L. Ko, and Bu Sung Lee. Formal Concept Discovery in Semantic Web Data. In 10th International Conference, ICFCA 2012, Leuven, Belgium, 2012.

[52] Gaihua Fu. FCA based ontology development for data integration. *Information Processing and Management*, vol. 52, issue 5, pp. 765–782, 2016.

[53] C. D. Maio, G. Fenza, M. Gaeta, V. Loia, F. Orciuoli, and S. Senatore. RSS-based e-learning recommendations exploiting fuzzy FCA for knowledge modeling. *Applied Soft Computing*, vol. 12, issue 1, pp. 113-124, 2012.

[54] Amel Grissa Touzi, Hela Ben Massoud, and Alaya Ayadi. Automatic Ontology Generation for Data Mining Using FCA and Clustering. arxiv.org, no. 1311.1764.

[55] Liya Fan and Tianyuan Xiao. An automatic method for ontology mapping. B. Apolloni, et al. (Eds.), KES/WIRN, Part III, LNAI, 4694, Springer, pp. 661-669, 2007.

[56] Guoxuan Li. DeepFCA: Matching biomedical ontologies using formal concept analysis embedding techniques. In *Proceedings of the 4th International Conference on Medical and Health Informatics*. pp. 259-265. 2020.

[57] Mengyi Zhao, Songmao Zhang, Weizhuo Li, and Guowei Chen. Matching biomedical ontologies based on formal concept analysis. *Journal of Biomedical Semantics* 9, 11, 2018. doi:10.1186/s13326-018-0178-9

[58] Jones Poelmans, Dmitry I. Ignatov, Sergei O. Kuznetsov, and Guido Dedene. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, vol. 40, issue 16, pp. 6538-6560, 2013.

[59] Jonas Poelmans, Dmitry I. Ignatov, Sergei O. Kuznetsov, and Guido Dedene. Formal concept analysis in knowledge processing: A survey on models and techniques. *Expert systems with Applications*. vol. 40, issue 16, pp. 6601-6623, 2013.

[60] Prem Kumar Singh, Cherukuri Aswani Kumar, and Abdullah Gani. A comprehensive survey on formal concept analysis, its research trends and applications. *International Journal of Applied Mathematics and Computer Science* ol.26, no.2, 3916, pp.495-516, 2016. doi: 10.1515/amcs-2016-0035

[61] Stratos Idreos, Olga Papaemmanouil, and Surajit Chaudhuri. Overview of Data Exploration Techniques. In *SIGMOD'15: Proceedings of the 2015 ACM SIGMOD*

International Conference on Management of Data. pp. 277–281, 2015. doi: 10.1145/2723372.2731084

[62] Jeffrey Heer and Ben Shneiderman. Interactive Dynamics for Visual Analysis. Commun. ACM, 55(4), pp 45–54, 2012. doi: 10.1145/2133806.2133821

[63] Ben Shneiderman. Extreme Visualization: Squeezing a Billion Records into a Million Pixels. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, SIGMOD 2008, 2008. doi:10.1145/1376616.1376618

[64] Jeffrey Heer and Sean Kandel. Interactive Analysis of Big Data. ACM Crossroads, 19(1), pp. 50-54, 2012. doi: 10.1145/2331042.2331058

[65] Kristi Morton, Magdalena Balazinska, Dan Grossman, and Jock Mackinlay. Support the Data Enthusiast: Challenges for Next-Generation Data-Analysis Systems. In *Proceedings of the VLDB Endowment*, 7(6), pp 453–456, 2014. doi: 10.14778/2732279.2732282

[66] Eugene Wu, Leilani Battle, and Samuel R. Madden. The Case for Data Visualization Management Systems. In *Proceedings of the VLDB Endowment*, 7(10), pp. 903-906, 2014. doi: 10.14778/2732951.2732964

[67] Parke Godfrey, Jarek Gryz, and Piotr Lasek. Interactive Visualization of Large Data Sets. In *IEEE Transactions on Knowledge and Data Engineering*, vol. 28(8), pp. 2142-2157, 2016. doi: 10.1109/TKDE.2016.2557324

[68] Enrico G. Caldarola, and Antonio M. Rinaldi. Big Data Visualization Tools: A Survey - The New Paradigms, Methodologies and Tools for Large Data Sets Visualization. In *6th International Conference on Data Science, Technology and Applications (DATA 2017)*. 2017.

[69] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Symposium on Visual Languages*, 1996. doi: 10.1109/VL.1996.545307

[70] Aba-Sah Dadzie and Matthew Rowe. Approaches to visualising Linked Data: A survey. *Semantic Web*, vol. 2, no. 2, pp. 89-124, 2011. doi: 10.3233/SW-2011-0037

[71] Nicolas Marie and Fabien Lucien Gandon. Survey of Linked Data Based Exploration Systems. In *IESD 2014 - Intelligent Exploitation of Semantic Data*, 2014.

[72] Fahad Alahmari, James A. Thom, Liam Magee and Wilson Wong. Evaluating Semantic Browsers for Consuming Linked Data. In *Australasian Database Conference*, 2012.

[73] Suvodeep Mazumdar, Daniela Petrelli, Khadija Elbedweihy, Vitaveska Lanfranchi, and Fabio Ciravegna. Affective graphs: The visual appeal of Linked Data. *Semantic Web*, vol. 6, no. 3, pp. 277-312, 2015. doi: 10.3233/SW-140162

[74] Nikos Bikakis and Timos Sellis. Exploration and Visualization in the Web of Big Linked Data: A Survey of the State of the Art. *6th International Workshop on Linked Web Data Management (LWDM 2016)*, 2016.

[75] Milena Frtunić Gligorijević, Miloš Bogdanović, Nataša Veljković, Leonid Stoimenov. Tool for Interactive Visual Analysis of Large Hierarchial Data Structures. *FACTA UNIVERSITATIS Series: Automatic Control and Robotics*, Vol. 20, No 2, 2021, pp. 111 - 121, doi: 10.22190/FUACR210715009F, Online ISSN: 1820-6425

[76] D3.js. Доступно на: https://d3js.org/

[77] Bernhard Ganter and Rudolph Wille. Formal Concept Analysis: Mathematical Foundations. Springer, Berlin-Heidelberg, 1999.

[78] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. 2014.

[79] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*. 41(6), 391-407, 1990.

[80] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics. 2013.

[81] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. vol. 5, pp. 135–146, 2017. doi: 10.1162/tacl_a_00051.

[82] Zhe Zhao, Tao Liu, Shen Li, Bofang Li, and Xiaoyong Du. Ngram2vec: Learning Improved Word Representations from Ngram Co-occurrence Statistics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 244–253, Copenhagen, Denmark. Association for Computational Linguistics. 2017.

[83] Julien Tissier, Christophe Gravier, and Amaury Habrard. Dict2vec: Learning Word Embeddings using Lexical Dictionaries. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Copenhagen, Denmark. Association for Computational Linguistics. 2017.

[84] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, C.-C. Jay Kuo. Evaluating word embedding models: Methods and experimental results. In *APSIPA Transactions on Signal and Information Processing*. vol. 8: no. 1, e19, 2019. doi: 10.1017/ATSIP.2019.12.

[85] Common Crawl. Доступно на: https://commoncrawl.org/

[86] Navid Rekabsaz, Mihai Lupu and Allan Hanbury. Exploration of a Threshold for Similarity Based on Uncertainty in Word Embedding. In *Advances in Information Retrieval*. ECIR 2017. Lecture Notes in Computer Science, vol 10193. Springer, Cham. 2017. doi: 10.1007/978-3-319-56608-5_31

[87] Канадски портал отворених података. Доступно на: https://open.canada.ca/en/open-data, 2022.

ПРИЛОГ А - ПРЕГЛЕД ПОРТАЛА ОТВОРЕНИХ ПОДАТАКА КОРИШЋЕНИХ У НАУЧНОМ ИСТРАЖИВАЊУ

У наставку је дат преглед свих портала отворених података коришћених у оквиру овог научног истраживања. За сваки портал приказане су информације о називу портала, URL адреси портала и коју платформу портал користи.

Табела 10 – Листа портала отворених података коришћена у анализама

Портал	URL адреса	Платформа
Africa	https://africaopendata.org/	CKAN
Alberta Government	https://open.alberta.ca/	CKAN
Aragon	https://opendata.aragon.es/	CKAN
Austria	https://data.gv.at/katalog/	CKAN
Brazil	http://www.dados.gov.br/	CKAN
Brisbane	https://www.data.brisbane.qld.gov.au/data/	CKAN
Buenos Aires	https://data.buenosaires.gob.ar/	CKAN
Canada	https://open.canada.ca/data/en/	CKAN
Chile	https://datos.gob.cl/	CKAN
Colorado	https://data.colorado.gov/	Socrata
Dutch	https://data.overheid.nl/data/	CKAN
Germany	https://ckan.govdata.de/	CKAN
Ireland	https://data.gov.ie/	CKAN
Italy	https://www.dati.gov.it/opendata/	CKAN
London	https://data.london.gov.uk/	CKAN
Maryland	https://opendata.maryland.gov/	Socrata
Melbourne	https://data.melbourne.vic.gov.au/	Socrata
Mexico	https://datos.gob.mx/busca/	CKAN
New Jersey	https://data.nj.gov/	Socrata
New South Wales	https://data.nsw.gov.au/data/	CKAN
New York State	https://data.ny.gov/	Socrata

New Zealand	https://catalogue.data.govt.nz/	CKAN
Niagara	https://niagaraopendata.ca/	CKAN
Oregon	https://data.oregon.gov/	Socrata
Pennsylvania	https://data.pa.gov/	Socrata
Queensland	https://www.data.qld.gov.au/	CKAN
Romania	https://data.gov.ro/	CKAN
Slovakia	https://data.gov.sk/	CKAN
South Australia	https://data.sa.gov.au/data/	CKAN
Surrey	http://data.surrey.ca/	CKAN
Swiss	https://ckan.opendata.swiss/en/	CKAN
Texas	https://data.texas.gov/	Socrata
United Kingdom	https://data.gov.uk/	CKAN
Uruguay	https://catalogodatos.gub.uy/	CKAN
USA	https://catalog.data.gov/	CKAN
Vermont	https://data.vermont.gov/	Socrata
Victoria	https://discover.data.vic.gov.au/	CKAN
Washington	https://data.wa.gov/	Socrata
Western Australia	https://catalogue.data.wa.gov.au/	CKAN
Western Pennsylvania	https://data.wprdc.org/	CKAN

ПРИЛОГ Б – ПРИМЕР МЕТАПОДАТАКА ПОДАТАКА СА ПОРТАЛА ОТВОРЕНИХ ПОДАТАКА

У наставку је дат пример са новозеландског портала отворених података који користи CKAN платформу. У примеру се виде метаподаци које корисник добија за један скуп података са овог портала. Пример који се налази у наставку приказује метаподатке скупа података који припада категорији *Education*.


```
{
    "author": "Ministry of Education",
    "author_email": "",
    "creator_user_id": "df58f452-5866-4edb-b658-49e55ab53831",
    "frequency_of_update": "Continuously updated",
    "id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
    "isopen": true,
    "issued": "2011-05-11",
    "language": "",
    "license_id": "CC-BY-4.0",
    "license_title": "Creative Commons Attribution 4.0 International",
    "license_url": "https://creativecommons.org/licenses/by/4.0/",
    "maintainer": "Education Data Requests",
    "maintainer_email": "*******@education.govt.nz",
    "maintainer_phone": "",
    "metadata_created": "2020-08-12T03:09:39.230277",
    "metadata_modified": "2023-09-07T18:00:37.847008",
    "modified": "2020-08-19",
    "name": "directory-of-educational-institutions",
    "notes": "The Ministry of Education maintains the following
databases that are updated regularly that include contact details,
institution information and regional information.",
    "num_resources": 8,
    "num_tags": 6,
    "organization": {
        "id": "6be2dce8-6b51-48d2-b561-0b75337f06af",
        "name": "ministry-of-education",
        "title": "Ministry of Education",
        "type": "organization",
        "description": "The following key areas summarise the overall
focus of the Ministry's work:\r\n\r\n1) More children gaining strong
learning foundations;\r\n\r\n2) More students participating in and
achieving in education;\r\n\r\n3) Provision of services - directly and
indirectly - to children and young people with special education
needs;\r\n\r\n4) Better quality schools and teachers;\r\n\r\n5) More
people continuing to develop their capabilities;\r\n\r\n6) Maori
```

```
education;\r\n\r\n7) Families and communities more strongly engaged in
education;\r\n\r\n8) Increasing Ministry capability\r\n\r\nThe Ministry
aims to build on the results already achieved so that we can make a
significant difference to further the achievement of improved outcomes
for our students, families and communities and enhance the country's
long-term economic prosperity and social wellbeing.",
        "image_url": "2017-09-25-222139.948094download-1.png",
        "created": "2017-02-18T21:11:52.754448",
        "is_organization": true,
        "approval_status": "approved",
        "state": "active"
    },
    "owner_org": "6be2dce8-6b51-48d2-b561-0b75337f06af",
    "private": false,
    "rights": "",
    "source_identifier":
"https://www.educationcounts.govt.nz/directories",
    "spatial": "",
    "state": "active",
    "temporal": "",
    "theme": "Education",
    "title": "Directory of educational institutions",
    "type": "dataset",
    "url": "https://www.educationcounts.govt.nz/directories",
    "version": null,
    "extras": [
        {
            "key": "harvest_object_id",
            "value": "33cb91e2-510a-4672-9d54-e49705d462f2"
        },
        {
            "key": "harvest_source_id",
            "value": "b1467f8c-a7e4-4742-bef4-bf4af88604ad"
        },
        {
            "key": "harvest_source_title",
            "value": "Ministry of Education migration harvest - DO NOT
RUN"
        }
    ],
    "groups": [
        {
            "description": "",
            "display_name": "Education",
            "id": "9c197059-f4c8-4df7-b175-d8f8936859b6",
            "image_display_url":
"https://catalogue.data.govt.nz/uploads/group/2018-10-03-
010228.692387MAI249946TePapaNormal-School-Christchurchfull.jpg",
```

```
            "name": "education",
            "title": "Education"
        }
    ],
    "resources": [
        {
            "cache_last_updated": null,
            "cache_url": null,
            "created": "2020-08-12T03:09:39.240678",
            "datastore_active": false,
            "description": "A list of ECE including contact details,
institution and regional information and management contact details.",
            "format": "XLSX",
            "hash": "",
            "id": "5171453e-3b37-4d26-98e7-4a5f295ff078",
            "last_modified": null,
            "metadata_modified": "2020-08-12T03:09:39.240678",
            "mimetype": "application/vnd.openxmlformats-
officedocument.spreadsheetml.sheet",
            "mimetype_inner": null,
            "name": "Early Childhood Services (ECE) Directory",
            "package_id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
            "position": 0,
            "resource_type": null,
            "size": null,
            "state": "active",
            "url":
"https://www.educationcounts.govt.nz/__data/assets/excel_doc/0020/62570
/Directory-ECE-Current.xlsx",
            "url_type": ""
        },
        {
            "cache_last_updated": null,
            "cache_url": null,
            "created": "2020-08-12T03:09:39.240688",
            "datastore_active": true,
            "format": "CSV",
            "hash": "",
            "id": "f65dfeb4-94be-4879-957c-e081d9570216",
            "last_modified": "2023-09-07T18:00:37.827056",
            "metadata_modified": "2023-09-07T18:00:37.856346",
            "mimetype": "text/csv",
            "mimetype_inner": null,
            "name": "Early Childhood Services (ECE) Directory",
            "package_id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
            "position": 1,
            "resource_type": null,
            "size": 2241150,
```

```
            "state": "active",
            "url": "https://catalogue.data.govt.nz/dataset/c1923d33-
e781-46c9-9ea1-d9b850082be4/resource/f65dfeb4-94be-4879-957c-
e081d9570216/download/ecedirectory-08-09-2023-060037.csv",
            "url_type": "upload"
        },
        {
            "cache_last_updated": null,
            "cache_url": null,
            "created": "2020-08-12T03:09:39.240692",
            "datastore_active": false,
            "description": "A list of New Zealand schools where all, or
some, of their students are taught curriculum subjects in the
M\u0101ori language for at least 51 percent of the time.",
            "format": "XLSX",
            "hash": "",
            "id": "ad297e71-05ff-4974-a058-f369d41918fa",
            "last_modified": null,
            "metadata_modified": "2023-04-18T06:04:20.333421",
            "mimetype": null,
            "mimetype_inner": null,
            "name": "M\u0101ori Medium Schools",
            "package_id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
            "position": 2,
            "resource_type": null,
            "size": null,
            "state": "active",
            "url":
"https://www.educationcounts.govt.nz/directories/maori-medium-
schools/directory-maori-medium-current.xlsx",
            "url_type": ""
        },
        {
            "cache_last_updated": null,
            "cache_url": null,
            "created": "2020-08-12T03:09:39.240695",
            "datastore_active": false,
            "description": "A list of New Zealand schools where all, or
some, of their students are taught curriculum subjects in the
M\u0101ori language for at least 51 percent of the time.",
            "format": "CSV",
            "hash": "",
            "id": "c3502ea6-e7dd-4f8d-8201-1f92bbd46596",
            "last_modified": null,
            "metadata_modified": "2023-04-18T06:12:25.637882",
            "mimetype": "text/csv",
            "mimetype_inner": null,
            "name": "M\u0101ori Medium schools",
```

```
            "package_id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
            "position": 3,
            "resource_type": null,
            "size": null,
            "state": "active",
            "url":
"https://www.educationcounts.govt.nz/directories/maori-medium-
schools/directory-maori-medium-current.csv",
            "url_type": ""
        },
        {
            "cache_last_updated": null,
            "cache_url": null,
            "created": "2020-08-12T03:09:39.240698",
            "datastore_active": true,
            "format": "CSV",
            "hash": "",
            "id": "20b7c271-fd5a-4c9e-869b-481a0e2453cd",
            "last_modified": "2023-09-06T20:30:38.445291",
            "metadata_modified": "2023-09-06T20:30:38.476461",
            "mimetype": "text/csv",
            "mimetype_inner": null,
            "name": "New Zealand Schools",
            "package_id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
            "position": 4,
            "resource_type": null,
            "size": 1582092,
            "state": "active",
            "url": "https://catalogue.data.govt.nz/dataset/c1923d33-
e781-46c9-9ea1-d9b850082be4/resource/20b7c271-fd5a-4c9e-869b-
481a0e2453cd/download/schooldirectory-07-09-2023-083037.csv",
            "url_type": "upload"
        },
        {
            "cache_last_updated": null,
            "cache_url": null,
            "created": "2020-08-12T03:09:39.240711",
            "datastore_active": false,
            "description": "A list of New Zealand tertiary providers,
including their contact details and institutional information.",
            "format": "XLSX",
            "hash": "",
            "id": "6fa89d95-6f28-4e02-bc63-cb2fd25a7664",
            "last_modified": null,
            "metadata_modified": "2023-04-18T06:14:11.193066",
            "mimetype": null,
            "mimetype_inner": null,
            "name": "Tertiary Providers",
```

```
            "package_id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
            "position": 5,
            "resource_type": null,
            "size": null,
            "state": "active",
            "url":
"https://www.educationcounts.govt.nz/directories/list-of-tertiary-
providers/directory-tertiary-current.xlsx",
            "url_type": ""
        },
        {
            "cache_last_updated": null,
            "cache_url": null,
            "created": "2020-08-12T03:09:39.240714",
            "datastore_active": true,
            "description": "A list of New Zealand tertiary providers,
including their contact details and institutional information.",
            "format": "CSV",
            "hash": "",
            "id": "4c718954-5098-4499-bf74-198885fc1aa8",
            "last_modified": null,
            "metadata_modified": "2023-04-18T06:15:19.624874",
            "mimetype": "text/csv",
            "mimetype_inner": null,
            "name": "Tertiary Providers",
            "package_id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
            "position": 6,
            "resource_type": null,
            "size": null,
            "state": "active",
            "url":
"https://www.educationcounts.govt.nz/directories/list-of-tertiary-
providers/directory-tertiary-current.csv",
            "url_type": ""
        },
        {
            "cache_last_updated": null,
            "cache_url": null,
            "created": "2020-08-12T03:09:39.240720",
            "datastore_active": false,
            "description": "Every year some schools may be merged,
closed and opened.  The 'most recent' workbook on this page consists of
the latest information available. Only changes that occurred in that
year are included in each workbook.",
            "format": "ZIP",
            "hash": "",
            "id": "1201eb38-8221-429b-b999-253f120f6332",
            "last_modified": null,
```

```
            "metadata_modified": "2023-03-22T00:51:48.731023",
            "mimetype": "application/zip",
            "mimetype_inner": null,
            "name": "School mergers, closures and new schools",
            "package_id": "c1923d33-e781-46c9-9ea1-d9b850082be4",
            "position": 7,
            "resource_type": null,
            "size": null,
            "state": "active",
            "url":
"https://www.educationcounts.govt.nz/directories/school-mergers/school-
mergers-closures-and-new-schools.zip",
            "url_type": ""
        }
    ],
    "tags": [
        {
            "display_name": "Education",
            "id": "65399142-f945-4162-a7e6-2bfac78f86c1",
            "name": "Education",
            "state": "active",
            "vocabulary_id": null
        },
        {
            "display_name": "address",
            "id": "89bc72c9-1a09-4983-a251-8faaf4c7f816",
            "name": "address",
            "state": "active",
            "vocabulary_id": null
        },
        {
            "display_name": "directory",
            "id": "7afd50c8-4b89-44a3-8aba-3990b355d938",
            "name": "directory",
            "state": "active",
            "vocabulary_id": null
        },
        {
            "display_name": "locations",
            "id": "71e5125f-3467-460f-aa51-c2242ce9f36b",
            "name": "locations",
            "state": "active",
            "vocabulary_id": null
        },
        {
            "display_name": "roll",
            "id": "d84dd6e7-214f-4899-b118-e713acca18e9",
            "name": "roll",
```

```
            "state": "active",
            "vocabulary_id": null
        },
        {
            "display_name": "schools",
            "id": "6a92c528-5c1b-4d5d-ac74-4f8210f54834",
            "name": "schools",
            "state": "active",
            "vocabulary_id": null
        }
    ],
    "relationships_as_subject": [],
    "relationships_as_object": []
}
```

БИОГРАФИЈА АУТОРА

Милена Фртунић Глигоријевић рођена је у Скопљу 25.02.1989. године. Основну школу и гимназију „Бора Станковић“ завршила је у Нишу. Добитница је дипломе Вук Караџић за остварен успех у основној и средњој школи. Електронски факултет у Нишу, на Универзитету у Нишу, уписала је 2008. године на смеру Рачунарство и информатика, одсек Софтверско инжењерство. Дипломирала је 2013. године са просечном оценом 9,66. На крају студија стекла је звање Мастер инжењер електротехнике и рачунарства из области за рачунарску технику и информатику. Као стипендиста програма билатералне сарадње између Универзитета у Нишу и Универзитета НТНУ у Трондхајму, провела је семестар пете године студија у Норвешкој.

Докторске студије на Електронском факултету у Нишу уписала је 2013. године на смеру за Рачунарство и информатику.

У периоду од 09.05.2014. до 30.09.2014. године била је стипендиста Министарства просвете, науке и технолошког развоја Републике Србије као истраживач-сарадник. Од 01.10.2014. до 30.06.2016. године ради на Електронском факултету у Нишу као стручни сарадник за научноистраживачки рад у оквиру пројекта Министарства просвете, науке и технолошког развоја Републике Србије „Инфраструктура за електронски подржано учење у Србији“. Од 01.07.2016. до 30.06.2017. године ради на Електронском факултету у Нишу као сарадник у настави. Од 13.07.2017. године ради на Електронском факултету у Нишу као асистент.

Ангажована је у извођењу рачунских и лабораторијских вежби из више предмета. Такође, ангажована је на пројекту Министарства просвете, науке и технолошког развоја Републике Србије „Инфраструктура за електронски подржано учење у Србији“, као и на пројекту Министарства енергетике, развоја и заштите животне средине „Норвешка помоћ енергетској политици Републике Србије у области локалног енергетског планирања“.

Као аутор и коаутор објавила је 6 научних радова у часописима, 13 радова на националним и међународним конференцијама и 4 техничка решења. Коаутор је једног помоћног уџбеника.

ИЗЈАВА О АУТОРСТВУ

Изјављујем да је докторска дисертација, под насловом

УНАПРЕЂЕЊЕ УПОТРЕБЉИВОСТИ ОТВОРЕНИХ ПОДАТАКА ДЕФИНИСАЊЕМ МЕТОДЕ КАТЕГОРИЗАЦИЈЕ ЗАСНОВАНЕ НА МЕТАПОДАЦИМА ПОРТАЛА ОТВОРЕНИХ ПОДАТАКА

која је одбрањена на Електронском факултету Универзитета у Нишу:

- резултат сопственог истраживачког рада;
- да ову дисертацију, ни у целини, нити у деловима, нисам пријављивао/ла на другим факултетима, нити универзитетима;
- да нисам повредио/ла ауторска права, нити злоупотребио/ла интелектуалну својину других лица.

Дозвољавам да се објаве моји лични подаци, који су у вези са ауторством и добијањем академског звања доктора наука, као што су име и презиме, година и место рођења и датум одбране рада, и то у каталогу Библиотеке, Дигиталном репозиторијуму Универзитета у Нишу, као и у публикацијама Универзитета у Нишу.

У Нишу, _____.

Потпис аутора дисертације:

М. Фртунић Глигоријевић

Милена Б. Фртунић Глигоријевић

ИЗЈАВА О ИСТОВЕТНОСТИ ШТАМПАНОГ И ЕЛЕКТРОНСКОГ ОБЛИКА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Наслов дисертације:

УНАПРЕЂЕЊЕ УПОТРЕБЉИВОСТИ ОТВОРЕНИХ ПОДАТАКА ДЕФИНИСАЊЕМ МЕТОДЕ КАТЕГОРИЗАЦИЈЕ ЗАСНОВАНЕ НА МЕТАПОДАЦИМА ПОРТАЛА ОТВОРЕНИХ ПОДАТАКА

Изјављујем да је електронски облик моје докторске дисертације, коју сам предао/ла за уношење у Дигитални репозиторијум Универзитета у Нишу, истоветан штампаном облику.

У Нишу, _____.

Потпис аутора дисертације:

М. Фртунић Глигоријевић

Милена Б. Фртунић Глигоријевић

ИЗЈАВА О КОРИШЋЕЊУ

Овлашћујем Универзитетску библиотеку „Никола Тесла“ да у Дигитални репозиторијум Универзитета у Нишу унесе моју докторску дисертацију, под насловом:

УНАПРЕЂЕЊЕ УПОТРЕБЉИВОСТИ ОТВОРЕНИХ ПОДАТАКА ДЕФИНИСАЊЕМ МЕТОДЕ КАТЕГОРИЗАЦИЈЕ ЗАСНОВАНЕ НА МЕТАПОДАЦИМА ПОРТАЛА ОТВОРЕНИХ ПОДАТАКА

Дисертацију са свим прилозима предао/ла сам у електронском облику, погодном за трајно архивирање.

Моју докторску дисертацију, унету у Дигитални репозиторијум Универзитета у Нишу, могу користити сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons), за коју сам се одлучио/ла.

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. **Ауторство – некомерцијално – без прераде (CC BY-NC-ND)**
4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)
5. Ауторство – без прераде (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

У Нишу, _____.

Потпис аутора дисертације:

М. Фртунић Глигоријевић

Милена Б. Фртунић Глигоријевић