Природно-математички факултет
Универзитет у Крагујевцу

Марија Ђокић Петровић

Биоинформатичка платформа за извршавање Federated SPARQL упита над онтолошким базама података и детектовање сличних података утврђивањем њихове семантичке повезаности

документ

Крагујевац, 2019.

ИДЕНТИФИКАЦИОНА СТРАНИЦА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ
I. Аутор
Име и презиме: Марија Ђокић Петровић
Датум и место рођења: 11.02.1986. Крагујевац, Србија
Садашње запослење: „Das Virtuelle Fahrzeug“ Forschungsgesellschaft mbH, Graz, Österreich, Gesellschafter (Technische Universität Graz, AVL List GmbH, Magna Steyr Fahrzeug AG & Co KG, Siemens AG Österreich, Joanneum Research Forschungs Forschung-GmbH); позиција: истраживач
II. Докторска дисертација
Наслов: Биоинформатичка платформа за извршавање Federated SPARQL упита над онтолошким базама података и детектовање сличних података утврђивањем њихове семантичке повезаности
Број страница: 169
Број слика: 62
Број листинга: 1
Број табела: 25
Број библиографских података: 233
Установа и место где је рад израђен: Природно-математички факултет Универзитета у Крагујевцу
Научна област (УДК): 004 (рачунарство)
Ментор: др Владимир Цвјетковић, доцент Природно-математичког факултета Универзитета у Крагујевцу, ужа научна област: Информатика у физици
III. Оцена и одбрана
Датум пријаве теме: 29.08.2018
Број одлуке и датум прихватања докторске дисертације:
Комисија за оцену подобности теме и кандидата: **1.** др Мирјана Ивановић, редовни професор Природно-математичког факултета Универзитета у Новом Саду **2.** др Ненад Стефановић, ванредни професор Природно-математичког факултета Универзитета у Крагујевцу **3.** др Снежана Марковић, доцент Природно-математичког факултета Универзитета у Крагујевцу **4.** др Татјана Стојановић, доцент Природно-математичког факултета Универзитета у Крагујевцу **5.** др Владимир Цвјетковић, доцент Природно-математичког факултета Универзитета у Крагујевцу
Комисија за оцену докторске дисертације: **1.** др Мирјана Ивановић, редовни професор Природно-математичког факултета Универзитета у Новом Саду **2.** др Милош Ивановић, ванредни професор Природно-математичког факултета Универзитета у Крагујевцу **3.** др Ненад Стефановић, ванредни професор Природно-математичког факултета Универзитета у Крагујевцу **4.** др Ана Капларевић-Малишић, доцент Природно-математичког факултета Универзитета у Крагујевцу **5.** др Снежана Марковић, доцент Природно-математичког факултета Универзитета у Крагујевцу
Комисија за одбрану докторске дисертације: **1.** др Мирјана Ивановић, редовни професор Природно-математичког факултета Универзитета у Новом Саду **2.** др Милош Ивановић, ванредни професор Природно-математичког факултета Универзитета у Крагујевцу **3.** др Ненад Стефановић, ванредни професор Природно-математичког факултета Универзитета у Крагујевцу **4.** др Ана Капларевић-Малишић, доцент Природно-математичког факултета Универзитета у Крагујевцу **5.** др Снежана Марковић, доцент Природно-математичког факултета Универзитета у Крагујевцу
Датум одбране дисертације:

Оцу Велизару

Захвалност

Ова докторска дисертација је резултат вишегодишњег истраживања под менторством проф. др Владимира Цвјетковића, доцента на Институту за физику, Природно-математичког факултета Универзитета у Крагујевцу. Због тога се, управо њему, захваљујем на пружању знања, искуства и конструктивне подршке током читавог истраживачког рада и израде ове дисертације. Захваљујем му се на благонамерним и критичким саветима који су омогућили да ова дисертација буде приведена крају.

Један део експерименталног истраживања реализован је у Лабораторији за ћелијску и молекуларну биологију Института за биологију и екологију Природно-математичког факултета Универзитета у Крагујевцу. Истраживање је представљало део пројекта Министарства просвете, науке и технолошког развоја Републике Србије (ПИБАС, број III41010) под руководством Доц. др Снежане Марковић. Због тога се посебно захваљујем колегама др Марку Живановићу и др Милени Милутиновић из Лабораторије на великој помоћи приликом спровођења експерименталног истраживања.

Такође, желим да се захвалим колегама Бранку Арсићу са Природно-математичког факултета Универзитета у Крагујевцу, проф. др Дејвиду Вилду (David Wild) и Џермију Јангу (Jeremy Yang) са Индијана Универзитета (Indiana University) из Сједињених Америчких Држава за њихов допринос у истраживањима која су омогућила објављивање научних радова везаних за тему ове дисертације.

Посебно се захваљујем и проф. др Драгани Бечејски-Вујаклији са Факултета организационих наука Универзитета у Београду, проф. емеритусу др Херману Мауреру (Hermann Maurer) и др Барбари Кетеле (Barbara Kettele) са Техничког Универзитета у Грацу (Technische Universität Graz) у Аустрији. Њихова искуства, пријатељство и подршка били су ми од великог значаја да истрајем у напорном раду. Захваљујем се и колегама Евалду Моитциу (Ewald Moitzi) и Еви Улбрих (Eva Ulbrich) из аустријске фирме mything GmbH, на подршци коју су ми пружили током израде ове дисертације. Додатно се захваљујем и др Дејвиду Притчарду (David Pritchard) из компаније Google, што је проширио моје програмерске видике које сам искористила током истраживања везаних за ову дисертацију. Захваљујем се и колегама Кристијану Кајсеру (Christian Kaiser) и др Герноту Лехнеру (Gernot Lechner) из истраживачког центра Virtual Vechile у Грацу на пружању подршке током израде дисертације.

Желим да се захвалим и куми Јадранки, пријатељима Николи, Весни, Нини, Јеци и Цимију што су ме увек подржали и дали ми снаге да истрајем.

Посебно се захваљујем супругу Данку и мајци Даници, на љубави и пружању безрезервне подршке у сваком тренутку. Хвала вам што сте веровали у мене!

Информације су данас кључ успеха! Само онај ко уме правилно да их користи, може допринети себи и човечанству!

Крагујевац, мај 2019.
Марија Ђокић Петровић

Резиме

Значај биоинформатике, као интердисциплинарне области, базира се на великом броју биолошких података који се могу адекватно употребити и процесирати применом актуелних информатичких технологија. Оно што је од виталног значаја у домену биоинформатике данас, јесте доступност података релевантних за истраживања, као и сазнање о томе да такви подаци већ постоје. Значајан предуслов за то је да су потребни подаци јавно доступни, интегрисани и да су развијени механизми за њихову претрагу. У циљу решавања датих проблема биоинформатичка заједница користи технологије семантичког веба. У том погледу развијени су многи семантички репозиторијуми и софтверска решења, који су изразито потпомогли истраживачким активностима на биоинформатичкој сцени. Међутим, ови приступи често се суочавају са проблемима јер су се многе базе података развијале у изолованом окружењу, без поштовања основних стандарда биоинформатичке заједнице. Ове хетерогене базе, које су карике многих високо специјализованих и независних ресурса, често користе различите конвенције, речнике и формате за представљање података. Због тога се актуелна софтверска решења суочавају са различитим изазовима у циљу претраге и откривања релевантних података. Такође, многе базе података се преклапају, чиме се покривају, односно прикривају слични подаци, формирајући на тај начин полу-хомогене или хомогене изворе података. У таквим случајевима семантичка корелација оваквих база често је нејасна и неопходно је применити одговарајуће методе за анализу података, како би се утврдили слични подаци. Ова дисертација је настала као резултат истраживања у циљу превазилажења недостатака постојећих решења.

У дисертацији је приказан допринос у развоју биоинформатичке платформе, која се огледа у низу оригиналних софтверских приступа који представљају основу кључних функционалности: извршавање Federated SPARQL упита над иницијалним (и кориснички селектованим) базама података у циљу откривања података релевантних за биоинформатичка истраживања, као и детектовање сличних података које је засновано на утврђивању семантичке повезаности података. Извршавање Federated SPARQL упита изводи се над базама података које користе *Resource Description Framework* (RDF) као модел података. Резултати упита се могу накнадно филтрирати, чиме се доприноси побољшању њихове значајности. Филтрирање подразумева одабир специфичних својстава (предиката) приликом динамичке пројекције RDF структуре базе података и извршавање динамички генерисаних *star-shaped* SPARQL упита. Алгоритам, који је развијен за потребе детекције сличних података, презентује оригиналан приступ и примењује се над инстанцама онтолошких база података. Он користи принципе онтолошког поравнања, рударење текстуалних података, модел векторског простора за математичку репрезентацију података и меру косинусне сличности за нумеричко одређивање сличности података.

Треба напоменути да је Платформа настала као последица вишегодишњег истраживања у оквиру CPCTAS (*Centre for PreClinical Testing of Active Substances*) и Лабораторије за ћелијску и молекуларну биологију као део Института за биологију и екологију Природно-математичког факултета Универзитета у Крагујевцу. Активност Лабораторије покрива једну од важних биоинформатичких подграна - преклиничко тестирање биоактивних супстанци (потенцијалних лекова за канцер). Примарни циљ Платформе је да истраживања у оквиру Лабораторије учини продуктивнијим и ефикаснијим.

Валидација Платформе је спроведена над тестним и реланим биоинформатичким изворима података, указујући на високу искоришћеност ресурса. Захваљујући ефикасним методама Платформе отворен је пут за нова истраживања у области биоинформатике, али и у било којој другој области која покрива онтолошко моделовање података.

Кључне речи

Биоинформатика, семантички веб, онтологије, RDF, SPARQL, Federated SPARQL, семантичка повезаност података, технике онтолошког поравнања, мера косинусне сличности, модел векторског простора, процесирање текстуалних података

Abstract

The importance of bioinformatics, as an interdisciplinary field, is based on a large number of biological data that can be adequately used and processed using current information technology. What is of vital importance in the field of bioinformatics today is the availability of data relevant to the research, as well as the knowledge that such data already exists. An important prerequisite for this is that the necessary data is publicly available, integrated and that mechanisms for their search have been developed. In order to solve these problems, the bioinformatics community uses semantic web technologies. In this respect, many semantic repositories and software solutions have been developed, which have significantly contributed to the research activities in the bioinformatic scene. However, these approaches often face problems because many databases have developed in an isolated environment, without respecting the basic standards of the bioinformatics community. These heterogeneous databases, which links a number of highly specialized and independent resources, often use different conventions, vocabularies and formats for presenting data. Therefore, current software solutions face different challenges in order to search for and discover relevant data. Also, many databases overlap, covering or concealing similar data, thus forming a homogeneous or semi-homogenous data sources. In such cases, the semantic correlation of such databases is often unclear and it is necessary to apply appropriate methods for data analysis, to determine similar data. This dissertation was created as a result of research in order to overcome the shortcomings of existing solutions.

The dissertation presents a contribution to the development of the bioinformatics platform, which presents a number of genuine software approaches that are the basis of key functionalities: executing Federated SPARQL queries over initial (and user selected) databases in order to discover data relevant to bioinformatics research, and the detection of similar data based on determining the semantic relatedness of data. Execution of Federated SPARQL queries is performed over databases that use the *Resource Description Framework* (RDF) as a data model. Query results can be subsequently filtered, thereby contributing to the improvement of their significance. Filtering involves selecting specific properties (predicates) during the dynamic projection of the RDF database structure and executing dynamically generated *star-shaped* SPARQL queries. The algorithm, developed for the detection of similar data, presents the original approach and is applied to instances of ontological databases. It uses the principles of ontological alignment, text data mining, the vector space model for the mathematical representation of data, and the cosine similarity measure for the numerical determination of the similarity of data.

It should be noted that the Platform was the result of long-term research within the CPCTAS (*Center for Pre-Clinical Testing of Active Substances*) Laboratory for Cellular and Molecular Biology as part of the Institute of Biology and Ecology at the Faculty of Science, University of Kragujevac. Laboratory activity covers one of the important bioinformatics subgroups - preclinical testing of bioactive substances (potential drugs for cancer). The primary goal of the Platform is to make Laboratory research more productive and more efficient.

Platform validation was conducted over real and test bioinformatic data sources, indicating high utilization of resources. Thanks to effective Platform methods, a new path for new research in the field of bioinformatics has been opened, but also in any other area that covers ontological data modelling.

Keywords

Bioinformatics, semantic web, ontologies, RDF, SPARQL, Federated SPARQL, semantic relatedness, ontology alignment techniques, cosine similarity measure, vector space model, text data processing

Садржај

Захвалност **4**
Резиме **5**
Кључне речи **5**
Abstract **6**
Keywords **6**
Листа слика **11**
Листа табела **14**
Листа скраћеница **16**
1 Увод **18**
1.1 Циљеви 19
1.2 Полазне хипотезе 20
1.3 Преглед садржаја 20
2 Биоинформатика **22**
2.1 Дефиниција и значење 22
2.2 Гране биоинформатике 22
2.2.1 Базе података и веб сервиси 22
2.2.2 Рационални дизајн лекова 23
2.3 Центар за преклиничка тестирања активних супстанци - CPCTAS 25
2.4 Проблеми у домену биоинформатике 27
2.4.1 Веб и биоинформатика 28
3 Технологије семантичког веба **30**
3.1 Архитектура семантичког веба 30
3.2 Uniform Resource Indentifier - URI 31
3.3 eXtensible Markup Language - XML 31
3.3.1 eXtensible Markup Language Schema - XML Schema 32
3.4 Resource Description Framework - RDF 33
3.4.1 Resource Description Framework Schema - RDF Schema 34
3.5 Онтологије 36
3.5.1 Типови онтологија 36
3.5.2 Компоненте онтолошких модела 36
3.5.3 Конструкција онтологија 37
3.5.4 Ontology Web Language - OWL 37
3.5.4.1 OWL синтакса 37
3.5.4.2 Закључивање 38
3.6 SPARQL Protocol and RDF Query Language – SPARQL 38
3.6.1 SPARQL синтакса 38

3.6.2 SPARQL endpoint 39
3.6.3 Federated SPARQL упити 40
4 Развој PIBAS онтологије, CPCTAS онтолошке базе података и прототипа софтвера за претрагу базе 41
4.1 Онтологије експеримената 41
4.2 Конструкција PIBAS онтологије 42
4.2.1 Таксономија концепата PIBAS онтологије 44
4.3 CPCTAS база података 46
4.3.1 SPARQL упити над CPCTAS базом података 48
4.4 Прототип софтвера за претрагу CPCTAS базе података 49
4.5 Интеграција CPCTAS базе података 51
4.5.1 Биоинформатички репозиторијуми 51
4.5.2 Поступак интеграције 53
4.5.3 Циљеви претрага биоинформатичких база података 54
5 Биоинформатичка платформа - PIBAS FedSPARQL 57
5.1 Принципи развоја Платформе 57
5.2 Архитектура програмског решења 57
5.2.1 Управљачки слој 58
5.2.2 Слој података 59
5.2.3 Слој корисничког интерфејса 63
5.3 Ток рада 66
5.4 Детаљи имплементације програмског решења 66
6 Основне методе Платформе 68
6.1 Креирање и извршавање предефинисаних упита 68
6.1.1 Методологије за креирање упита 69
6.1.2 Процес креирања предефинисаних упита 70
6.1.2.1 Селекција иницијалних упита 71
6.1.2.2 Утврђивање таксономије 73
6.1.2.3 Онтолошко поравнање и селекција извора података 75
6.1.2.4 Модификација иницијалних упита - креирање Federated SPARQL упита 76
6.1.2.5 Процена и вредновање упита 80
6.1.3 Извршавање предефинисаних упита 81
6.1.3.1 Презентација резултата 82
6.2 Динамичко филтрирање резултата упита 84
6.2.1 Звездасти SPARQL упити 85
6.2.2 Презентација резултата 86
6.3 Додавање кориснички селектованог скупа података 87
6.4 Компаративни преглед литературе 88

7 Метода детекције сличних података на Платформи 93
7.1 Детекција сличних података у домену биоинформатике 93
7.1.1 Детекција сличних података над онтолошкима базама података 95
7.2 Онтолошко поравнање 96
7.2.1 Технике онтолошког поравнања 97
7.2.1.1 Терминолошке технике 98
7.2.1.2 Екстензијске технике 100
7.2.1.3 Комбиновање техника и улога експерата 100
7.2.2 Семантичка сличност и семантичка повезаност података 101
7.3 Рударење текстуалних података 103
7.3.1 Процеси рударења текстуалних података 103
7.3.2 Претпроцесирање текста 105
7.4 Преглед литературе 107
7.5 Алгоритам детекције сличних података на Платформи 110
7.5.1.1 Модел детекције сличних података 111
7.5.2 Улазни параметри 112
7.5.3 Утврђивање тежине термина 114
7.5.3.1 Селекција предиката 116
7.5.3.2 Избор терминолошке технике за рачунање сличности 118
7.5.4 Претпроцесирање текстуалних података 119
7.5.5 Трансформација текстуалних у векторске вредности 121
7.5.6 Рачунање сличности 122
7.5.7 Излазни параметри 123
7.5.8 Презентација резултата 124
7.5.9 Математички алати 124
7.5.9.1 Модел векторског простора 124
7.5.9.2 Мера косинусне сличности 125
7.5.9.3 Портеров алгоритам 125
8 Резултати и дискусија 126
8.1 Анализа и дискусија резултата основних метода 126
8.1.1 Компарација резултата основних метода Платформе са резултатима актуелних софтверских решења 137
8.2 Анализа резултата методе за детекцију сличних података 140
8.2.1 Компарација алгоритма за детекцију сличних података на Платформи са алгоритмима актуелних решења 144
8.3 Ограничења Платформе 147
8.4 Дискусија постигнутих резултата 148
9 Закључак 149

Додатак **151**
Библиографија **152**
Биографија **167**

Листа слика

Слика 2.1 Хијерархијска структура експеримената који се изводе у оквиру Лабораторије 26

Слика 2.2 Хронолошки развој веба (извор: веб) 28

Слика 3.1 Архитектура семантичког веба (извор: веб) 30

Слика 3.2 Пример XML документа са елементима који представљају подсегмент структуре експеримената (Слика 2.1) 32

Слика 3.3 Пример коришћења именских простора у XML документу који представља подсегмент структуре експеримената (Слика 2.1) 32

Слика 3.4 Пример XMLS документа са елементима који представљају подсегмент структуре експеримената (Слика 2.1) 33

Слика 3.5 Графичка репрезентација RDF изјаве (pibas:Animals, pibas:isTypeOf, pibas:ModelSystem) 34

Слика 3.6 Репрезентација RDF изјаве (pibas:Animals, pibas:isTypeOf, pibas:ModelSystem) у форми XML-а 34

Слика 3.7 RDFS документ (а) и одговарајући графички приказ (б) са елементима који представљају подсегмент структуре експеримената (Слика 2.1) 35

Слика 3.8 Пример SELECT SPARQL упита чији резултат извршавања чине инстанце класе pibas:modelSystem дефинисане у PIBAS онтологији 39

Слика 3.9 Пример Federated SPARQL упита помоћу кога се врши претрага лекова (у DrugBank/Bio2RDF и Chembl/EMBL-EBI базама података) који делују на биолошку мету издвојену као резултат тестирања активне супстанце у CPCTAS бази података 40

Слика 4.1 Део таксономије концепата EXPO онтологије 43

Слика 4.2 Основна таксономија концепата PIBAS онтологије 45

Слика 4.3 Приказ експеримента (инстанце) из Protégé едитора (лево) у корелацији са екстерним онтолошким фајловима (ModelSystem.owl, ActiveSubstance.owl, Protocol.owl и expHr137.owl) у CPCTAS бази података 48

Слика 4.4 Декомпозиција CPCTAS базе података 48

Слика 4.5 Пример SPARQL упита дефинисаног над CPCTAS базом података чији резултат извршавања чине експерименти који задовољавају услове дефинисане у WHERE клаузули 49

Слика 4.6 Кориснички интерфејс прототипа софтвера за претрагу CPCTAS базе података 50

Слика 4.7 Резултат поређења онтолошких база података Drugbank/Bio2RDF и PIBAS/CPCTAS применом OWLDiff алата Protégé едитора 53

Слика 4.8 Графички приказ интеграције PIBAS/CPCTAS базе података са DrugBank/Bio2RDF базом података 54

Слика 5.1 Архитектура Платформе 58

Слика 5.2 Основна таксономија концепата DataSources онтологије 60

Слика 5.3 Иницијални кориснички интерфејс за извршавање предефинисаних упита на Платформи 64

Слика 5.4 Табеларни приказ резултата извршавања предефинисаног упита на Платформи 64

Слика 5.5 Квантитативни приказ резултата извршавања предефинисаног упита на Платформи 64

Слика 5.6 Форма за додавање кориснички селектоване базе података на Платформи 65

Слика 5.7 Пример панела PIBAS/CPCTAS базе података за селектовање предиката за примену методе

динамичког филтрирања резултата упита 66

Слика 6.1 SPARQL упити за претрагу биолошких мета (а), есеја и ћелијских линија (б) у Chembl/EMBL-EBI бази података. Променљива chemblID представља ID лека (активне супстанце) 72

Слика 6.2 SPARQL упит за претрагу биолошких мета у BindingDB/Chem2Bio2RDF бази података. Променљива ?cid представља ID лека (активне супстанце) 72

Слика 6.3 SPARQL упити за претрагу биолошких мета у Drugbank/Bio2RDF бази података (а) и публикација у PubMed/Bio2RDF бази података (б). Променљива ?compound_uri представља URI спецификацију лека (активне супстанце). Променљива ?author_name представља име аутора 72

Слика 6.4 Приказ класа првог хијерархијског нивоа Chembl/EMBL-EBI базе података добијен применом софтверског решења представљеног у [108] 73

Слика 6.5 Примери SPARQL упита за откривање InChiKey и SMILES параметара. Упит (а) се извршава над инстанцом типа chembl:SmallMolecule, а упит (б) над инстанцом типа bindingdb:bindingdb_ligand 75

Слика 6.6 Предефинисани Federated SPARQL упити који се користе за откривање биолошких мета (а), есеја (б), ћелијских линија (ц), лекова (д) и публикација (е) на Платформи 79

Слика 6.7 SELECT SPARQL упит за преузимање објектне вредности предиката pibas:hasInitialQuery из DataSources онтологије за селектовани шаблон на Платформи. Променљива $templateid означава ID шаблона 81

Слика 6.8 Делимични приказ JSON објекта добијеног извршавањем предефинисаног упита за откривање биолошких мета које су у интеракцији са леком Fluorouracil на Платформи 82

Слика 6.9 Делимични приказ резултата извршавања предефинисаног упита за откривање биолошких мета које су у интеракцији са леком Fluorouracil на Платформи 82

Слика 6.10 Делимични приказ дескрипције drugbank:BE0000324 инстанце 83

Слика 6.11 Кориснички панели методе динамичког филтрирања резултата предефинисаног упита за откривање биолошких мета које су у интеракцији са леком Fluorouracil на Платформи 84

Слика 6.12 Пример SPARQL упита за добијање предиката за попуњавање PIBAS/CPCTAS панела за примену методе динамичког филтрирања резултата на Платформи 85

Слика 6.13 Примери звездастих SPARQL упита за примену методе динамичког филтрирања резултата предефинисаних упита за откривање биолошких мета које су у интеракцији са леком Fluorouracil на Платформи 86

Слика 6.14 Резултат примене методе динамичког филтрирања резултата за селектовани предикат pibas:targetType из PIBAS/CPATAS панела и chembl:targetType из Chembl/EMBL-EBI панела 86

Слика 6.15 Форма за додавање кориснички селектоване базе података на Платформи попуњена тест подацима 87

Слика 6.16 Резултат извршавања методе предефинисаног упита са кориснички селектованом базом података за откривање биолошких мета које су у интеракцији са леком Fluorouracil на Платформи 88

Слика 6.17 Резултат извршавања методе динамичког филтрирања резултата са кориснички селектованом базом података за откривање биолошких мета које су у интеракцији са леком Fluorouracil Платформи 88

Слика 7.1 Шема класификације техника поравнања на основу врсте улазних података [140] 98

Слика 7.2 Модел детекције сличних података предложен на Платформи 112

Слика 7.3 Део кода алгоритма за детекцију сличних података који извршава RunningSparql оператор у циљу креирања корпуса (текстуалних датотека) за улазне параметре ?instance и ?instance_endpoint

..... 113

Слика 7.4 Део кода алгоритма за детекцију сличних података за рачунање tf, idf и tf-idf тежинских мера 115

Слика 7.5 Део кода алгоритма за детекцију сличних података који извршава TextProcessData оператор у циљу претпроцесирања текстуалних података 120

Слика 7.6 Део кода алгоритма за детекцију сличних података који извршава VectorTransforming оператор који обавља процес трансформације текстуалних података у векторске вредности..... 121

Слика 7.7 Део кода алгоритма за детекцију сличних података који извршава DataSimilarity оператор који обавља израчунавање CSM-а. Функција get_cosine имплементира формулу (4) 122

Слика 7.8 Резултат примене методе детекције сличних података на Платформи за улазне параметре pibas:TestTarget1, drugbank:BE0000324, kegg:5f47d0b54b4d81097410bcc4cf01cf71, chembl_target:3160 и chembl_target:CHEMBL1075416..... 124

Слика 8.1 Резултат извршавања предефинисаног упита за откривање информација о леку Fluorouracil на Платформи 127

Слика 8.2 Резултат извршавања методе динамичког филтрирања резултата упита за откривање информација о леку Fluorouracil на Платформи (за селектоване предикате drugbank_vocabulary:calculated-properties и pibas:sameAs)..... 128

Слика 8.3 Резултат примене методе динамичког филтрирања резултата упита за откривање информација о тест супстанци 2 на Платформи (за селектовани предикат chembl_molecule:Xref) 129

Слика 8.4 Резултат примене методе динамичког филтрирања резултата упита за откривање биолошких мета које су у интеракцији са леком Fluorouracil на Платформи (за селектовани предикат chembl:taxonomy) 131

Слика 8.5 Резултат примене методе динамичког филтрирања резултата упита за откривање биолошких мета које су у интеракцији са тест супстанцом 2 на Платформи (за селектовани предикат chembl_target:targetType)..... 132

Слика 8.6 Резултат примене методе динамичког филтрирања резултата упита за откривање есеја који су у интеракцији са леком Fluorouracil на Платформи (за селектовани предикат dcterms:description), филтриран по кључној речи MTT 133

Слика 8.7 Резултат примене методе динамичког филтрирања резултата упита за откривање есеја који су у интеракцији са тест супстанцом 2 на Платформи (за селектовани предикат dcterms:description), филтриран по кључној речи DNA 133

Слика 8.8 Резултат примене методе динамичког филтрирања резултата упита за откривање ћелијских линија које су у интеракцији са леком Fluorouracil на Платформи (за селектоване предикате rdfs:label и dcterms:description), филтриран по кључној речи L1210..... 134

Слика 8.9 Резултат примене методе динамичког филтрирања резултата упита за откривање ћелијских линија које су у интеракцији са тест супстанцом 2 на Платформи (за селектовани предикат chembl:isCellLineForAssay), филтриран по кључној речи CHEMBL3254890 135

Слика 8.10 Резултат примене методе динамичког филтрирања резултата упита за откривање публикација које су у интеракцији са леком Fluorouracil на Платформи (за селектовани предикат dc:tittle), филтриран по кључној речи cancer 135

Слика 8.11 Резултат примене методе динамичког филтрирања резултата упита (за селектоване предикате drugbank_vocabulary:x-uniprrot и testOntology:hasSynonim) након додавања кориснички селектоване базе података за откривање биолошких мета које су у интеракцији са леком Fluorouracil на Платформи 137

Листа табела

Табела 2.1 Значења неких концепата хијерархијске структуре експеримената (извор [36]) 26

Табела 4.1 Домени (rdfs:domain) и кодомени (rdfs:range) објектних својстава (object property) и својстава типа података (datatype property) дефинисаних у PIBAS онтологији 46

Табела 4.2 Новодефинисане класе, објектна својства (object property) и својства типа података (datatype property) у CPCTAS бази података 47

Табела 5.1 Релације између класа pibas:Topics, pibas:SubTopic и pibas:Templates у DataSources онтологији 61

Табела 5.2 Преглед инстанци свих класа за предефинисне упите у DataSources онтологији 61

Табела 5.3 Преглед инстанци класе pibas:Templates са приказом релевантних предиката у DataSources онтологији 62

Табела 6.1 Број класа првог хијерархијског нивоа потенцијалних кандидата (база података) за креирање предефинисаних Federated SPARQL упита за шаблоне (*) 74

Табела 6.2 Број инстанци класа које припадају потенцијалним кандидатима (базама података) за креирање предефинисаних упита за шаблоне (*) 74

Табела 6.3 Валидација резултата извршених предефинисаних упита 80

Табела 7.1 Примери стандардизованих и кориснички дефинисаних предиката Bio2RDF, Chem2Bio2RDF, EMBL-EBI и CPCTAS репозиторијума који су типа података (datatype property) и чији су кодомени стринг подаци (xsd:string) 97

Табела 7.2 Преглед формула неких од терминолошких техника онтолошког поравнања (извор [141]) 99

Табела 7.3 Резултат рада RunningSparql оператора за улазне параметре (биолошке мете) у циљу формирања корпуса 113

Табела 7.4 Резултат рада TF-IDFMeasure оператора у циљу формирања words листе 115

Табела 7.5 Резултат рада PredicateSelection оператора у циљу селекције предиката (selected_predicates) 117

Табела 7.6 Резултати примене терминолошких техника над тестним паровима у циљу селекције најадекватније технике 119

Табела 7.7 Резултат рада TextProcessData подоператора у циљу претпроцесирања текстуалних података 120

Табела 7.8 Делимични приказ резултата рада VectorTransforming оператора 121

Табела 7.9 Резултат рада DataSimilarity оператора 122

Табела 7.10 Резултат рада SortingResult оператора у циљу одређивања излазних параметара 123

Табела 8.1 Компарација резултата рада методе извршавања предефинисаних упита на Платформи са резултатима рада софтверских решења Open PHACTS [118], BioSearch [111] и BioCarian [116] 138

Табела 8.2 Резултати валидације методе за детекцију сличних података над улазним параметрима који су добијени као резултат извршавања предефинисаних упита 140

Табела 8.3 Резултати валидације алгоритма за детекцију сличних података за улазне параметре селектоване на основу релација у PubChem, UniProt, UniChem и canSAR базама података 142

Табела 8.4 Резултати примене алгоритма за детекцију сличних података за два основна приступа: са селекцијом и без селекције предиката 143

Табела 8.5 Резултати примене алгоритма за детекцију сличних података за улазне параметре селектоване из ChEMBL/EMBL-EBI базе података и различите вредности прага сличности (th)..... 144

Табела 8.6 Преглед резултата рада алгоритма за детекцију сличних података представљеног у дисертацији и алгоритама представљених у истраживањима [131,212,205,198]..... 145

Листа скраћеница

CPCTAS	*Centre for PreClinical Testing of Active Substances*
NCBI	*National Center for Biotechnology Information*
EMBL	*European Molecular Biology Laboratory*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
UniProt	*Universal Protein Resource*
IC_{50}	*Half Maximal Inhibitory Concentration*
EMA	*European Medicines Agency*
FDA	*Food and Drug Administration*
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
URN	*Uniform Resource Notation*
LOD	*Linked Open Data*
MGED	*Microarray Gene Expression Data*
MSI	*Chemical Methods Ontology*
CHMO	*Metabolomics Standards Initiative*
OBI	*Ontology for Biomedical Investigations*
XML	*eXtensible Markup Language*
XML Schema (XMLS)	*eXtensible Markup Language Schema*
RDF	*Resource Description Framework*
RDF Schema (RDFS)	*Resource Description Framework Schema*
OWL	*Ontology Web Language*
SPARQL	*SPARQL Protocol and RDF Query Language*
MGED	*Microarray Gene Expression Data*
MSI	*Metabolomics Standards Initiative*
LCT	*Linked Clinical Trials*
HCLS IG	*W3C Semantic Web Health Care and Life Sciences Interest Group*
UCV	*UniProt Core Vocabulary*
CCO	*ChEMBL Core Ontology*
BAO	*BioAssay Ontology*
CHEMINF	*Chemical Infomation Ontology*
UO	*Unit Ontology*
SIO	*Semantic Science Ontology*
LODD	*Linking Open Drug Data*
GFBio	*The German Federation for Biological Data*
UMLS	*Unified Medical Language System*
CSM	*Cosine Similarity Measure*
VSM	*Vector Space Model*
DC	*Dice coefficient*
JM	*Jaccard measure*
LD	*Levenshtein distance*
NLP	*Natural Language Processing*
IR	*Information Retrieval*
IE	*Information Extraction*
DM	*Data Mining*

OAEI	*Ontology Alignment Evaluation Initiative*
LCS	*Lowest Common Subsumer*
IC	*Information Content*
NER	*Named Entity Recognition*
TF	*Term Frequency*
IDF	*Inverse Document Frequency*

1 Увод

Током неколико последњих деценија убрзан развој биолошких, хемијских и технологија молекуларног истраживања, допринели су стварању велике количине података. „*Заједно са великим бројем података о структури протеина, експресији гена, лекова и других типова података настала је и велика разноврсност биолошких информација*“ [1]. У циљу презентације, интеграције, обраде и анализе ових података рачунари су заузели примарно место, па је отуда и настала потреба за биоинформатиком. У савременој литератури евидентирано је много различитих дефиниција биоинформатике, а једну од најсажетијих формулисао је Национални центар за биотехнолошке информације (*National Center for Biotechnology Information*[1] - NCBI): „*Биоинформатика је поље науке у којој су се биологија, рачунарска наука и информационе технологије спојиле у једну дисциплину*“ [2]. „*Биоинформатика представља назив којим се описују математички и информатички приступи коришћени у циљу потпунијег разумевања биолошких процеса. Основни циљ биоинформатике је повећање разумевања биолошких процеса. Оно што је издваја од других процеса је фокус на развој и примену информатички интензивних техника за постизање тог циља. Биоинформатика данас подразумева стварање и развој база података, алгоритама, информатичких и статистичких техника, као и теоријске основе за решавање формалних и практичних проблема који се јављају у управљању и анализи биолошких података*“ [3].

Истраживачи у биоинформатичком домену могу да изводе закључке и доносе конкретне одлуке неопходне за даља истраживања, ако су у стању да на једноставан начин приступају свим релевантним подацима. Овај циљ је једино могућ ако су подаци јавно доступни и интегрисани. Већ на самом почетку развоја биоинформатике уочена је предност интегрисаних података и многе базе које се развијале у овом домену настојале су да задовоље критеријуме доступности и интегрисаности [4]. Ово правило се активно примењује и данас, што је свакако допринело убрзаном и успешном развоју биоинформатике. Ипак, мора се узети у обзир и чињеница да су се многе базе развијале у изолованом окружењу, не поштујући принципе биоинформатичке заједнице. Ове хетерогене базе, које су у саставу многих високо специјализованих и независних ресурса, често користе различите конвенције, формате и вокабуларе за представљање интерних података. Често многи мулти-дисциплинарни пројекти у биоинформатичком домену спроводе истраживања над оваквим типовима података и њихова финална реализација и коначни успех у великој мери зависе од њихове доступности и употребе. Ипак, рапидна и константна акумулација података постала је примарни изазов са којим се биоинформатичка заједница суочава у циљу откривања и стицања нових знања [5]. Додатно, кључан проблем је и непостојање добро дефинисаних стандарда, односно модела неопходних за егзактно представљање података [5]. Њихово одсуство експлицитно утиче на процес интеграције и ефикасне претраге података. У циљу решавања наведених проблема биоинформатичка заједница је усвојила технологије семантичког веба [4]. Семантички веб [6] представља трећу генерацију веба и надограђује се на претходне две генерације. Он омогућава прецизно дефинисање појмова и знања на начин адекватан за аутоматско процесирање.

Многе научне институције и истраживачке организације су усвојиле семантичке стандарде и допринеле својим конструктивним решењима у домену биоинформатике. Технологије семантичког веба решавају актуелне проблеме хетерогених база података дефинишући стандарде који олакшавају процесе имплементације, интеграције и аутоматског откривања биоинформатичких података [4]. Као последица тога развијен је велики број софтверских решења (представљених у одељку 6.4) која обезбеђују разумљив кориснички интерфејс за претрагу, визуелизацију и анализу података у различитим биоинформатичким областима. Ове функционалности су доступне захваљујући комбинацији онтологија, SPARQL упита и различитих програмских алата за визуелизацију података. Ипак, постојећа софтверска решења су често ограничена јер не подржавају могућност проширења постојећих упита на претрагу нових база података, не пружају приступ ажурним подацима, не нуде могућност дефинисања нових упита и подршка за претрагу података је лимитирана на употребу одређених команди, у смислу да је

[1] *https://www.ncbi.nlm.nih.gov/*

могуће креирати искључиво једноставне SPARQL упите. Како би се делимично отклонили наведени проблеми, развијена је Платформа која је представљена у дисертацији. Она настоји да применом предефинисаних Federated SPARQL упита који приступају *remote endpoint*-има обезбеди корисницима ажурне податке. Такође, предефинисани упити се могу проширити на претрагу нових база података на захтев корисника. На тај начин се омогућава популаризација мање познатих база и откривање комплементарног знања. Додатно, Платформа се увек може проширити новим предефинисаним упитима који би били од важности за разна биоинформатичка истраживања. Као новитет, Платформа нуди и опцију детекције сличних података добијених извршавањем предефинисаних упита, што је постигнуто комбинацијом SPARQL упита и одређених информатичких, односно математичких приступа. Резултати примене ове функционалности могу утицати на реализацију будућих експерименталних приступа у смислу уштеде ресурса или као потврда кориснику да се његова истраживања одвијају у добром смеру.

Главни предмет истраживања дисертације јесте биоинформатичка платформа[2] [7], која својим методама олакшава истраживања у домену биоинформатике, користећи предности семантичких технологија. Методе које су имплементиране на Платформи имају за циљ да обезбеде извођење претрага семантичких репозиторијума, које уз одговарајуће софтверске приступе омогућавају откривање различитих информација неопходних за планирање биоинформатичких истраживања. Једна од значајних метода Платформе јесте и детекција сличних података, која има за циљ да истраживања учини још ефикаснијим и продуктивнијим откривањем сличних (и идентичних) података. Платформа је првенствено развијена да подржи сакупљање знања у процесу дизајна лекова (преклиничком тестирању активних супстанци), као једној од кључних грана биоинформатике, али се може применити и у било којој другој области која подржава онтолошке базе података.

1.1 Циљеви

Циљ дисертације јесте побољшање и потпора истраживању у домену биоинформатике, коришћењем развијене Платформе, базиране на семантичким технологијама, за откривање значајних и комплементарних информација у рационалном дизајну лекова. Аутоматизованом претрагом више различитих биоинформатичких семантичких репозиторијума, директно се користи огроман потенцијал ових извора на начин који истраживачи дефинишу сагласно захтевима својих истраживања. Конкретно, могуће је открити биолошке мете[3] (биолошке тест системе) који се користе за експериментална истраживања, као и дефинисање молекуларних механизама биолошког деловања потенцијалног лека, односно одговарајуће активне супстанце. Ове мете, које могу припадати различитим базама података, могу имати епитет „повољних“, односно могу означавати оне које су коришћене у успешно спроведеним експериментима и које потенцијално могу имати позитиван ефекат на даљу активност у процесу преклиничког тестирања лекова. Успешност експеримента се процењује на основу цитотоксичности одређене активне супстанце, а одређује помоћу IC_{50} вредности, која представља концентрацију тестиране активне супстанце која индукује 50% смртности ћелија. У општем смислу, овај квантитативни показатељ је индикатор мере у којој је тестирана супстанца неизоставна да се дати биолошки процес инхибира за половину [8]. Биолошке мете које имају епитет „повољних“ могу се искористити за будућа експериментална истраживања. Шанса успешности извођења таквих експеримената сигурно је већа од оних са насумично одабраним метама. На сличан начин могуће је извршити селекцију ћелијских линија или есеја (тестова). Биоинформатички семантички репозиторијуми омогућавају откривање информација и о одређеним активним супстанцама (лековима), а нуде и информације о литератури, која је од потенцијалне важности за неко биоинформатичко истраживање.

Релевантна функционалност Платформе јесте и метода детектовања сличних података, која се заснива на утврђивању семантичке повезаности података. Алгоритам који је развијен за потребе ове функционалности представља иновативан приступ за детекцију сличних података (инстанци) над

[2] У даљем тексту користи се термин *Платформа*.

[3] Најчешће коришћен термин за биолошке мете у српској стручној литератури је *таргет*. У даљем тексту се под биолошким тест системом подразумева *биолошка мета* (*таргет*) и не прави се разлика између ових термина.

онтолошким базама података. Алгоритам користи принципе онтолошког поравнања[4] (енгл. *ontology alignment; ontology matching*), методе текстуалног рударења података (енгл. *text-data mining*), модел векторског простора (енгл. *vector space model*) за математичку репрезентацију података и меру косинусне сличности (енгл. *cosine similarity measure*) за нумеричко одређивање сличности података. Утврђивање сличних података може утицати на планирање наредних експерименталних приступа у домену биоинформатике. Конкретно, када се утврди група биолошких мета (ћелијских линија или есеја), применом алгоритма могу се издвојити слични ентитети, иако они могу припадати различитим базама података, односно могу имати различите URI (*Uniform Resource Identifier*) [9] спецификације. На тај начин се сужава претрага података и штеде ресурси који су неопходни за извођење будућих експеримената. Имплементирани алгоритам се може применити не само у биоинформатичком, него и у било ком другом домену који подржава онтолошко моделовање података.

1.2 Полазне хипотезе

Полазну основу дисертације представљају актуелна истраживања на пољу примена технологија семантичког веба у развоју биоинформатичких софтверских приступа. Основне хипотезе истраживања [10] су да технологије семантичког веба:

- нуде богате и добро дефинисане моделе за представљање и интеграцију података;
- омогућавају агрегацију хетерогених података коришћењем експлицитне семантике;
- олакшавају претрагу података;
- омогућавају поновну употребу података у циљу извођења закључака круцијалних за даља биоинформатичка истраживања.

Развој метаподатака (енгл. *metadata*) у области биологије изузетно је важан за употребу семантичких технологија у домену биоинформатике, јер омогућава семантичку интеграцију података коришћењем RDF (*Resource Description Framework*) [11] оквира као основе за семантичко представљање података. Онтологије [12] се као формални модели за експлицитно представљање концепата и релација користе за решавање проблема хетерогености у изворима података. Упитни језик SPARQL (*SPARQL Protocol and RDF Query Language*) [13] као стандардни упитни језик за RDF, омогућава извршавање глобалних упита над различитим изворима података који могу бити дистрибуирани и третирани као велика семантичка база података на вебу. Генерално, технологије семантичког веба олакшавају процес представљања, интеграције, поновне употребе и претраге података у домену биоинформатике [10]. Употреба семантичких технологија олакшава проналажење релевантних података, са могућношћу примене различитих приступа и метода за даљу анализу и процесирање података. Ово свакако може допринети квалитетнијим биоинформатичким истраживањима.

1.3 Преглед садржаја

Дисертација се састоји из девет поглавља.

Друго поглавље уводи појам биоинформатике, представља њен значај у домену савременог истраживања, као и изазове са којима са биоинформатичка заједница сусреће. У овој целини укратко су представљене и неке од најважнијих грана биоинформатике - рационални дизајн лекова и базе података. У овом поглављу је представљен и концепт веба, хронолошки развој веба и његов утицај на домен биоинформатике. У склопу ових истраживања презентована је и основна делатност Лабораторије за ћелијску и молекуларну биологију[5] (као део Института за биологију и екологију Природно-математичког факултета, Универзитета у Крагујевцу) која је усвојила принципе семантичких технологија у процесу преклиничког тестирања активних супстанци.

[4] *Технике поравнања* се у српском језику називају и *техникама упоређивања* или *мечирања онтологија*. У неким другим језицима често се користе и термини *поравнавања* или *упоређивања*.

[5] У даљем тексту се користи термин *Лабораторија* (уколико није другачије наглашено).

Тема **Трећег поглавља** оријентисана је на трећу генерацију веба - семантички веб. У складу са тим, представљањена је архитектура семантичког веба, а потом и анализа његових технологија које се активно промовишу у домену биоинформатике. Овај сегмент дисертације се може користити и као водич за примену семантичких технологија у различитим доменима. У овом поглављу је представљен и потенцијал који технологије семантичког веба генерално нуде у циљу решавања биоинформатичких проблема - представљања и претраге података.

У **Четвртом поглављу** су детаљно објашњене фазе истраживања које су претходиле развоју Платформе. Најпре је представљена PIBAS онтологија, чија је намена моделовање структуре варијетета експеримената који се обављају у Лабораторији дефинисањем релација и веза између концепата (биолошких термина) који карактеришу разноврсне експерименталне процедуре. Затим, представљен је развој CPCTAS онтолошке базе података Лабораторије, као и функционалност прототипа софтвера за претрагу дате базе, који је претходио развоју Платформе. У овом поглављу су представљени и истраживачки захтеви, који доприносе откривању релевантних информација у процесу преклиничког тестирања активних супстанци, а које Платформа подржава.

У **Петом поглављу** су представљени принципи развоја Платформе, архитектура са детаљним приказом компоненти и начином њиховог функционисања, као и детаљи имплементације.

Основне методе Платформе су представљене у **Шестом поглављу**. Детаљно је представљено извршавање предефинисаних упита над иницијалним и кориснички селектованим базама података, као и динамичко филтрирање резултата упита. За сваку од метода је приказан начин примене као и интерпретација резултата. Посебна пажња је усмерена на опис методологије која је у овом истраживању искоришћена за креирање предефинисаних упита. У склопу истраживања указује се и на комплексност ентитета и релација у онтолошким биоинформатичким базама података, као и изазовима који се појављују у оквиру њихових анализа и претрага, а у циљу креирања SPARQL упита.

Главни допринос који доноси ова дисертација јесте метода детектовања сличних података заснована на утврђивању семантичке повезаности између података, која је описана у **Седмом поглављу**. Један део овог поглавља посвећен је студији литературе са циљем утврђивања чињеничног стања овог присутног проблема у домену биоинформатике. У овом поглављу врши се анализа и поређење термина семантичке сличности и семантичке повезаности података. Такође, дат је преглед информатичких приступа и математичких алата, који су коришћени за дефинисање алгоритма. Са тим циљем представљене су технике онтолошког поравнања, технике текстуалног рударења података, основне процедуре претпроцесирања текстуалних података, а затим и модел векторског простора коришћен за математичку репрезентацију података, као и мера косинусне сличности за нумеричко одређивање сличности података.

Осмо поглавље представља дискусију и анализу резултата актуелних метода Платформе над реалним и тестним примерима. У циљу верификације, резултати су анализирани у сарадњи са истраживачима Лабораторије. У овом поглављу су представљена и ограничења Платформе.

Анализа доприноса Платформе, као и предлози за побољшање и даљи развој метода и самог софтверског решења дата су у **Деветом поглављу**.

2 Биоинформатика

У овом поглављу је представљен концепт биоинформатике (основни задаци и циљеви) и најзначајније гране (области) биоинформатике, са акцентом на развој (и одржавање) биолошких база података и рационални дизајн лекова. У овом поглављу је представљена и област рада Лабораторије, која покрива једну од подграна у рационалном дизајну лекова – преклиничко тестирање биоактивних супстанци. У овом сегменту су представљени и проблеми са којима се Лабораторија суочава у процедури представљања интерних биоинформатичких података. Такође, образложен је концпет веба, хронолошки развој веба, однос веба и биоинформатике, као и потенцијал који технологије семантичког веба генерално нуде у циљу решавања биоинформатичких проблема.

2.1 Дефиниција и значење

„*Биоинформатика је мултидисциплинарна наука која интегрише развој информатичких и рачунарских наука примењених у биотехнолошким и биолошким наукама*“ [1]. Биоинформатика је тренутно у фази велике експанзије и развоја. Као високо интердисциплинарна наука, биоинформатика не само да користи технике и концепте информатике, већ и статистике, математике, хемије, биологије, биохемије, физике па и лингвистике [1]. Циљ биоинформатике је да изведе знање из биолошких података користећи рачунарске анализе, при чему су подаци најчешће информације записане у генетском коду, експериментални резултати из разних (онтолошких) извора, статистике пацијената и стручна литература [14]. Уопштено, биоинформатика се може схватити као информациони систем управљања у молекуларној биологији који има многе практичне примене [15].

2.2 Гране биоинформатике

Постоје различити типови података које биоинформатика обрађује, а три примарна су: нуклеотидне и протеинске секвенце, структуре макромолекула и резултати експеримената [1]. У складу са тим, биоинформатика се дели на неколико основних грана (области) [16]:

- **Геномика** - грана која се бави секвенцирањем, мапирањем, анализом функција и структуре генома[6];
- **Протеомика** - грана која се бави анализом функција и структуре протеома[7];
- **Рационални дизајн лекова** (енгл. *rational drug design*) - грана која се бави инвентивним процесом откривања или дизајнирања нових медикамената;
- **Базе биолошких података и анализа података** (енгл. *bio data bases and data mining*) - грана која се бави развојем и одржавањем база података, као и развојем нових метода за анализу података;
- **Молекуларна филогенеза** - грана која одређује квантитативан критеријум за класификацију организама преко молекуларне, односно биоинформатичке анализе протеинских и нуклеотидних секвенци;
- **Системска биологија**

С обзиром на релевантност у овом истраживању, у наставку су концизније размотрене базе података (укључујући и веб сервисе), као и рационални дизајн лекова.

2.2.1 Базе података и веб сервиси

Биоинформатичко истраживање се углавном завршава одређеним резултатима: сазнањима и подацима, који сами по себи имају своју тежину. Међутим, ови подаци се вреднују тек када се упореде са

[6] Геном је комплетан сет наследне информације неког организма; обухвата укупну ДНК хаплоидносг сета хромозома, а код еукариота и геноме органела [233].

[7] Протеом је комплетан сет протеина присутан у некој ћелији у одређеном тренутку [233].

резултатима других истраживања и када се могу обрадити тако да се из њих извлаче квалитетни и аргументовани закључци. Због тога се ови подаци, као и све остале информације које су битне за истраживање, чувају у базама података. Задатак биоинформатике је да развије нове и да притом одржава старе базе података које садрже такве информације. Поред овога, задатак биоинформатике је да развије нове начине и методе који утичу на складиштење, анализу и дељење података.

С обзиром на типове података које се користе у домену биоинформатике, може се закључити да постоји велика разлика у величини и комплексности база података. Европска лабораторија за молекуларну биологију је 1980. године основала EMBL (*European Molecular Biology Laboratory*) [17] базу података, односно библиотеку нуклеотидних секвенци. Једна од првих база нуклеотидних секвенци која је 1982. године добила статус јавне базе података је GenBank [18]. Прва база протеинских секвенци, PIR (*Protein Information Resource*) [19], креирана је 1984. године. Године 1986. креирана је Swiss-Prot [20] база протеинских секвенци на Швајцарском институту за биоинформатику. Године 1991. основана је GenomeNet мрежа, која садржи KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [21] базу података (подаци из области генома, ензимских путева и биолошких једињења), као и везе ка другим јавним базама као што су PubChem [22] (подаци молекула са особинама сличним лековима) и PubMed [23] (библиографски подаци). Године 2002. креирана је UniProt (*Universal Protein Resource*) [24] база - јединствена светска база протеинских секвенци и функција.

Због проблема редундантности и многострукости података, као и интеграције међу подацима датих база, често је била потребна организација информација великих размера. Како постоје разни извори и врсте информација које је потребно комбиновати, било је неопходно развити одређене сервисе, који омогућавају интегрисан приступ кроз више извора података и разне упите [1]. Један од најпопуларнијих веб сервиса је ENTREZ [25], део NBCI центра, који поседује једноставан интерфејс за флексибилно и прецизно претраживање нуклеотидних секвенци, малих молекула, генома, гена, експресије гена, протеина и ћелијских сигналних путева. Други, такође популаран систем за приступ базама је DBGET [26], који приступа мрежи база GenomeNet. NCBI је развио сервис IVR (*Influenza Virus Resource*) [27] који пружа јавну услугу претраге и анализе података из Америчког националног института за алергије и инфективне болести (енгл. *National Institute of Allergy and Infectious Diseases*) и GenBank базе података.

Са раним сазревањем биоинформатике настављало се интензивно усавршавање постојећих и развој нових база података. Многе базе су се константно ажурирале или интегрисале са другим базама података. Интеграција је свакако била примарни циљ, јер је било све више повезаних података, који су имали утицај у откривању знања. Такође, број софтверских решења (сервиса) који користе дате базе био је у сталном порасту и њихова намена и начин коришћења зависио је од специфичних корисничких захтева. Биолошке базе података су своју велику примену нашле у области рационалног дизајна лекова.

2.2.2 Рационални дизајн лекова

Дизајн и откривање лекова једна је од виталних грана биоинформатике. Развој лека је скуп и временски захтеван процес, због чега је битан развој методолошких приступа који га могу убрзати и појефтинити [1]. Идеја која стоји иза процеса дизајна лекова је да се помоћу познатих структура идентификује биолошка мета и предвиди начин везивања потенцијалног лека за биолошку мету [28]. Помоћу ових сазнања могу се предвидети структуре нових молекула које се чвршће везују и успешније моделују активност мете. Због тога је један од најзначајнијих корака у развоју новог лека идентификација и валидација биолошке мете. „*Биолошка мета, циљно место или таргет је биополимер, протеин*“ [1] или нуклеинска киселина. Протеинске биолошке мете могу бити ензими, G протеин-спрегнути рецептори, јонски канали, мембрански транспортери, нуклеарни хормонски рецептори или структурни протеини [29]. Квалитетна биолошка мета мора имати карактеристику лековитости, ефикасности, безбедности и требало би да задовољи клиничке и комерцијалне потребе [28]. Компетенција „лековитости“ подразумева да биолошка мета мора бити доступна предложеном леку и да мерљива биолошка реакција мора бити изазвана као последица интеракције лека и биолошке мете [28]. Биолошка мета у интеракцији са леком мења своју активност и тиме утиче на спречавање болести, чиме се постиже терапеутски ефекат [1]. Идентификација мете омогућава повећан кредибилитет у односу саме мете и болести. „*Циљ*

идентификације биолошке мете је разумевање биолошких процеса везаних за болест и идентификација механизама и структура појединачних елемената болести“ [1]. Процес идентификације раног тестирања терапеутских мета траје у просеку од 12 до 15 година и самим тим подразумева велики утрошак ресурса [1]. „*Зато је веома битно разумети процес међу-молекулских интеракција, који је основа идентификације протеинских и нуклеотидних мета и предикције интеракција лек-протеин и лек-нуклеотид, што би омогућило скраћење периода и смањење трошкова процеса*“ [1]. Валидација биолошких мета подразумева верификацију њених особина. Особине се анализирају *in vitro* (на ћелијским културама) и *in vivo* моделима (на моделима болести код животиња) [30]. Биолошка мета је успешно прошла верификацију када одређено деловање на њу покаже адекватне ефекте на моделу. Валидација биолошких мета може имати позитиван ефекат на даљу активност у процесу преклиничког тестирања лекова, јер када је биолошки систем валидиран, он се може укључити у додатна истраживања, која могу открити неке нове хемијске супстанце или нове компоненте које би могле модификовати активност мете [30]. Овим се може потврдити новитет и комерцијална способност биолошке мете.

Након овог корака врши се одређивање хемијског једињења (будућег лека) које поседује одговарајућу биолошку активност, а „*чија структура представља полазну тачку за даљу модификацију*“ [1]. Најпре се дизајнирају и синтетишу библиотеке једињења, а потом се врши њихов „*скрининг (прегледање и сортирање) кроз експерименталне тестове, коришћењем идентификоване мете. Она једињења која покажу активност у тестовима се означавају као „погоци“ (енгл. hits), и наставља се њихова провера. Лек мора бити сигуран, ефикасан и мора да поседује: могућност добре апсорпције у организму, метаболизам који му омогућава довољно дуг полуживот, селективан ефекат на мету, малу токсичност и минималне нежељене ефекте.*“ [1].

На дугом путу стварања нових лекова су *in vitro* испитивања цитотоксичног (антипролиферативног) деловања према различитим туморским ћелијским линијама [31]. Ћелијске линије се добијају од примарних ћелијских култура успостављених директно из узорака ткива пацијената [32]. Селекција ћелијске линије за било који експеримент захтева успостављање баланса између избора одговарајућег модела и одабира ћелијске линије са којом се може радити. Ћелијски тестови (есеји) често се користе за скрининг колекције једињења да би се утврдило да ли тест молекули имају ефекте на пролиферацију (умножавање) ћелија или показују директне цитотоксичне ефекте који на крају доводе до ћелијске смрти [33]. Овим тестовима се на различитим типовима ћелија могу мерити: вијабилност, степен пролиферације, интегритет ћелијске мембране, синтеза ДНК и други параметри ћелијског метаболизма [32]. За процену биолошких ефеката се користе различите методе, које могу бити квалитативне и квантитативне. У најширој употреби су тестови са тетразолијумовим солима (MTT, XTT, MTS, WST и др.) за евалуацију ћелијског метаболизма [32]. Одабир есеја може бити изазован и мора одговарати одређеним потребама. Пре свега мора се знати какав тип мерења се може извести на крају експеримента - IC_{50}, EC_{50} итд. Есеји омогућавају мерења различитих маркера који указују на број мртвих ћелија (тест цитотоксичности), број живих ћелија (тест способности преживљавања), укупан број ћелија или механизам смрти ћелије (нпр. апоптоза) [34]. Такође, приликом одабира есеја у обзир се мора узети и време које је неопходно за његову примену. „*Једињења која при ниским концентрацијама значајно смањују пролиферисање туморских ћелија представљају неопходан предуслов за успешан развој нових, ефикаснијих терапеутика за лечење оболелих од малигних болести. Антипролиферативна активност, тј. способност неког једињења да смањи број ћелија у култури у односу на контролу квантитативно се изражава у процентима за одређену концентрацију и време излагања. На основу временске и дозне зависности цитотоксичног ефекта могу се извести закључци о антипролиферативној активности испитиваног једињења. Ефикасност инхибиције пролиферације туморских ћелија испитиваним једињењем током одређеног периода се изражава* IC_{50} *вредношћу, која представља концентрацију испитиване супстанце при којој се број третираних ћелија у датом времену смањује за 50% у односу на нетретирани контролни узорак. На основу ове вредности могуће је упоређивати ефикасност различитих једињења на истој ћелијској линији. Мада не постоји строга подела једињења на основу цитотоксичног ефекта, међу научницима који се баве оваквим истраживањима прихваћене су неке смернице у вези квантификације јачине антипролиферативног (цитотоксичног) ефекта: Уколико је* IC_{50} *вредност нижа од 10 μM, сматра се да то једињење показује добру цитотоксичност. Ако је* IC_{50} *у*

интервалу 10-100 μM, супстанца показује умерен до слаб цитотоксични ефекат, док се једињења са IC_{50} *вредношћу преко 100 μM сматрају практично нетоксичним*“ [32].

Преклиничка истраживања лекова представљају слојевит и дуготрајан процес, који у просеку траје око 10 година. Пре него што доспе на тржиште, односно прође одређене контроле од стране надлежних агенција као што су EMA[8] (*European Medicines Agency*) или FDA[9] (*Food and Drug Administration*), он мора проћи кроз адекватно клиничко истраживање [1]. Овај корак подразумева проверу безбедности лека над узорцима (добровољцима) са циљем утврђивања ефикасности и споредних ефеката лека током дужег периода коришћења [1]. Реакције људског организма на лек су од суштинског значаја и неретко могу бити негативне (са смртним исходом) и зато је неопходно извршити одговарајуће и прецизно преклиничко истраживање. Поред лидера у овој области, као што је NAMSA[10], многе мање институције и организације се, како у развијеним земљама, тако и у земљама у развоју, баве преклиничким тестирањем лекова. У наставку је представљена једна од лабораторија, која се бави овим важним кораком у процесу рационалног дизајна лекова.

2.3 Центар за преклиничка тестирања активних супстанци - CPCTAS

Лабораторија за ћелијску и молекуларну биологију у оквиру CPCTAS [35], главни је носилац националног пројекта „Преклиничка испитивања биоактивних суспатнци“ (PIBAS 14010). Она „*представља централни и координирајући сегмент активности Центра за преклиничка тестирања активних супстанци, па су, сходно томе њена организација и активности предмет акредитације у складу са стандардом SRPS ISO/IEC 17025:2006, за период 04/2012-04/2016. Основне активности Лабораторије имају за циљ испитивање значаја физиолошких, генетичких, молекуларно-биолошких и туморских маркера у процени ефеката активних супстанци и предвиђању патолошких стања код људи. Дефинисање механизма деловања активних супстанци у биолошким системима са јасним повратним информацијама за хемијску синтезу нових варијанти активних супстанци побољшаних особина, као и дефинисаним излазом за дистрибуцију, клиничко тестирање и примену испитиваних активних супстанци, представља циљеве реализације активности Лабораторије. Мисија Лабораторије је да кроз научно-истраживачки и едукативни процес, уз максимално и стално иновирање истраживачког рада, примену савремених метода и техника истраживања, допринесе унапређењу постојећих и развоју нових сазнања о начинима деловања различитих активних супстанци и могућностима њихове примене на живе системе (ћелије, ткива, органи, организми).*“ [35].

Биоинформатичка област рада Лабораторије подразумевала је реализацију различитих комплексних тестова и експеримената. Експерименти који се изводе укључују праћење *in-vitro* ефеката активних супстанци у ћелијским линијама различитог порекла (нарочито ћелијских линија канцера) и примарних ћелија изолованих из различитих ткива [36], што је један од важних корака у рационалном дизајну лекова. Извођење експеримената је резултовало јединственом и сложеном структуром (Слика 2.1), која је синтетизована од стручних, углавном биолошких термина, односно концепата, који су приказани на различитим нивоима, тј. хијерархијски. Структура је представљена у раду [36], који представља један од научних доприноса ове дисертације.

[8] *http://www.ema.europa.eu/ema/*

[9] *https://www.fda.gov/*

[10] *https://www.namsa.com/services/medical-device-testing/*

EXPERIMENTS

A) TYPE OF EXPERIMENTS
- A. MODEL SYSTEMS
  - I. ANIMALS
    1. RATS
    2. FISH
  - II. CELL LINES
    1. PRIMARY
    2. CANCER
  - III. PATIENTS
- B. METABOLIC SYSTEMS
  - I. ENERGY METABOLISM
  - II. OXIDATIVE STRESS
    1. REACTIVE SPECIES
    2. OXIDATIVE DAMAGE
    3. ANTIOXIDATIVE SYSTEM
  - III. HEMATOLOGY
  - IV. ENDOCRINE SYSTEM
  - V. IMMUNE SYSTEM
- C. TREATMENT
  - *I. IN VIVO*
  - *II. IN VITRO*

B) THE AIM

C) TYPE OF TREATMENTS
- A. ACTIVE SUBSTANCES
- B. DRUGS
- C. TREATMENT
  - *I. IN VIVO*
  - *II. IN VITRO*
- D. DOSES
  - I. ACUTE
  - II. CHRONIC
- E. PATHOLOGY

D) METHODS
- A. MATERIAL
- B. MODEL SYSTEMS
  - I. ANIMALS
    1. RATS
    2. FISH
  - II. CELL LINES
    1. PRIMARY
    2. CANCER
  - III. PATIENTS
- C. ANALYTICAL METHODS
  - I. HEMATOLOGICAL PARAMETERS
  - II. OXIDATIVE/ANTIOXIDATIVE METABOLISM
    1. REACTIVE OXYGEN SPECIES
       - A. SUPEROXIDE ANION RADICAL
       - B. HYDROGEN PEROXIDE
    2. REACTIVE NITROGEN SPECIES
       - A. NITRIC OXIDE
       - B. PEROXYNITRITE
       - C. NITROXYL ANION
    3. OXIDATIVE DAMAGE PARAMETERS
       - A. METHEMOGLOBIN
       - B. HEINZ BODIES
       - C. LIPID PEROXIDES
    4. NON-ENZYMATIC ANTIOXIDATIVE COMPONENTS
       - A. REDUCED GLUTATHIONE
       - B. OXIDIZED GLUTATHIONE
       - C. VITAMIN C
       - D. VITAMIN E
    5. ANTIOXIDATIVE ENZYMES
       - A. SUPEROXIDE DISMUTASE
       - B. CATALASE
       - C. GLUTATHIONE PEROXIDASE
       - D. GLUTATHIONE REDUCTASE
       - E. GLUTATHIONE S TRANSFERASE
  - III.ENERGY METABOLISM
    1. OXIDATIVE PHOSPHORYLATION PARAMETERS
    2. GLYCOLYSIS PARAMETERS
- D. CELL CULTURE ASSAYS
  - I. CELL ASSAYS
  - II. MTT ASSAY
  - III. IMMUNOFLUERESCENCE
- E. PROTEIN ANALYSIS ASSAYS
  - I. PROTEIN EXTRACTION FROM TISSUES
  - II. PROTEIN EXTRACTION FROM RBCs
  - III. PROTEIN EXTRACTION FROM MITOCHONDRIA
  - IV. SDS PAGE
- F. DNA AND RNA ANALYSIS ASSAYS
  - I. RNA EXTRACTION FROM TISSUES
  - II. RT-PCR

E) RESULTS

F) RESULTS AND COMMENTS

Слика 2.1 Хијерархијска структура експеримената који се изводе у оквиру Лабораторије

Табела 2.1 представља оригинална тумачења неких концепата хијерархијске структуре.

Табела 2.1 Значења неких концепата хијерархијске структуре експеримената (извор *[36]*)

Коцепт	**Значење**
Experiments	An experiment is a methodical trial and error procedure carried out with the goal of verifying, falsifying, or establishing the validity of a hypothesis. Experiments vary greatly in their goal and scale, but always rely on repeatable procedure and logical analysis of the results. An experiment is a method of testing - with the goal of explaining - the nature of reality.
Type of Experiments	Assigning subjects to conditions by the experimenter.
Model Systems	Biological systems (animals, cell lines, patients) that are extensively studied with the expectation that discoveries will provide insight into specific biological phenomena.
Metabolic systems	The set of chemical reactions that happen in the cells of living organisms. These processes allow organisms to grow and reproduce, maintain their structures, and respond to their environments.
Treatments	Treatment is in vivo or in vitro application of defined doses of active substances/drugs on experimental model systems.
The aims	Generate measurable data that can be tested, and contribute to gradual accumulation of human knowledge.
Type of treatment	Specifies in vivo or in vitro application of active substances /drugs defined doses on experimental model systems.
Active substance	Chemical or plant substance that affects the physiology, the function of the body of a human or animal.
Drugs	A drug, broadly speaking, is any substance that, when absorbed into the body of a living organism, alters normal bodily function.

Doses	Administration of tested active substances/drugs in experimentally defined amounts.
Pathology	Pathology is the precise study and diagnosis of disease. Pathology addresses four components of disease: cause/etiology, mechanisms of development (pathogenesis), structural alterations of cells (morphologic changes), and the consequences of changes (clinical manifestations).
Methods	Techniques for phenomena investigation, new knowledge acquiring, or correcting and integrating previous knowledge. It is based on gathering empirical and measurable evidence subject to specific principles of reasoning and consisting in systematic observation, measurement, experiment, formulation, testing and modification of hypotheses.
Material	Material Tools or apparatus for the performance of a given task. Also, chemicals used for experiment.
Analytical Method	Techniques used to draw statistical inferences including multiple regression, path analysis, discriminate analysis and logistic analysis.
Cell Culture Assay	Assay employed in cell culturing.
Protein Analysis Assay	Assay employed in protein analysis.
DNA and RNA Analysis Assay	Protocol employed in molecular biology methods.
Result and comments	The final consequence of a sequence of actions or events expressed qualitatively or quantitatively.

Односи између концепата прилично су комплексни и утврђују се на основу извршених експеримената. Сваки концепт представља класу, која истовремено може бити и потклаcа, својство или инстанца неке друге класе. На пример, концепти *Methods* и *Type of Experiments* представљају класе, док концепт *Model Systems* означава истовремено и (пот)класу и својство класа *Methods* и *Type of Experiments*. Структура експеримената је требало да буде представљена на разумљив начин, тако да је корисници Лабораторије могу једноставно користити [36]. Додатно, било је потребно обезбедити могућност једноставне модификације структуре, као и претрагу експерименталних података, који су се обављали у складу са том структуром. Првобитна идеја репрезентације структуре експеримената подразумевала је одговарајући графички приказ[11], а затим и примену релационих база података [36]. Међутим, ови приступи нису пружили очекиване резултате. Релационе базе података нису биле погодне [37] за моделовање хијерархијске структуре концепата и комплексних односа између њих на различитим нивоима. Такође, поменути приступи су одбачени и због отежаног додавања нових концепата услед потенцијалне проширивости структуре, као и због потребе за евентуалним извођењем закључака. Да би се ови проблеми на неки начин премостили било је неопходно истражити проблематику представљања биоинформатичких података.

2.4 Проблеми у домену биоинформатике

Као почетак развоја биоинформатике може се сматрати период настанка првих база података које су представљене у одељку 2.2.1. Примарни задатак биоинформатике био је управо креирање и одржавање тако великих база. За то је било потребно изградити комплексан механизам који омогућава коришћење већ постојећих информација (веб сервиси), као и проширење база новим или ревидираним подацима [38]. Међутим, многе базе које су се развијале у домену биоинформатике подржавале су податке различитих формата као што су текстуалне датотеке, чланци научних часописа, XML (*eXtensible Markup Language*) фајлови, релационе базе итд., а притом су и сами подаци често бити структурирани или неструктуирани [39]. Пресудни проблем свакако је био одсуство стандарда, односно модела за адекватно представљене хетерогених података.

Уз огромне количине података у базама нуклеотидних и протеинских секвенци, макро-молекуларних и библиографских база, јавила се и потреба за развојем нових метода за анализу тих података [1]. Циљ је био користити принципе метода из области анализе података. Међутим, овај процес се могао постићи

[11] *http://cpctas-lcmb.pmf.kg.ac.rs/lcmb/expGraphView.php*;
http://cpctas-lcmb.pmf.kg.ac.rs/lcmb/expGraph/cpctasTypeOfExperiments.htm;
http://cpctas-lcmb.pmf.kg.ac.rs/lcmb/expGraph/experimentalMethods.htm;
http://cpctas-lcmb.pmf.kg.ac.rs/lcmb/expGraph/treatmentTypes.htm

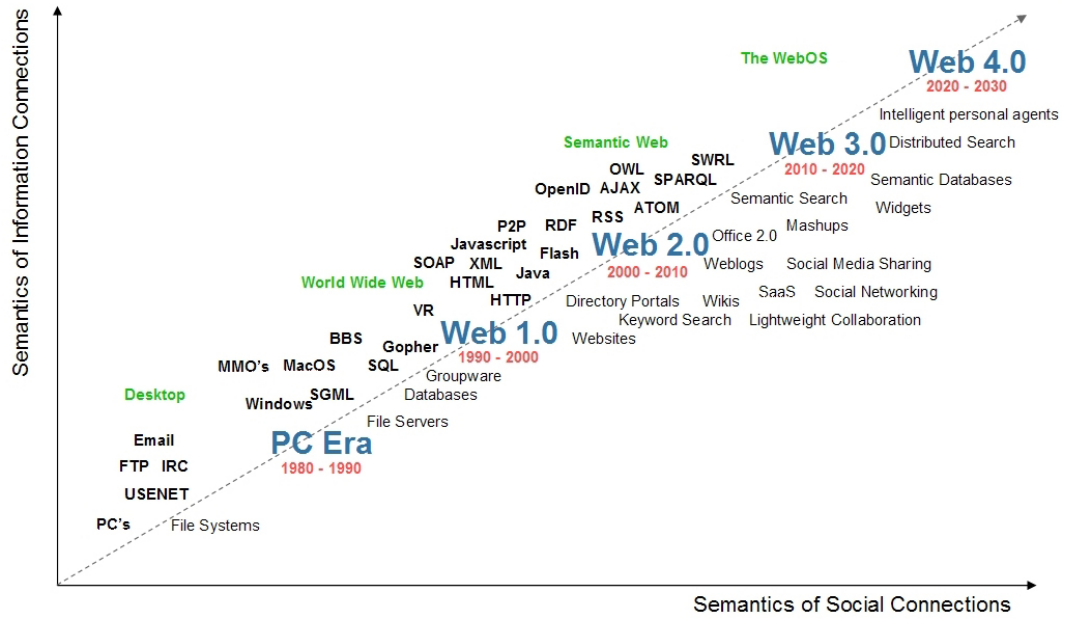
само ако су подаци семантички представљени и интегрисани, тј. ако се вишеструки извори података повежу у одређене системе [40,41]. Имплементација таквих интегрисаних система подразумева стварање великих мрежа повезаних јединења, биолошких мета, гена, ћелијских путева, лекова, болести и нежељених ефеката из вишеструких хетерогених извора [39]. Један од главних проблема интеграције јесте и тај што се многи извори података преклапају (енгл. *cross-reference data*) и покривају (прекривају) идентичне и сличне податке (хомогени или полу-хомогени извори података) о чему је детаљније размотрено у одељку 7.1.1. У таквим случајевима семантички однос оваквих скупова података често је нејасан [39].

Значајан задатак биоинформатике јесте и претрага информација над вишеструким изворима података. За извођење овог процеса није довољно бити само врсни познавалац одговарајућих програмерских техника и приступа, јер њихово познавање често није довољан предуслов да би се преузели адекватни подаци. Преузимање података о одговарајућим ентитетима (биолошким метама, ћелијским линијама, лековима итд.) условљено је и утврђивањем односа између самих ентитета који припадају хетерогеним изворима података. Такође, неопходно је поседовати и одређено доменско знање како би се разумела семантика ентитета.

Биоинформатичка заједница је зато била принуђена да уведе одговарајуће стандарде и технологије, који омогућавају адекватно представљање, интеграцију и претрагу података. Како је развој биоинформатике текао паралелно са развојем и усавршавањем веба, тако је биоинформатичка заједница покушавала да искористи његове квалитете у циљу решавања актуелних проблема. У наставку рада је описан хронолошки развој веба и његов утицај на домен биоинформатике.

2.4.1 Веб и биоинформатика

Тим Бернерс-Ли (*Tim Berners-Lee*) је 1989. године формулисао термин веб (*енгл. World Wide Web*) као иновативни концепт, чија је основна идеја стварање заједничког информационог простора у коме људи могу да комуницирају разменом информација [42]. Генерално, егзистирају четири генерације веб технологија: веб 1.0 (веб информација), веб 2.0 (комуникациона мрежа), веб 3.0 (семантички веб) и веб 4.0 (симбиотски веб) што приказује Слика 2.2. Веб 4.0 је у зачетку и није од важности за ово истраживање.



Слика 2.2 Хронолошки развој веба (извор: веб[12])

[12] *http://www.sidar.org/ponencias/2008/egyrs/rioja/graf/RadarNetworksTowardsAWebOS.jpg*

Веб 1.0 је махом био намењен читању докумената. Ова генерација веба је омогућавала корисницима да прикажу своје каталоге или да представе продукте и сервисе на интернету [43]. Веб сајтови нису били интерактивни, већ су више личили на неку врсту интернет брошура, које су углавном укључивале статичке HTML странице које су ретко када биле ажуриране [43]. Главни циљ сајтова је био саопштити информације и успостави присуство на интернету, због чега је веб имао атрибуте статичности и једносмерности. За разлику од прве генерације веба, друга генерација третира веб као платформу која корисницима омогућује интеракцију, једноставно праћење и сарадњу у креирању садржаја, а корисници веба више нису само пасивни посматрачи и примаоци информација [43]. Они сада могу учествовати у њиховом стварању, допуњавању, модификацијама и преношењу [43]. Веб 2.0 је заправо мрежа докумената, при чему је главни проблем што су веб документа дизајнирана искључиво за употребу од стране људи - семантика садржаја и линкови су имплицитни и повезаност између објеката је прилично ниска [43].

У почетној фази развоја биоинформатика је била заснована на технологијама прве и друге генерације веба, тако да је, као и веб, имала ограничене могућности за отворену и непосредну комуникацију између истраживача. Квантитативно повећање података доступних путем веба донело је низ предности за биоинформатичку заједницу. Међутим, такви подаци су најчешће били разумљиви истраживачима, али не и самим рачунарима. Ово је посебно било изражено са другом генерацијом веба. Тада се проблеми у домену биоинформатике продубљују, јер је било све теже одржавати, чувати и претраживати биоинформатичке податке који су се рапидно нагомилавали. Притом, многе базе података су се развијале у релативној изолацији, не поштујући биоинформатичке стандарде. Истраживачи су били принуђени да прелазе са једне веб локације на другу и да прате путање међусобно повезаних података, како би открили релевантне информације.

Предлог Тим Бернерс-Лија, да створи глобалну мрежу података, која би се темељила на основној идеји веба 2.0 али уз додатак семантике, био је велики плус за биоинформатичку заједницу. Развој нове генерације веба - семантичког веба, решио је многе биоинформатичке проблеме укључујући проблеме представљања, интеграције и претраге података [44]. Кључне компоненте семантичког веба, RDF [11] и онтологије [12], имају најважнију улогу јер омогућавају ефикасно представљање и интеграцију података. Упитни језик SPARQL [13], као стандардни упитни језик за RDF, омогућава претрагу семантички представљених података. Семантички веб је омогућио структуру садржаја веб страница, стварајући на тај начин средину у којој софтверски агенти могу да прелазе са једне веб странице на другу и да обављају софистициране задатке за истраживаче. Ово је свакако био велики допринос за биоинформатичку заједницу, јер је омогућена имплементација различитих софтверских алата за претрагу података. Неки од најзначајнијих су анализирани у одељку 6.4. У следећем поглављу детаљније су представљене технологије семантичког веба, како би се разумео начин њихове примене и деловање на домен биоинформатике.

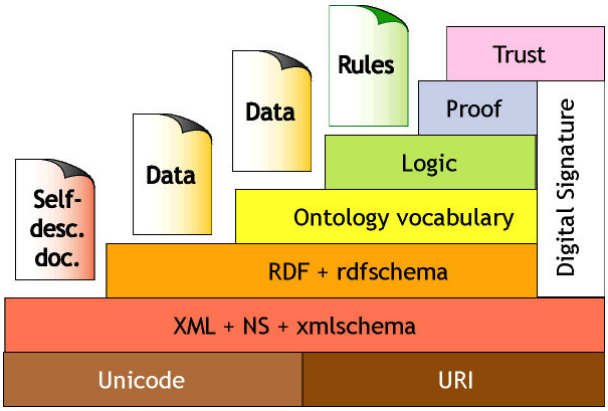
3 Технологије семантичког веба

У овом поглављу је представљена архитектура семантичког веба као и његове основне технологије. Примери који се користе у оквиру описа технологија базирани су на истраживачком раду у циљу примене семантичких технологија за потребе Лабораторије. Ово поглавље може послужити као својеврстан водич за употребу датих технологија у било ком домену. Циљ је представити слојеве архитектуре семантичког веба и показати на које све начине се они могу искористити за процесе представљања и претраге података.

3.1 Архитектура семантичког веба

Семантички веб или веб 3.0 је развијен како би се превазишли проблеми веба 2.0. и сматра се његовим проширењем [43]. За разлику од друге генерације веба, информације представљене технологијама семантичког веба имају јасно дефинисано значење, што поједностављује њихову употребу. Семантички веб се дефинише као мрежа података, при чему је дизајн података прилагођен захтеву да су подаци читљиви од стране људи и машина [43]. Основна концепција је уклонити јаз између људи и машина путем прогресивних технологија које представљају знање. Зато је битно представити знање у формату који је разумљив машинама, а које се може лако процесирати и користити за резоновање. Коначни циљ је обезбедити да све те задатке спроводе софтверски агенти (апликације) и на тај начин учинити веб разумљивим за машину [43]. Семантички веб се као термин често користи и за идентификацију скупа технологија, алата и стандарда који чине основне елементе система који би могао да оствари визију веба прожетог семантиком [6]. Технологије семантичког веба омогућавају да се подаци анотирају (означе) метаподацима, који описују значење тих података [43]. Метаподаци су од изузетне релевантности за софтверске агенте у смислу интерпретације, односно разумевања тих података.

Као и већина информационих технологија и семантички веб има слојевиту архитектуру. Слика 3.1 представља дијаграм са главним елементима ове архитектуре, који је предложио Тим Бернерс-Ли [6], а који је касније доживео бројне варијације.



Слика 3.1 Архитектура семантичког веба (извор: веб[13])

Сваки слој, односно ниво архитектуре, има јасну и стриктну дефиницију, која се користи за „изградњу“ слоја изнад. Виши слојеви користе функционалности нижих слојева. Аутори истраживања [45] дефинишу специфичности појединачних слојева:

- ***Unicode и URI*** - има задатак да јединствено дефинише појмове;
- ***XML NS и XML Schema*** - представља темељну структуру записа података;
- ***RDF*** - модел који повезује податке; ***RDF Schema*** - део слоја који даје базични смисао и значење везама и самим тим омогућава хијерархију;

[13] *https://www.w3.org/RDF/Metalog/images/sw-tower.png*

- ***Ontology vocabulary*** - надограђује претходни слој, али са могућношћу већег броја семантичких веза између података;
- ***Logic*** - задужен је за разумљиву екстракцију информација и доношење одлука од стране апликација;
- ***Proof*** - пружа механизме за проверу тачности пронађених података;
- ***Trust*** - задужен је за проверу у смислу да ли је могуће веровати извору да су подаци тачни.

Иначе, слојеви технологија се не развијају истом брзином: последња три слоја се активно промовишу и унапређују, док су нижи слојеви у великој мери стандардизовани.

Тим Бернерс-Ли [6] је навео да су основне технологије семантичког веба XML [46] и RDF [11]: XML је усвојен као *de facto* стандард за пренос и размену података на вебу, док је RDF усвојен као стандард за презентацију информација о ресурсима на вебу, односно за опис значења података. Такође, Тим Бернерс-Ли [6] сматра да су онтологије [12] кључни механизам за представљање и поновну употребу знања, као и за генерисање нових чињеница кроз механизме закључивања на основу експлицитно датих чињеница. У наставку рада су представљене стандардизоване технологије.

3.2 Uniform Resource Indentifier - URI

Основни слој архитектуре семантичког веба чине Unicode [47] и URI [9]. Unicode је стандард који омогућава да се различити језици на вебу могу користити примењујући стандардизоване формате. Један такав формат је URI. Он представља низ знакова који једнозначно идентификују неки ресурс на вебу, при чему се под ресурсом подразумева било који објекат који може (али и не мора) бити повезан са нечим што постоји на интернету [47]. Постоји неколико подскупова URI идентификатора, а два основна су URL (*Uniform Resource Locator*) и URN (*Uniform Resource Notation*) [48]. Адресе докумената на вебу, односно URL примери, заправо су примери URI идентификатора са којима се најфреквентније сусрећемо. Основна разлика између URI и URL идентификатора је у тзв. идентификатору фрагмента. Идентификатор фрагмента је низ карактера смештених након домена (ознаке #), који означава да је тај фрагмент, односно секундарни ресурс, подређен домену, тј. примарном ресурсу. Идентификатор фрагмента може представљати и неки део документа. На пример, URI *http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#modelSystem* се користи за идентификацију класе *Model systems* из структуре експеримената (Слика 2.1). Код записивања URI идентификатора често се због прегледности избегава писање домена и користе се именски простори (енгл. *namespace*) [49], који могу бити стандардизовани (подразумевани) или кориснички дефинисани. Тако је претходни пример записа еквивалентан са *pibas:modelSystem*, где *pibas* означава кориснички дефинисан именски простор, а *modelSystem* идентификатор фрагмента. У наставку се примењује скраћени вид записа URI идентификатора, а Додатак садржи примере свих коришћених именских простора у дисертацији. За разлику од URL идентификатора, URN омогућава идентификацију докумената без навођења локације документа. Идентификатор URN користи тзв. *urn scheme*. Пример URN идентификатора је *urn:isbn:0-262-01242-1*, која представља одговарајућу књигу, где је *isbn* ознака за *International Standard Book Number*[14].

Дакле, Unicode се користи за представљање било ког симбола на јединствен начин без обзира на то који се језик користи, док је URI јединствени идентификатор ресурса свих врста [43]. Функционалност базичног слоја семантичког веба може се окарактерисати као пружање јединственог механизма за идентификацију ресурса, који се користи у даљем развоју технологија.

3.3 eXtensible Markup Language - XML

Према речима Тим Бернерс-Лија [6] XML је једна од основих технологија семантичког веба. Идеја везана за XML је да он дефинише стандард којим се размењују подаци на вебу, али тако да се сва пажња усмери на садржај, а не на приказ података [46]. Најјасније речено, XML је скуп података из неког домена

[14] *https://www.isbn.org/about_isbn_standard*

записаних на структуриран и стандардан начин. XML је језик који не поседује предефинисан скуп кључних речи (елемената и атрибута), већ је то језик за дефинисање других језика (мета језик).

XML документ се састоји од низа угњеждених елемената - ознака (енгл. *tag*), унутар једне изворне ознаке (енгл. *root element*). Свака ознака може имати произвољан број атрибута или својстава (енгл. *property*). XML документ фактички представља означено (лабелирано) стабло (енгл. *tree*), при чему је свака ознака у релацији са означеним чвором у моделу података, а свака угњеждена ознака представља чвор „дете“ у стаблу. Слика 3.2 представља пример XML документа који садржи податке подсегмента структуре експеримената. Овај XML документ има елементе *modelSystem*, *name* и *hasInstance*.

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
 <modelSystem>
  <name>Animals</name>
   <hasInstance>Rats</hasInstance>
   <hasInstance>Fish</hasInstance>
 </modelSystem>
 <modelSystem>
  <name>Cell lines</name>
   <hasInstance>Primary</hasInstance>
   <hasInstance>Cancer</hasInstance>
 </modelSystem>
 <modelSystem>
  <name>Patients</name>
 </modelSystem>
</root>
```

Слика 3.2 Пример XML документа са елементима који представљају подсегмент структуре експеримената (Слика 2.1)

Основни проблем XML-а је немогућност јединственог именовања елемената и атрибута [50]. Механизам који омогућава решење овог проблема назива се XML именски простор, а који се дефинише кључном речи *xmlns* као атрибут неког елемента.

```
<?xml version="1.0"?>
<modelSystems xmlns:PIBAS="http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#">
 <PIBAS:modelSystem>
  <PIBAS:name>Animals</PIBAS:name>
  <PIBAS:hasInstance>Rat</PIBAS:hasInstance>
  <PIBAS:hasInstance>Fish</PIBAS:hasInstance>
 </PIBAS:modelSystem>
 <PIBAS:modelSystem>
  <PIBAS:name>Cell lines</PIBAS:name>
  <PIBAS:hasInstance>Primary</PIBAS:hasInstance>
  <PIBAS:hasInstance>Cancer</PIBAS:hasInstance>
 </PIBAS:modelSystem>
 <PIBAS:modelSystem>
  <PIBAS:name>Patients</PIBAS:name>
 </PIBAS:modelSystem>
</modelSystems>
```

Слика 3.3 Пример коришћења именских простора у XML документу који представља подсегмент структуре експеримената (Слика 2.1)

У примеру (Слика 3.3) се употребљава кориснички дефинисан именски простор *PIBAS*. За сваки елемент који припада овом именском простору, а налази се унутар елемента *PIBAS:modelSystem*, мора се експлицитно означити припадање том именском простору. XML именски простори омогућавају јединствено именовање извесних елемената и атрибута, али помоћу њих није могуће дефинисати да су ти елементи (нпр. *PIBAS:name*) типа стринг, нити да тај стринг не сме бити дужи од 15 карактера. Тај недостатак решава XML Schema (*eXtensible Markup Language Schema*) [51].

3.3.1 eXtensible Markup Language Schema - XML Schema

Језик за дефинисање структуре XML документа јесте XML Schema (XMLS). Она има два важне карактеристике. Прва карактеристика се огледа у дефинисању структуре XML документа, што обухвата: дефинисање елемената и атрибута који се појављују у документу, дефинисање редоследа елемената,

дефинисање хијерархије елемената и дефинисање броја појављивања једног елемента [50]. Друга важна карактеристика је могућност дефинисања типова података које атрибут или елемент садржи [50]. Ово подразумева да је могуће дефинисати да ли је елемент или атрибут празан или садржи одређени текст, одредити тип података који тај текст садржи (цео број, децимални број, датум итд.), дефинисати подразумевану вредност или константу вредност елемената и атрибута и одредити скуп вредности које неки елемент или атрибут може имати [50]. Слика 3.4 представља пример XMLS документа (Слика 3.3 садржи податке дефинисане у складу са том шемом). Конкретно, овом шемом је дефинисано да елемент *PIBAS:ModelSystem* има елементе *PIBAS:name* и *PIBAS:hasInstance* који су типа стринг (*xs:string*) и који се могу користити неограничен број пута (*xs:string*).

```
<?xml version="1.0" encoding="UTF-8"?>
 <xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified" attributeFormDefault="unqualified">
  <!-- XML Schema Generated from XML Document on Sat Mar 23 2019 13:17:22 2MT+0100 (Central Europe Standard Time) -->
  <!-- with XmlGrid.net Free Online Service http://xmlgrid.net -->
  <xs:element name="modelSystems">
   <xs:complexType>
    <xs:sequence>
     <xs:element name="PIBAS:modelSystem" maxOccurs="unbounded">
      <xs:complexType>
       <xs:sequence>
        <xs:element name="PIBAS:name" type="xs:string"></xs:element>
        <xs:element name="PIBAS:hasInstance" maxOccurs="unbounded" type="xs:string"></xs:element>
       </xs:sequence>
      </xs:complexType>
     </xs:element>
    </xs:sequence>
    <xs:attribute name="xmlns:PIBAS" type="xs:string"></xs:attribute>
   </xs:complexType>
  </xs:element>
 </xs:schema>
```

Слика 3.4 Пример XMLS документа са елементима који представљају подсегмент структуре експеримената (Слика 2.1)

Основа карактеристика XMLS-а јесте да је она утемељена на XML-у па самим тим има могућност поновног коришћења постојећих шема. Она има могућност наслеђивања типова података, што омогућује креирање нових типова проширивањем или надоградњом већ постојећих [50]. Ипак, иако XML и XMLS омогућавају општу, добро дефинисану синтаксу лаку за процесирање, они се не баве семантиком података које описују. То подразумева да је изнад XML-а морао бити креиран адекватни стандард који се бави овом тематиком. Први корак у том смеру јесте RDF [11], општи модел на слоју метаподатака и RDF Schema (*Resource Description Framework Schema*) [52], језик на нивоу модела.

3.4 Resource Description Framework - RDF

Модел података који користи URI спецификацију да идентификује ресурсе на вебу и опише релације између ресурса према дефинисаним својствима и вредностима назива се RDF. У својим почецима RDF се користио за представљање мета података о HTML страницама, као што су информације о аутору или ауторским правима (енгл. *copyright*) [53]. Оваква примена RDF-а није допринела жељеној дефиницији семантичког веба. Зато се приступило генерализацији термина „ресурс“. Данас се ресурс изражава помоћу URI спецификације, која пружа довољну проширивост, па се може рећи како све што може поседовати јединствени идентификатор представља ресурс [53]. На пример, ресурс може бити HTML документ, XML документ, колекција докумената, део документа, веб сајт итд. Поред ресурса основни градивни блок RDF-а су својства (енгл. *properties*) и изјаве (енгл. *statement*) [11]. Својства представљају специфичан аспект неког ресурса. Они описују релације између ресурса и идентификују се помоћу URI спецификације. Ресурси и својства се комбинују како би се креирале једноставне изјаве. RDF модел користи три начина записивања изјава: триплетима (енгл. *triples*), графовима и XML-ом [49]. Посматрајмо чињеницу *Animals is a type of ModelSystem* из структуре експеримената (Слика 2.1). RDF изјава ове чињенице у форми триплета је облика (*pibas:Animals, pibas:isTypeOf, pibas:ModelSystem*). Триплети су облика (*објекат - атрибут - вредност*, односно (*субјекат-предикат-објекат*)[15]. Субјекат је

[15] У већини литертурних извора се не прави разлика између термина *предикат*, *атрибут* и *својство*. Исти принцип важи и у дисертацији.

увек ресурс, док објекат може бити ресурс или литерал (константа). У датом примеру објекат је представљен као ресурс. Слика 3.5 представља графичку репрезентацију изјаве. RDF графови су директни графови са означеним чворовима и усмереним гранама, при чему се гране користе за повезивање субјекта изјаве са објектом изјаве. Представљање објекта као ресурса омогућава креирање комплексних графова, јер објекат једне RDF изјаве може бити субјекат друге изјаве.

Слика 3.5 Графичка репрезентација RDF изјаве (*pibas:Animals*, *pibas:isTypeOf*, *pibas:ModelSystem*)

Графови јесу моћан механизам, посебно за људско разумевање, али визија семантичког веба захтева репрезентацију информација у облику који је разумљив рачунарима. У складу са тим постоји и трећа могућност интерпретације изјаве, у форми XML-а. Слика 3.6 представља RDF/XML документ који одговара примеру изјаве. Ови типови докумената користе *rdf:RDF* ознаке, који садрже *rdf:Description* ознаке, а који представљају изјаве. У *rdf:RDF* ознакама дефинишу се и атрибути који представљају именске просторе. У нашем случају дефинисан је кориснички именски простор *pibas*.

```
<?xml version="1.0"?>
<rdf:RDF pibas:http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
 <rdf:Description about="pibas:Animals">
  <pibas:isTypeOf>pibas:ModelSystem</pibas:isTypeOf>
 </rdf:Description>
</rdf:RDF>
```

Слика 3.6 Репрезентација RDF изјаве (*pibas:Animals*, *pibas:isTypeOf*, *pibas:ModelSystem*) у форми XML-а

Модел података RDF је синтаксно-независан, апстрактни модел који одређује стандард метаподатака и који служи за опис ресурса на вебу [53]. Иако је овим моделом остварен несумњив напредак у смислу семантике, он поседује одговарајуће недостатке. Конкретно, са RDF моделом није могуће дефинисати типове података и њихове узајамне односе, тј. није могуће дефинисати класе и установити њихову хијерархију. Тај проблем надомешћује RDF Schema [52].

3.4.1 Resource Description Framework Schema - RDF Schema

Модел података RDF може се формулисати кроз више машински читљивих језика, а најчешћи је језик шеме - RDF Schema (RDFS). RDFS се посматра као проширење RDF-а са речником за дефинисање класа, хијерархију класа, својстава (бинарних релација), хијерархије и рестрикције својстава [52]. Изворне класе у RDFS-у су: *rdfs:Resource* (класа свих ресурса), *rdfs:Class* (класа свих класа), *rdfs:Literal* (класа свих литерала - константи), *rdf:Property* (класа свих својстава) и *rdf:Statement* (класа свих изјава). Појединачни објекти који припадају класи представљају инстанце те класе. Однос између инстанци и класе, као и односи између самих класа дефинисани су својствима [52]:

- *rdf:type* - повезује ресурс са припадајућом класом (ресурс је дефинисан као инстанца те класе);
- *rdfs:subClassOf* - повезује класу са неком од њених надкласа, све инстанце класе су инстанце својих надкласа;
- *rdfs:subPropertyOf* - повезује својство са неким од његових надсвојстава.

Својства која се користе за дефинисање ограничења су [52]:

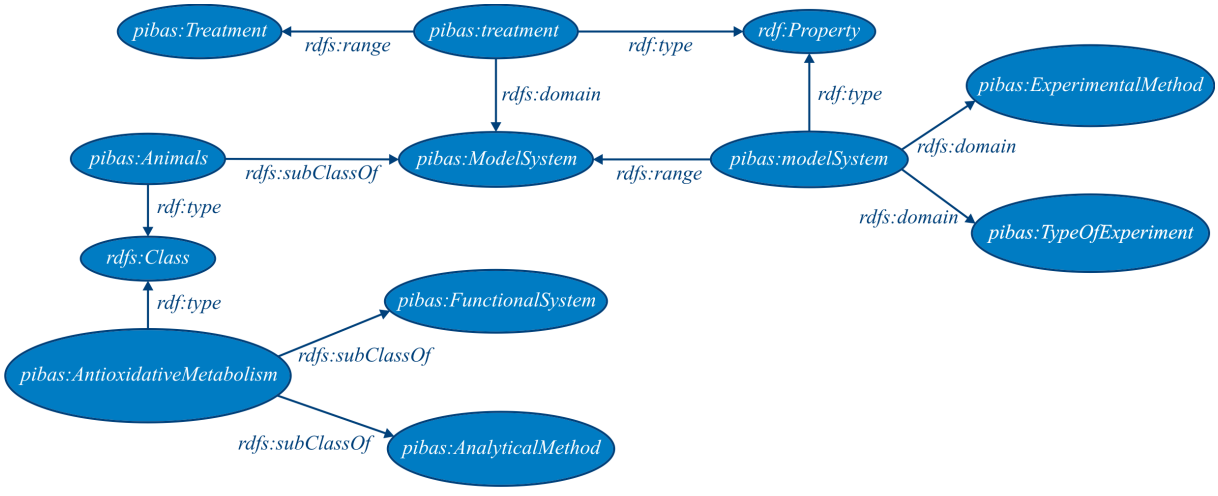
- *rdf:domain* - прецизира домен својства и дефинише да је сваки ресурс који представља вредност својства инстанца класе домена;
- *rdfs:range* - прецизира опсег (кодомен) својства и наводи да су вредности својства инстанце класе опсега (кодомена).

Слика 3.7 а) представља пример RDFS документа, који се односи на подсегмент структуре експеримената (Слика 2.1). У примеру су дефинисане класе *pibas:Animals* и

pibas:AntioxidativeMetabolism. Класа *pibas:Animals* је потклаcа класе *pibas:ModelSystem,* а класа *pibas:AntioxidativeMetabolism* је потклаcа класа *pibas:AnalitycalMethod* и *pibas:FunctionalSystem*. Можемо уочити и својства *pibas:treatment* и *pibas:modelSystem*. Својство *pibas:treatment* има за домен класу *pibas:ModelSystem*, а за опсег класу *pibas:Treatment*. Домен дефинисаног својства *pibas:modelSystem* су класе *pibas:ExperimentalMethod* и *pibas:TypeOfExperiment*, а опсег је класа *pibas:ModelSystem*. Одговарајући графички приказ датог RDFS документа представља Слика 3.7 б).

```
<rdf:RDF xml:lang="en"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#”
   xmlns:base="http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#">

<rdf:Description rdf:ID="Animals">
   <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
   <rdfs:subClassOf rdf:resource="#ModelSystem"/>
</rdf: Description>

<rdf: Description rdf:ID="AntioxidativeMetabolism">
   <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
   <rdfs:subClassOf rdf:resource="#AnalitycalMethod"/>
   <rdfs:subClassOf rdf:resource="#FunctionalSystem"/>
</rdf: Description>

<rdf:Description rdf:ID="treatment">
 <rdf:type resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
    <rdfs:domain rdf:resource="#ModelSystem"/>
    <rdfs:range rdf:resource="#Treatment"/>
</rdf: Description>

<rdf:Description rdf:ID"modelSystem"›
 <rdf:type resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
 <rdfs:domain rdf:resource="#ExperimentalMethod"/>
 <rdfs:domain rdf:resource="#TypeOfExperiment"/>
 <rdfs:range rdf:resource="#ModelSystem"/>
</rdf:Description>
</rdf:RDF>
```

а)



б)

Слика 3.7 RDFS документ (а) и одговарајући графички приказ (б) са елементима који представљају подсегмент структуре експеримената (Слика 2.1)

Модел података RDF, заједно са RDFS-ом, представља темељни слој семантичког веба са гледишта размене метаподатака на нивоу података [43]. Ипак, RDF и RDFS су релативно ограничени у својим могућностима. У RDFS-у, рецимо, није могуће дефинисати дисјунктне класе (класе без заједничких елемената) или ограничити колико различитих вредности својство може имати. Зато је било неопходно

направити нови искорак у домену семантичког веба и обезбедити још „богатију“ семантику. Тако су настале онтологије [12] које се сматрају кулминацијом стандардизованих технологија семантичког веба.

3.5 Онтологије

Термин „онтологија“ потиче из V. века п. н. е. и тада је означавао науку о бићу, која системски објашњава биће, проучава природу и организацију стварности [53]. Управо је таква дефиниција утицала на то да се термин онтологија „позајмљује“ и примењује у различитим доменима. Деведесетих година XX века онтологије су постале и предмет информатичких и рачунарских истраживања. У овом пољу онтологија се формално дефинише као систем појмова (концепата) и релација између њих [53]. Тачније, онтологија је образац података који представља концепте унутар неког домена и односе између тих концепата [54]. Такође, онтологије су усвојене и као примарно средство за моделовање и репрезентацију знања. Аутори у [55] наглашавају да онтологије рефлектују фундаментални, монолитни поглед на знање, при чему се знање посматра као целина на коју се нови делови могу континуирано надовезивати. Једна од кључних одлика онтологија, као механизама за моделовање знања, јесте потенцијал поновне употребе и дељења знања (енгл. *shared knowledge*) за нова истраживања [55].

3.5.1 Типови онтологија

У сфери репрезентације знања маневрише се онтологијама различитог нивоа сложености. Stevens и др. [55] истичу да се онтологије генерално класификују у генеричке, доменске и онтологије задатка. Генеричке онтологије су заправо детерминисане као опште онтологије и аутори истраживања [55] наводе да се оне конструишу са циљем апсорбовања знања из више домена. Концепти које ове онтологије моделују подложни су адаптацијама у онтолошком моделу, па су ове онтологије посебно адекватне у ситуацијама када се захтева поновна употреба знања [56]. Доменске онтологије садрже речнике одређеног домена и одређене релације међу њима [56]. Дистинкција између доменских и општих онтологија огледа се у томе што концепти у доменским онтологијама презентују спецификацију концепата дефинисаних у општим онтологијама, и што се доменске онтологије могу изнова употребити само у домену у коме су дефинисане [56]. Онтологије задатка описују речнике који су везани за неки специфичан задатак или активност, а концепти ових онтологија такође настају спецификацијом концепата презентованих у општим онтологијама [56]. У пракси је понекад тешко одредити тип онтологије, јер је класификација само смерница која помаже финијем разумевању и прихватању онтологије.

3.5.2 Компоненте онтолошких модела

Елементарне компоненте онтолошког модела су концепти, релације, инстанце и аксиоми (постулати) [56,57]. Концепти су базне компоненте онтолошких модела, које представљају класе ентитета који се налазе у оквиру разматраног домена [55]. Они могу бити једноставни или комплексни, конкретни или мисаони, стварни или фиктивни [57]. На пример, у структури експеримената (Слика 2.1) као примери концепта могу се издвојити термини *Type of Experiments*, *Model Systems*, *Active Substances*, *Cell lines* итд. Релације су врсте ентитета које илуструју спону између концепата. Разликују се две врсте релација: таксономије и асоцијативне релације [55]. Таксономијом се утврђују релације које наглашавају врсту концепата, а концепти се даље организују у (концептуална) стабла, која су хијерархијски профилисана. На пример, канцер (*Cancer*) је представник ћелијских линија (*Cell lines*), а ћелијске линије су модел системи (*Model Systems*). Асоцијативне релације имају задатак да повежу концепте унутар самог стабала и да представе међусобни положај концепата (нпр. активна супстанца (*ActiveSubstance*) има одређену улогу у експерименту (*Experiment*)) као и функцију концепата. Инстанце су фактички ентитети најнижег реда, који су су потчињени концептима вишег реда [56]. На пример, *IN VIVO* може бити инстанца концепта *Treatment*. Аксиоми имају облик сажетих реченица које су увек веродостојне и примењују се у онтологијама да би се успоставили одговарајући лимити над класама и инстанцама [57]. Пример аксиома је да се сваки експеримент (*Experiment*) изводи над одговарајућом активном супстанцом (*Experiment*). Аксиоми се најчешће дефинишу коришћењем неког математичког механизма као што је логика првог реда. Иначе, компоненте онтологије често нису статичке. Оне подлежу променама током развоја саме

онтологије. Промене на доменима и прилагођавање различитим условима захтевају модификацију онтологија.

3.5.3 Конструкција онтологија

Конструкција онтологија је комплексан и итеративан процес и постоје различити истраживачки приступи који описују ову методологију [58,59]. Конструкција онтологија, у зависности од домена онтологије и типова података који се представљају, најчешће обухвата следеће кораке [60]:

- **Дизајнирање** - у овом кораку потребно је одредити обим, сврху онтологије и везе између концепата неопходне за изградњу таксономије;
- **Имплементацију** - овај корак подразумева креирање онтологије при чему се евентуално могу користити друге актуелне онтологије;
- **Интеграцију** - овај корак подразумева комбиновање имплементиране онтологије са већ актуелним онтологијама (уколико се оне могу адаптирати виталном концепту изворне онтологије);
- **Валидацију** - верификација онтологије подршком експерта или уграђених аутоматских алата;
- **Итерацију** - итеративни приступ целог поступка до коначног резултата.

За конструкцију онтологија често је неопходно коришћење одређених софтверских алата, тзв. онтолошких едитора. Они се могу применити на разне фазе у животном циклусу онтологије, укључујући развој, интеграцију и валидацију [43]. Најпопуларнији онтолошки едитори су Protégé [61], OntoStudio [62], TopBraid Composer[16] итд. За потребе дисертације коришћен је Protégé едитор.

3.5.4 Ontology Web Language - OWL

Појава XML-а као стандарда за размену података на вебу, утицала је на то да већина онтолошких језика има синтаксу базирану на XML-у. Број онтолошких језика је временом растао и често се вршило њихово комбиновање да би се проблем неког домена могао записати. Као последица, настао је OWL (*Ontology Web Language*) [63], који се сматра основним онтолошким језиком. Овај семантички језик задужен је за публиковање и размену онтологија на вебу. Његова основна улога је приказивање значења термина у речницима и односа међу тим терминима, а то значи да су информације које се налазе у документима обрадиве од стране апликација и да је њихов садржај разумљив човеку [64].

3.5.4.1 OWL синтакса

OWL је језик великих могућности и у овој подсекцији представљене су само његове основе. Примери који илуструју примену OWL-а су представљени у наредном поглављу. Синтакса OWL-а је у великој мери заснована на RDFS-у, али OWL има више могућности за исказивање значења и семантике [53]. У OWL-у елементом *owl:Class* дефинише се класа, а постоје и унапред дефинисане класе: *owl:Thing* и *owl:Nothing*. Класа *owl:Thing* је најопштија класа (класа свих ресурса), а *owl:Nothing* је празна класа. Свака класа је потклаcа класе *owl:Thing* и надкласа класе *owl:Nothing*. OWL синтакса дефинише две врсте својстава: објектна својства (енгл. *object property*), која се односе на објекте других објеката и својства која се односе на тип податка вредности (енгл. *datatype property*). Синтакса OWL-а не садржи предефинисане типове података, нити обезбеђује дефиницију посебних објеката. Уместо тога, она омогућава коришћење RDFS типова података. Особине својстава се могу дефинисати директно, коришћењем синтаксних елемената:

- *owl:TransitiveProperty* - дефинише транзитивна својства;
- *owl:SymmetricProperty* - дефинише симетрична својства;
- *owl:FunctionalProperty* - дефинише функционална својства, а која имају најмање једну вредност за сваки објекат;
- *owl:InverseFunctionalProperty* - дефинише својства која за два различита објекта не могу имати

[16] *https://www.topquadrant.com/docs/tbc/AppDevQuickstartGuide.pdf*

исту вредност.

Синтакса OWL-а подржава и дефинисање нових класа применом скуповних операција (*owl:union*, *owl:complement* или *owl:intersection*) над постојећим класама. Коришћењем *owl:disjointWith* елемента могу се дефинисати дисјунктне класе. У OWL-у је могуће дефинисати синониме за инстанце, класе и својства, као и минималну и максималну кардиналност неког својства у односу на класу, итд [65]. Инстанце класа се дефинишу као у RDFS-у.

3.5.4.2 Закључивање

Закључивање (енгл. *reasoning*) подразумева извођење чињеница које нису експлицитно наведене. Са онтолошког становишта могуће је извести следеће чињенице [64]:

- **Припадност класи** - Ако је *x* инстанца класе *C*, а *C* је потклаcа класе *D*, онда се може закључити да је *x* инстанца класе *D*;
- **Еквиваленција класа** - ако је класа *A* еквивалент класе *B*, а класа *B* је еквивалент класе *C*, онда је класа *A* еквивалент класе *C*;
- **Доследност** - ако је *x* инстанца класе *A* и класа *A* је потклаcа класе *B*, а притом је *A* потклаcа класе *C*, а *B* и *C* су дисјунктне класе, онда постоји недоследност, јер би класа *A* требало да буде празна, али она има инстанцу *x*. Ово је тзв. индикација грешке у онтологији;
- **Класификација** - ако инстанца *x* испуњава услове и ако су одређени парови својство-вредност довољан услов за чланство у класи *A*, може се закључити да је *x* инстанца класе *A*.

Претходна извођења се углавном спроводе применом уграђених семантичких закључивача (енгл. *semantic reasoner*s). Они су често алати онтолошких едитора. Најпопуларнији су FACT[17], Pellet[18] и Jena[19]. Подршка закључивању је јако важна јер омогућава проверу доследности онтологије и знања, проверу односа између класа и аутоматску класификацију инстанци класа [64].

Дакле, OWL описује информације неког домена, али када му се придода и могућност извођења чињеница које нису експлицитно наведене, он постаје и језик за представљање знања [64]. OWL се између осталог, користи као декларативни језик у представљању знања разних домена као што су медицина, молекуларна биологија, софтверско инжењерство, конфигурација сложених система итд. Посебно место OWL заузима у домену биоинформатике и у каснијем прегледу литературе (одељак 4.5.1 и одељак 6.4) су представљена решења која користе предности овог језика. У наредном поглављу је представљен процес развоја PIBAS онтологије, која користи OWL синтаксу, а која је конструисана са циљем моделовања структуре експеримената (Слика 2.1).

3.6 SPARQL Protocol and RDF Query Language – SPARQL

Визија семантичког веба је да третира веб као једну велику базу отворених повезаних података (енгл. *Linked Open Data*) [64]. Претходно представљене технологије омогућавају презентацију и чување података, али је циљ обезбедити и њихову претрагу. За ту намену развијен је SPARQL [13] - стандардни упитни језик за претрагу података у RDF моделу. Он је сличан SQL-у што може бити изненађујуће с обзиром на чињеницу да су модели релационих база података и RDF-а потпуно другачији. Како је RDF заснован на триплету, тако је и SPARQL заснован на подударању графовских образаца. Најједноставнији графовски образац је троструки образац - патерн (енгл. *pattern*), који је сличан RDF изјави, али са могућношћу променљиве уместо RDF субјекта, предиката или објеката [64].

3.6.1 SPARQL синтакса

У SPARQL синтакси се користи резервисана реч PREFIX [13], која представља еквивалент именском

[17] *http://www.cs.man.ac.uk/~horrocks/FaCT/*

[18] *http://semanticweb.org/wiki/Pellet*

[19] *http://jena.apache.org/*

простору. Могу се користити стандардизовани или кориснички дефинисани именски простори. Додатак садржи примере неких префикса (именских простора). SPARQL упити се генерално могу поделити у четири групе:

- SELECT упити - имају синтаксу сличну SQL-у; ту спадају и упити типа UPDATE и DELETE;
- ASK упити - намењени су провери да ли неки упит има резултат, повратна вредност овог упита је логичка вредност (*True/False*);
- DECSRIBE упити - користе се за преузимање свих триплета (у форми фајла) о ресурсу који се наводи у WHERE клаузули;
- CONSTRUCT упити - користе се за конструисање RDF графова.

SPARQL синтакса омогућава коришћење резервисаних речи, које побољшавају резултате упита на следећи начин:

- DISTINCT - уклања дупликате из скупа решења (ову резервисану реч користе само SELECT упити);
- FILTER - филтрира резултат на основу задатих услова;
- LIMIT - ограничава број враћених резултата из упита;
- OPTIONAL - омогућава приказ променљивих и у случају да за њих нема резултата;
- OFFSET - изоставља првих *n* резултата;
- ORDER BY - сортира резултат.

```
PREFIX pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?ModelSystem
WHERE
  {
    ?ModelSystem rdf:type pibas:ModelSystem.
  }
LIMIT 2
```

Слика 3.8 Пример SELECT SPARQL упита чији резултат извршавања чине инстанце класе *pibas:modelSystem* дефинисане у PIBAS онтологији

Слика 3.8 представља пример једноставног SELECT SPARQL упита који преузима инстанце класе *pibas:modelSystem* дефинисане у PIBAS онтологији. Број повратних вредности у овом примеру је ограничен уз помоћ резервисане речи LIMIT.

3.6.2 SPARQL endpoint

SPARQL упити се изводе над тзв. SPARQL крајњим тачкама (енгл. *endpoint*), који представљају веб сервисе, који прихватају и обрађују SPARQL упите. Они су одређени наменским URL адресама и често нуде HTML форму, која корисницима омогућава директно креирање упита. Након извршавања упита, веб сервис прослеђује одговор кориснику у једном од неколико формата (HTML, RDF/XML, JSON итд.). Међу најпознатијим SPARQL серверима су Virtuoso[20], Joseki[21], Fuseki[22] итд. Један од најчешћих проблема у раду са SPARQL *endpoint*-има јесте честа блокираност проузрокована фреквентним коришћењем. Један од начина да се ови проблеми избегну јесте складиштење база података на локалном SPARQL серверу. У том случају би се подацима приступало коришћењем FROM клаузуле у оквиру упита, чиме би се повећала брзина извођења самих упита. Међутим, како ове базе података неретко поседују огроман број триплета, њихово локално складиштење захтева значајне рачунарске ресурсе.

[20] *http://vos.openlinksw.com/owiki/wiki/VOS/VOSSparqlProtocol*

[21] *http://www.joseki.org/*

[22] *https://jena.apache.org/documentation/serving_data/*

3.6.3 Federated SPARQL упити

Специјална врста SPARQL упита су Federated SPARQL упити [66], који се користе за претрагу како *remote endpoint*-а, тако и за комбиновање података са више извора. Ови упити су од изузетно велике важности за претрагу дистрибуираних база података, какве су и биоинформатичке базе података. За креирање Federated SPARQL упита користи се резервисана реч SERVICE (или чешће SERVICE SILENT како би се избегле грешке у раду са *endpoint*-има) која омогућава усмеравање одређеног дела упита на SPARQL *endpoint*. Резултат сваког подупита се затим спаја са остатком упита и резултат се приказује кориснику. Federated SPARQL упити омогућавају приступ неограниченом броју извора података, односно нуде коришћење неограниченог броја SERVICE клаузула. Међутим, један од предуслова за успешан Federated SPARQL упит јесте управо редослед SERVICE клаузула које су њему користе, јер овакви типови упита често користе променљиву из једне SERVICE клаузуле у другој, како би на тај начин остварили комбинацију SPARQL *endpoint*-а, односно извора података. Дакле, променљиве морају бити дефинисане у одговарајућем редоследу и SERVICE клаузуле морају бити позиване у коректном редоследу. Слика 3.9 представља пример Federated SPARQL упита који преузима лекове (из одговарајућих биоинформатичких извора података DrugBank/Bio2RDF и Chembl/EMBL-EBI) који делују на биолошку мету издвојену као резултат тестирања активне супстанце у бази података развијеној за потребе Лабораторије.

```
PREFIX pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX drugbank:<http://bio2rdf.org/drugbank/vocabulary:>
PREFIX cco:<http://rdf.ebi.ac.uk/terms/chembl#>

SELECT DISTINCT ?drug
WHERE
  {   { SERVICE SILENT <http://cpctas-lcthb.pmf.kg.ac.rs:3030/PIBAS/query>
        { ?drug pibas:hasInChiKey "DQLATOHUWYHOKH-UHFFFAOYSA-L".
          ?experiment pibas:hasActiveSubstance ?substance;
                      pibas:hasTarget ?target.
        }
      }
    UNION
      { SERVICE SILENT <http://drugbank.bio2rdf.org/sparql>
        { ?drug drugbank:target ?target. }
      }
    UNION
      { SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/chembl/sparq1/>
        { ?drug cco:hasTarget ?target. }
      }
  }
```

Слика 3.9 Пример Federated SPARQL упита помоћу кога се врши претрага лекова (у DrugBank/Bio2RDF и Chembl/EMBL-EBI базама података) који делују на биолошку мету издвојену као резултат тестирања активне супстанце у CPCTAS бази података

У наредном поглављу је представљен значај примене семантичких технологија кроз домен биоинформатике. Извршен је преглед актуелних биоинформатичких репозиторијума, као и начин на који они користе погодности семантичког веба. Централни део поглавља је резервисан за анализу имплементираних решења за потребе Лабораторије, који такође користе предности семантичких технологија.

4 Развој PIBAS онтологије, CPCTAS онтолошке базе података и прототипа софтвера за претрагу базе

У овом поглављу је представљена PIBAS онтологија чија је намена моделовање структуре варијетета експеримената у Лабораторији. Моделовање је постигнуто дефинисањем веза и релација између концепата који представљају разноврсне биолошке термине, а који се примењују у експерименталним процедурама. У складу са тим је извршен преглед постојећих онтологија које се баве сличном тематиком, односно моделовањем структуре експеримената. Доменска PIBAS онтологија детаљно је представљена својим компонентама и структуром. Она уважава концепте проширивости и интеграције, а има за циљ да послужи као средство за прикупљање знања. Такође, представљена је и онтолошка база података CPCTAS, која садржи податке појединачних експеримената који се изводе у Лабораторији, као и прототип софтвера за претрагу дате базе, који је претходио развоју Платформе. У последњем делу поглавља су разматрани и неки од актуелних биоинформатичких репозиторијума (онтолошких база података), њихова улога у домену савремених истраживања, као и однос са CPCTAS базом података у процесу интеграције. У овом поглављу су наведени и истраживачки захтеви које би биоинформатички семантички репозиторијуми требало да задовоље у процесу рационалног дизајна лекова.

4.1 Онтологије експеримената

Један од важних задатака природних наука јесте прогресивно повећање знања извођењем експеримената [67]. Могућ начин формализације знања јесте дефинисањем онтологија, које задовољавају мерила општих онтологија и прилагођавају се специфичним захтевима. У наставку су представљена нека решења.

Аутори рада [67] развили су EXPO онтологију за научне експерименте. Како се у свим наукама користе слични експериментални принципи, сродни инструменти и материјали, а експерименти се притом извршавају и анализирају на сличан начин, било је изводљиво и пожељно развити једну такву онтологију. EXPO онтологија се заснива на класичној теорији дизајна експеримента, статистике, теорији грешака, анализи постојећих доменских онтологија и анализи метаподатака експеримената. Ова онтологија предлаже стандардизовани речник за објашњење научних експеримената и притом омогућава дељење и поновну употребу заједничког знања.

Удружење MGED[23] (*Microarray Gene Expression Data*) покушало је да уз помоћ своје MO онтологије формализује описе експеримената [68]. Њихова онтологија је креирана са циљем да обезбеди стандардну терминологију за *microarray* експерименте[24]. Радне групе HUPO PSI General Proteomics Standards[25] и Mass Spectrometry[26] изградили су онтологију која подржава протеомске експерименте [69]. Радна група MSI[27] (*Metabolomics Standards Initiative*) развила је WG онтологију како би омогућила научној заједници да схвати, интерпретира и интегрише метаболичке експерименте [70]. Онтологија CHMO (*Chemical Methods Ontology*) описује методе који се користе за колекцију података у хемијским експериментима, припрему и одвајање материјала за даље анализе и синтезу материјала [71]. Она такође описује и инструменте који се користе у експериментима. Дата онтологија је комплементарна са OBI (*Ontology for Biomedical Investigations*) онтологијом [72], која пружа егзактно дефинисана значења свих аспеката

[23] *mged.sourceforge.net/*

[24] *https://discover.nci.nih.gov/microarrayAnalysis/Experimental.Design.jsp*

[25] *http://www.psidev.info/*

[26] *http://www.psidev.info/groups/mass-spectrometry*

[27] *http://www.metabolomics-msi.org/*

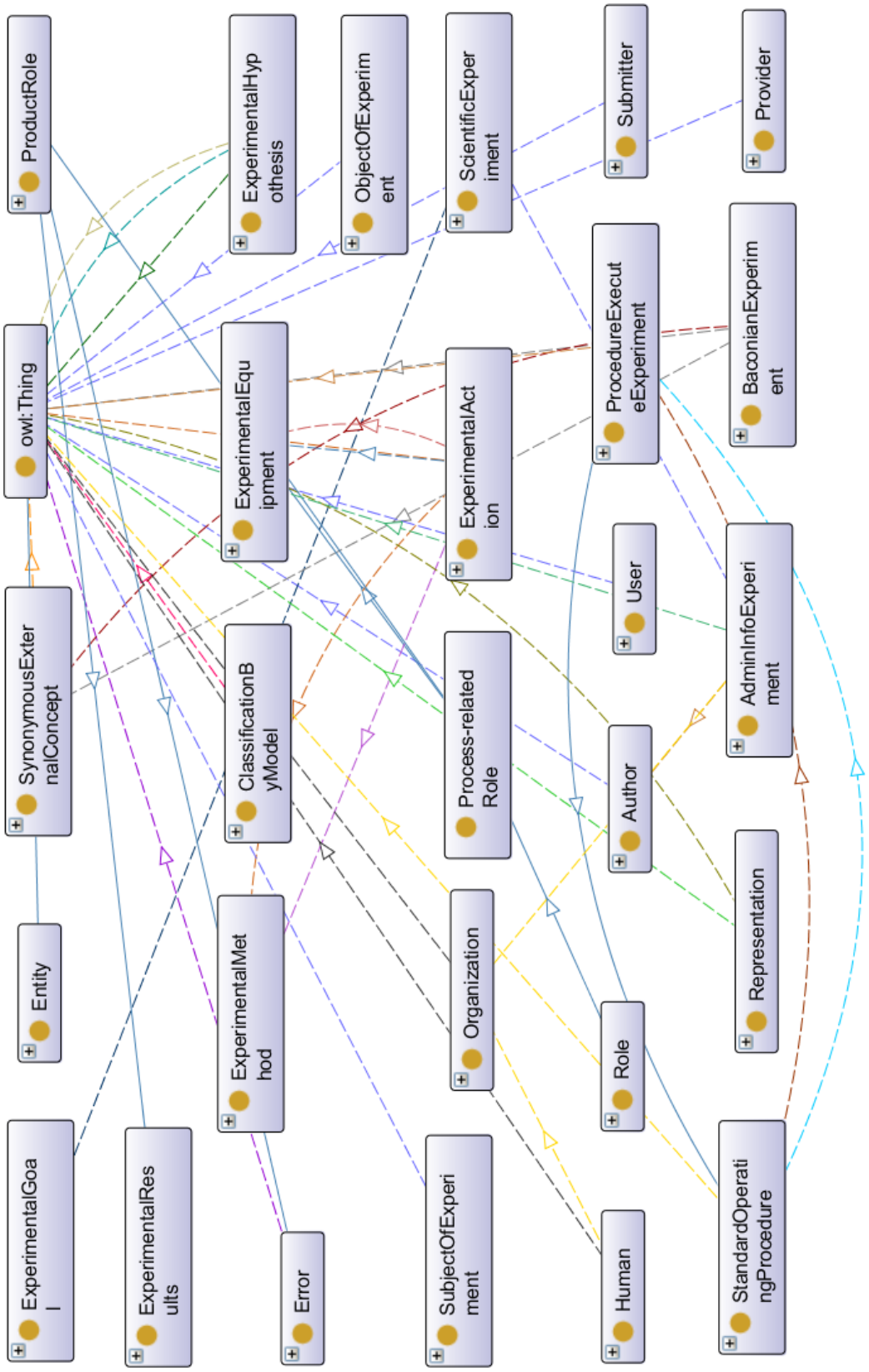
истраживања у медицинским и биолошким доменима. Она обухвата све етапе процеса истраживања: планирање, спровођење експеримента и креирање извештаја.

Иако наведене онтологије значајно доприносе формализацији експеримената у научним областима, оне су често неадекватне и не могу се усвојити као стандард, јер су првенствено оријентисане ка специјализованим доменима. Додатно, експериментални рад неретко зависи од одређених околности, тако да је тешко креирати онтологију за научне експерименте која је усвојена као стандард [36]. Како ниједна од поменутих онтологија не нуди онтолошки оквир за специфичну структуру експеримената који се обављају у Лабораторији, постојала је потреба за развојем посебне онтологије, која уважава концепте проширивости и интеграције и служи као средство за прикупљање знања, које је значајно у планирању будућих експеримената у Лабораторији.

4.2 Конструкција PIBAS онтологије

Анализирајући технологије семантичког веба и истраживања у одељку 4.1, дошло се до закључка да онтологије имају највећи потенцијал за моделовање структуре експеримената (Слика 2.1). Конструкција PIBAS (*Preclinical Investigation of BioActive Substances*) онтологије била је комплексан и итеративан процес имајући у виду сложеност саме структуре. Као што је већ представљено у одељку 2.3 структура експеримената се састоји од стручних (већином биолошких) термина, који су хијерархијски повезани. Тачна репрезентација те структуре није нимало лака јер се она састоји од више нивоа термина, који су прилично неправилно дистрибуирани [36]. Однос термина у структури може бити класа-потпкласа, класа-својство или класа-инстанца. У процесу дизајнирања било је пре свега неопходно разумети термине и дефинисати њихове међусобне односе неопходне за изградњу таксономије. Након тога је уследио процес имплементације. Због своје доступности (енгл. *open source*), поузданости и популарности, као алат за имплементацију, изабран је онтолошки едитор Protégé [61]. У овом кораку вршен је и процес верификације онтологија представљених у одељку 4.1 у циљу усвајања постојећих концепата унутар доменске PIBAS онтологије. У тренутку конструкције PIBAS онтологије није постојала идеја о могућој интеграцији са неком другом сличном онтологијом, већ само могућност усвајања и прилагођавања постојећих концепата. Након исцрпне анализе утврђено је да EXPO[28] онтологија [62] (Слика 4.1) има потенцијал за усвајање концепата због своје уопштености.

[28] *http://expo.sourceforge.net/*



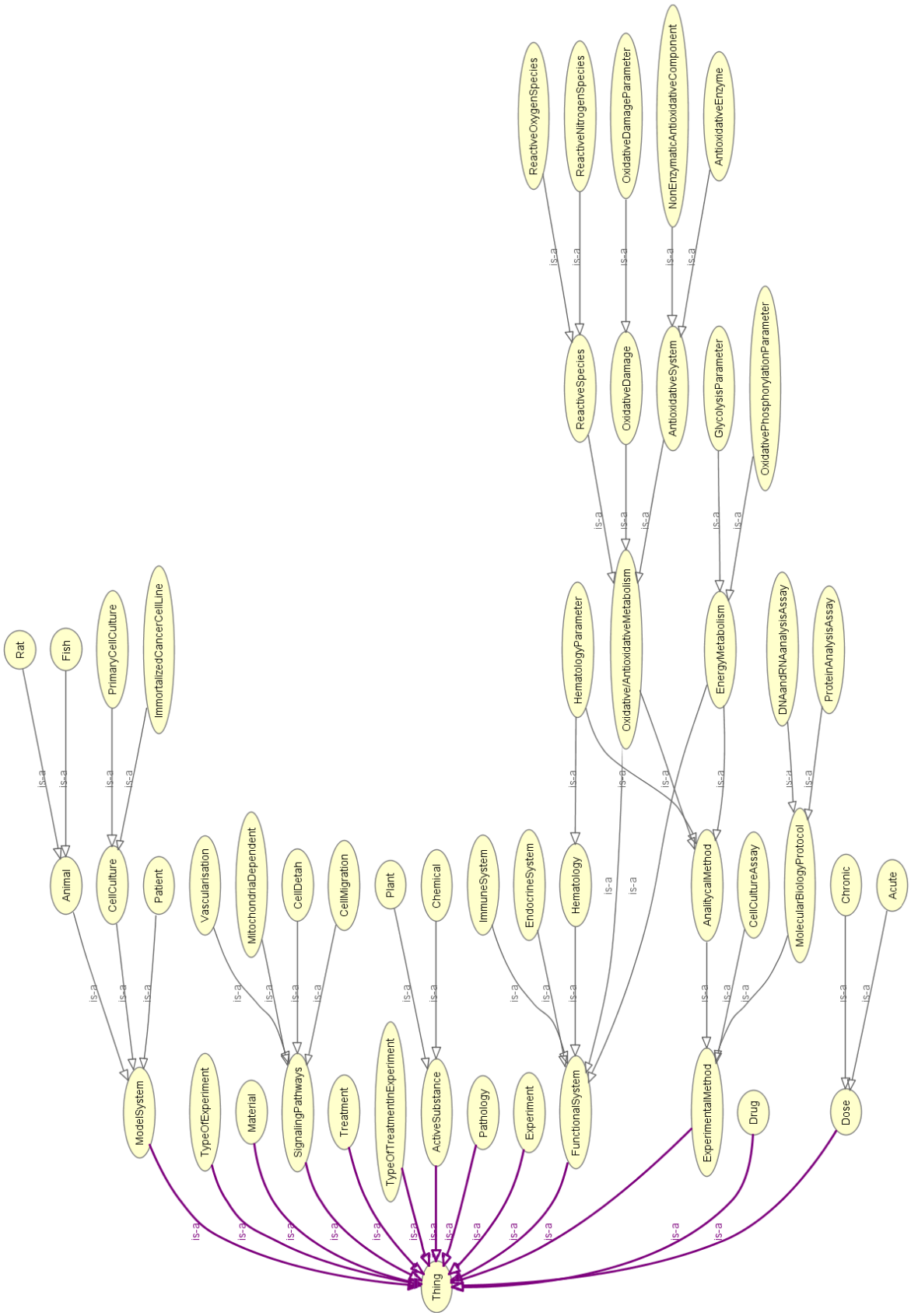
Слика 4.1 Део таксономије концепата EXPO онтологије[29]

[29] За визуелизацију је коришћен Protégé додатак OntoGraph.

Међутим, с обзиром на специфичну структуру експеримената Лабораторије, усвојени су и модификовани само одређени концепти EXPO онтологије. Конкретно, на основу класа *expo:ExperimentalResult* и *expo:ExperimentalGoal* креирани су одговарајући пандани, односно својства типа података: *pibas:result* и *pibas:theAimOfExperiment*. У EXPO онтологији су уочене и класе *expo:ScientificExperiment* и *expo:ExperimentalMethod*, које имају исту намену као и концепти *Experiment* и *ExperimentalMethod* дефинисани у структури експеримената. Остали концепти PIBAS онтологије имају епитет аутохтоних концепата. Процес валидације је обављен од стране експерата (особља Лабораторије), док је процес закључивања, у циљу провере конзистентности онтологије, обављен применом FACT алата Protégé едитора. Читав процес конструкције је прошао кроз неколико итерација док нису били задовољени сви критеријуми. У наставку је детаљно представљена таксономија PIBAS онтологије.

4.2.1 Таксономија концепата PIBAS онтологије

Слика 4.2 представља таксономију концепата PIBAS онтологије. Све класе у PIBAS онтологији су потклассе од *owl:Thing* и припадају одговарајућем хијерархијском нивоу. Међу класама првог хијерархијског нивоа издвојићемо класу *pibas:Experiment,* која је намењена за генерално представљање „Експеримент“ концепта. Ова класа објашњава експеримент као методичко испитивање која се спроводи с циљем провере, фалсификовања или утврђивања ваљаности хипотезе експеримента (Табела 2.1). Описи свих класа у онтологији се презентују помоћу *rdfs:comment* предиката и одговарају значењима које представља Табела 2.1. Класа *pibas:ExperimenatalMethod* описује експерименталну методу као технику за истраживање феномена, стицање нових знања или исправљање и интеграцију претходног знања. Класа *pibas:TypeOfExperiment* дефинише тип експеримента (True, Quasi- и Single-subject експерименте). Класа *pibas:ModelSystem* означава концепт који представља биолошке системе. Класа *pibas:ActiveSubstance* је концепт који описује хемикалију или супстанцу (потенцијални лек) која утиче на физиологију, функцију тела човека или животиње. Ове супстанце могу бити вештачке (*pibas:Chemical*) или природне (*pibas:Plant*). Класа *pibas:Treatment* дефинише тип третмана (*in vivo* или *in vitro*). Класа *pibas:Material* дефинише алате, уређаје и хемикалије који се користе за спровођење експеримента. Класа *pibas:Drug* се користи за дефиницију лека, али је заправо еквивалентна класи *pibas:ActiveSubstance*, иако то није експлицитно наглашено у онтологији. Класа *pibas:SignalingPathways* дефинише трансдукцију сигнала која се јавља када екстраћелијски сигнални молекул активира рецептор на површини ћелије. Класа *pibas:Pathology* дефинише прецизну студију дијагнозе болести. Класа *pibas:Dose* бележи тестиране активне супстанце (лекове) у експериментално дефинисаним количинама.



Слика 4.2 Основна таксономија концепата PIBAS онтологије[30]

[30] За визуелизацију је коришћен Protégé додатак OWLViz.

На другом хијерархијском нивоу су дефинисане поткласе (*owl:subClassOf*) одговарајућих класа из првог хијерархијског нивоа. Издвојићемо поткласе класа *pibas:ModelSystem* и *pibas:ExperimenatalMethod*. Поткласе, класе *pibas:ModelSystem* су *pibas:Animal*, *pibas:CellCulture* и *pibas:Patient*. Прва поткласа се бави концептом разноликости која се налази код животиња, а који се користи за експерименталне подухвате (*pibas:Rat* и *pibas:Fish*). Друга поткласа се користи за представљање ћелијских линија (*pibas:ImmortalizedCancerCellLine* и *pibas:PrimaryCellCulture*). Трећа поткласа представља пацијенте који учествују у експериментима. Међу потклasама класе *pibas:ExperimenatalMethod* издвајамо потклasу *pibas:MolecularBiologyProtocol* која представља протоколе (анализе), који се користе за тестирање биолошких система (*pibas:DNAandRNAanalysisAssay* и *pibas:ProteinAnalysisAssay*) и потклasу *pibas:CellCultureAssay* која дефинише есеје (ћелијске тестове), који се користе у експерименталним приступима. Класе трећег (четвртог и петог) хијерархијског нивоа су поткласе другог (трећег и четвртог) хијерархијског нивоа. Користећи предности OWL синтаксе дефинисане су и дисјунктне класе. На пример, класе *pibas:Rat* и *pibas:Fish* су представљене као дисјунктне. У оквиру PIBAS онтологије дефинисана су и одговарајућа објектна и својства типа података. Табела 4.1 представља њихове домене, односно кодомене.

Табела 4.1 Домени (*rdfs:domain*) и кодомени (*rdfs:range*) објектних својстава (*object property*) и својстава типа података (*datatype property*) дефинисаних у PIBAS онтологији

Објектна својства (*object property*)		
Назив	**Домен (*rdfs:domain*)**	**Кодомен (*rdfs:range*)**
pibas:activeSubstance	*pibas:TypeOfTreatmentInExperiment*	*pibas:ActiveSubstance*
pibas:dose	*pibas:TypeOfTreatmentInExperiment*	*pibas:Dose*
pibas:drug	*pibas:TypeOfTreatmentInExperiment*	*pibas:Drug*
pibas:experimentalMethod	*pibas:Experiment*	*pibas:ExperimentalMethod*
pibas:functionalSystem	*pibas:TypeOfTreatment*	*pibas:FunctionalSystem*
pibas:material	*pibas:ExperimentalMethod*	*pibas:Material*
pibas:modelSystem	*pibas:TypeOfTreatment* *pibas:ExperimentalMethod*	*pibas:ModelSystem*
pibas:patology	*pibas:TypeOfTreatmentInExperiment*	*pibas:Pathology*
pibas:protocol	*pibas:ExperimentalMethod*	*pibas:MolecularBiologyProtocol*
pibas:signalingPathways	*pibas:TypeOfExperiment*	*pibas:SignalingPathways*
pibas:treatment	*pibas:ModelSystem*	*pibas:Treatment*
pibas:typeOfExperiment	*pibas:Experiment*	*pibas:TypeOfExperiment*
pibas:typeOfTreatmentInExperiment	*pibas:ModelSystem*	*pibas:TypeOfTreatmentInExperiment*
Својства типа података (*datatype property*)		
Назив	**Домен (*rdfs:domain*)**	**Кодомен (*rdfs:range*)**
pibas:comment	*pibas:Experiment*	*xsd:string*
pibas:ID	*pibas:Experiment*	*xsd:integer*
pibas:protocolId	*pibas:ExperimentalMethod*	*xsd:string*
pibas:result	*pibas:Experiment*	*xsd:string*
pibas: *storingConditionOfActiveSubstance*	*pibas:Experiment*	*xsd:string*
pibas:theAimOfExperiment	*pibas:Experiment*	*xsd:string*

4.3 CPCTAS база података

Лабораторија обавља многе реалне експерименте у складу са PIBAS онтологијом, што је резултовало бројним скуповима података. Сваки спроведени експеримент имао је форму извештаја[31] (Word документа), који представља његов детаљни опис укључујући податке о називу експеримента, активним супстанцама, модел системима, анализама (тестовима), материјалу, коришћеним методама и експерименталним приступима, времену извршавања експеримента, наручиоцима експеримента, особљу Лабораторије које је било задужено за спровођење експеримента итд. Због овакве организације било је пожељно сместити податке појединачних експеримента у посебне OWL фајлове (онтологије). С обзиром да су извештаји поседовали неке додатне информације у односу на структуру PIBAS онтологије извршено је увођење нових класа и својстава (Табела 4.2).

[31] С обзиром да су извештаји експеримената заштићени од јавности нећемо се бавити њиховим детаљним садржајем.

Табела 4.2 Новодефинисане класе, објектна својства (*object property*) и својства типа података (*datatype property*) у CPCTAS бази података

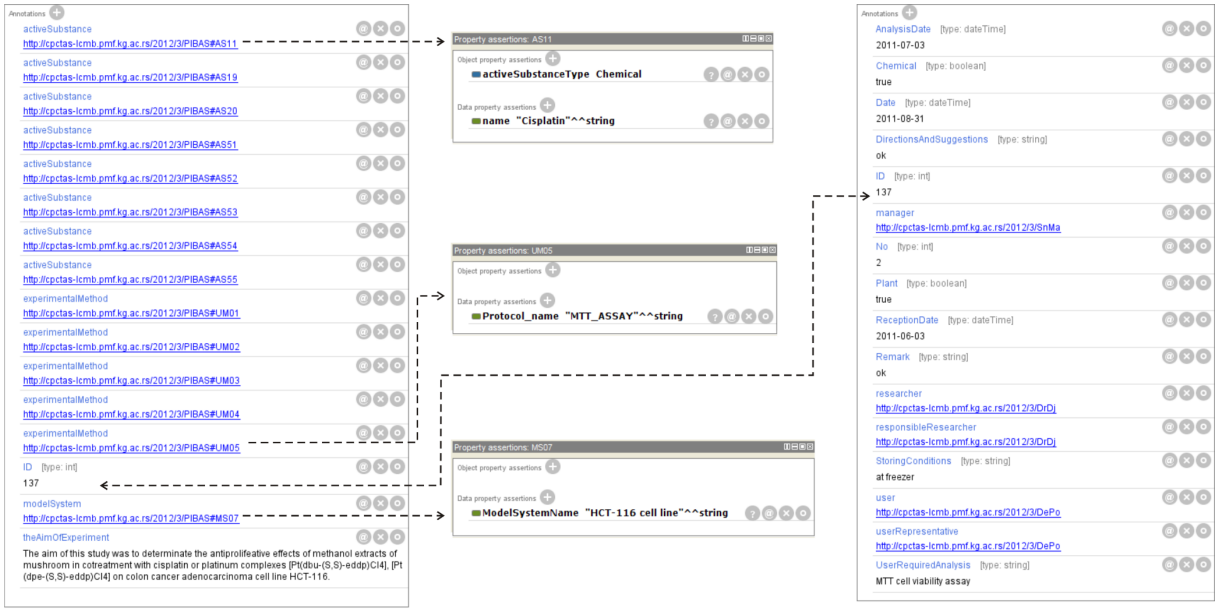
Класе			**Онтологија**
Назив			
pibas:Institution			*PibasHuman.owl*
pibas:Person			*PibasHuman.owl*
pibas:ModelSystemType			*ModelSystem.owl*
pibas:ActiveSubstanceType			*ActiveSubstance.owl*
pibas:Organization			*PibasAddress.owl*
Објектна својства (*object property*)			
Назив	**Домен (*rdfs:domain*)**	**Кодомен (*rdfs:range*)**	
pibas:institution	*pibas:Experiment*	*pibas:Institution*	*PibasHuman.owl*
pibas:manager	*pibas:Experiment*	*pibas:Person*	*PibasHuman.owl*
pibas:researcher	*pibas:Experiment*	*pibas:Person*	*PibasHuman.owl*
pibas:responsibleResearcher	*pibas:Experiment*	*pibas:Person*	*PibasHuman.owl*
pibas:user	*pibas:Experiment*	*pibas:Person*	*PibasHuman.owl*
pibas:userRepresentative	*pibas:Experiment*	*pibas:Person*	*PibasHuman.owl*
pibas:modelSystemType	*pibas:ModelSystem*	*pibas:ModelSystemType*	*ModelSystem.owl*
pibas:activeSubstanceType	*pibas:ActiveSubstance*	*pibas:ActiveSubstanceType*	*ActiveSubstance.owl*
Својства типа података (*object property*)			
Назив	**Домен (*rdfs:domain*)**	**Кодомен (*rdfs:range*)**	
pibas:AnalysisDate	*pibas:Experiment*	*xsd:date*	*PibasHuman.owl*
pibas:Date	*pibas:Experiment*	*xsd:date*	*PibasHuman.owl*
pibas:DirectionsAndSuggestions	*pibas:Experiment*	*xsd:string*	*PibasHuman.owl*
pibas:ExpTitle	*pibas:Experiment*	*xsd:string*	*PibasHuman.owl*
pibas:ReceptionDate	*pibas:Experiment*	*xsd:date*	*PibasHuman.owl*
pibas:UserRequiredAnalysis	*pibas:Experiment*	*xsd:string*	*PibasHuman.owl*
pibas:Web	*pibas:Experiment*	*xsd:string*	*PibasHuman.owl*
pibas:StoringConditions	*pibas:Experiment*	*xsd:string*	*PibasHuman.owl*
pibas:Remark	*pibas:Experiment*	*xsd:string*	*PibasHuman.owl*
pibas:Protocol_name	*pibas:ExperimentalMethod*	*xsd:string*	*Protocol.owl*
pibas:name	*pibas:ModelSystem*	*xsd:string*	*ModelSystem.owl*
pibas:name	*pibas:ActiveSubstance*	*xsd:string*	*ActiveSubstance.owl*

Инстанце класа *pibas:ModelSystem*, *pibas:ActiveSubstance, pibas:ExperimentalMethod, pibas:Institution*, *pibas:Person* и *pibas:Organization* смештене су у одговарајуће онтологије - *ModelSystem.owl*, *ActiveSubstance.owl*, *Protocol.owl*, *PibasHuman.owl* и *PibasAddress.owl*. У овим онтологијама инстанце су презентоване на основу својстава која представља Табела 4.2. Инстанце класе *pibas:Person* служе се предикатима *foaf*[32] речника, који се користи за представљање људских ресурса. Подаци појединачних експеримената су такође смештени у посебне OWL фајлове (*expID.owl*[33]), који су преко објектних својстава *pibas:activeSubstance, pibas:experimenatlMethod* и *pibas:modelSystem* повезани са инстанцама новодефинисаних онтологија: *ModelSystem.owl*, *ActiveSubstance.owl* и *Protocol.owl*. За сваки експеримент дефинисан је и додатни *expHrID.owl* фајл, који представља информације о људским ресурсима који садржи инстанце *PibasHuman.owl* онтологије. Овај фајл је у корелацији са *expID.owl* фајлом захваљујући својству *pibas:ID*. На овај начин се постиже ефекат треће нормалне форме, као код релационих база података. То заправо значи да уколико треба модификовати објектну вредност неког концепта, промена се обавља само на једном месту. Слика 4.3 представља пример једног експеримента[34] из Protégé едитора, укључујући и његову интеракцију ка екстерним онтологијама у CPCTAS бази података.

32 *http://xmlns.com/foaf/spec/*

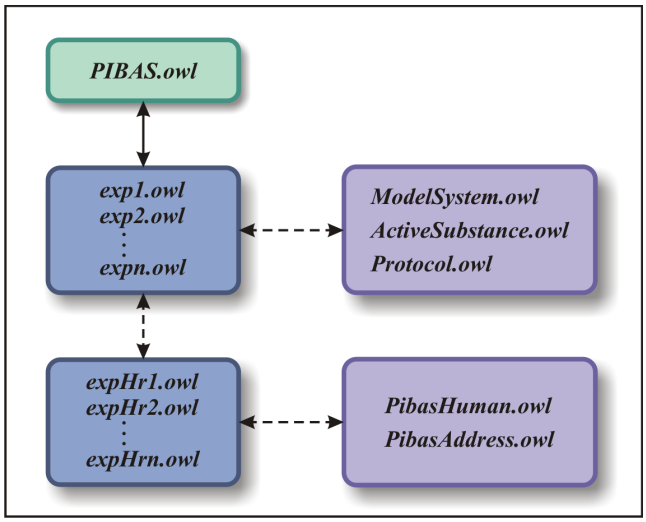
33 *ID* представља идентификациони број експеримента.

34 *pibas:Antiproliferative_effects_of_mushroom_methanol_in_cotreatment_with_platinum_complexes_on_HCT-116_cell_line*



Слика 4.3 Приказ експеримента (инстанце) из *Protégé* едитора (лево) у корелацији са екстерним онтолошким фајловима (*ModelSystem.owl*, *ActiveSubstance.owl*, *Protocol.owl* и *expHr137.owl*) у CPCTAS бази података

Дакле, CPCTAS база података садржи *PIBAS.owl* онтологију (која представља структуру експеримента без инстанци), онтологије *ModelSystem.owl*, *ActiveSubstance.owl*, *Protocol.owl*, *PibasHuman.owl* и *PibasAddress.owl* (са својим инстанцама), као и онтолошке фајлове (*expID.owl* и *expHrID.owl*) који садрже податке појединачних експеримената. Слика 4.4 представља декомпозицију базе података.



Слика 4.4 Декомпозиција CPCTAS базе података

4.3.1 SPARQL упити над CPCTAS базом података

Сви фајлови CPCTAS базе података су смештени на JOSEKI SPARQL серверу. Претрага базе је омогућена задавањем SPARQL упита на PIBAS *endpoint*[35]-у. Слика 4.5 представља пример SELECT SPARQL упита, који преузима све експерименте (називе експеримената представљене предикатом *pibas:ExpTitle*) и њихове линкове (PDF/Word извештаје представљене предикатом *pibas:Web*), који задовољавају услове дефинисане у WHERE клаузули. На пример, услов да је наручилац експеримената институција *Medical Herbs* представљен је обрасцем *?exp pibas:institution pibas:Medical_Herbs_in_Novi_Sad*; услов да су експерименти обављени применом протокола *UM05*

35 *http://cpctas-lcmb.pmf.kg.ac.rs:2020/*

представљен је обрасцем *?exp pibas:experimentalMethod pibas:UM05*; услов да су експерименти обављени применом аналитичке методе *MTT ASSAY* представљен је обрасцем *?exp pibas:UserRequiredAnalysis "MTT cell viability assay"*; услов да су експерименти обављени над активном супстанцом *Cisplatin* представљен је обрасцем *?exp pibas:activeSubstance pibas:AS11*. Такође, упит врши и филтрирање резултата по датуму извођења експеримента и за то се користи резервисана реч FILTER. Упит садржи кориснички дефинисан именски простор *pibas*, као и уграђене именске просторе *xsd* и *rdf*.

```
PREFIX pibas: <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX xsd: <http://mar.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://mar.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?value ?search
WHERE
  {
    ?exp rdf:type pibas:Experiment.
    ?exp pibas:ExpTitle ?search.
    ?exp pibas:Web ?value.
    ?exp pibas:Date ?date.
    ?exp pibas:institution pibas:Medical_Merbs_in_Novi_Sad.
    ?exp pibas:experimentalMethod pibas:UMO5.
    ?exp pibas:modelSystem pibas:MS07.
    ?exp pibas:activeSubstance pibas:AS11.
    ?exp pibas:user pibas:Dejan_Poslon.
    ?exp pibas:userRepresentative pibas:Dejan_Poslon.
    ?exp pibas:UserRequiredAnalysis "MTT cell viability assay".
    ?exp pibas:researcher pibas:Dragana_Seklic.
    ?exp pibas:responsibleResearcher pibas:Dragana_Seklic.
    ?exp pibas:Chemical ?a.
    ?exp pibas:experimentalMethod ?expmet.
    ?expmet pibas:protocolType pibas:CellCultureAssay.
    FILTER (?date > "2011-08-30"^^xsd:date)
    FILTER (?date < "2011-10-15"^^xsd:date)
  }
ORDER BY ?date
```

Слика 4.5 Пример SPARQL упита дефинисаног над CPCTAS базом података чији резултат извршавања чине експерименти који задовољавају услове дефинисане у WHERE клаузули

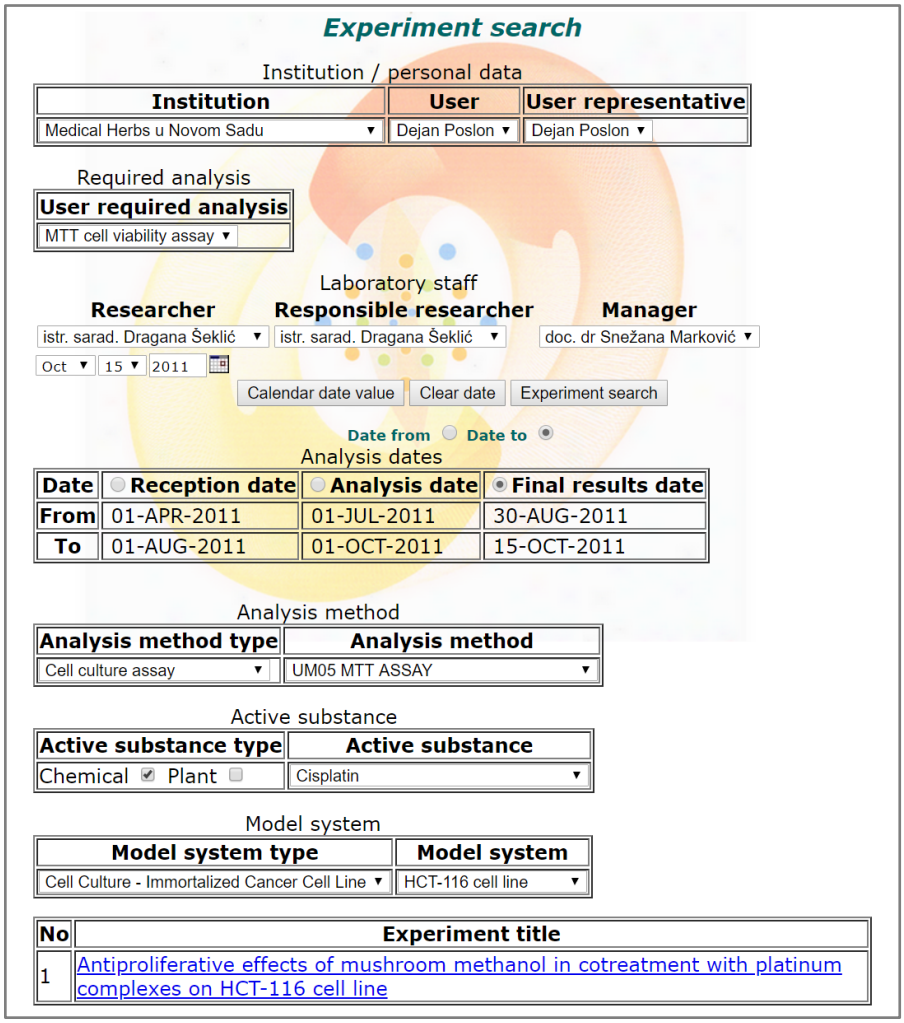
Уз помоћ SPARQL-а могуће је креирати читав дијапазон упита, који задовољавају различите захтеве и интересовања корисника Лабораторије. На пример, може се креирати упит који открива све експерименте који су спроведени за дату активну супстанцу, упит који открива све модел системе који су коришћени у оквиру неког експеримента, упит који открива све есеје који се користе у експериментима за дату активну супстанцу итд. Међутим, како SPARQL синтакса није једноставна за кориснике који немају искуства са семантичким технологијама, имплементиран је прототип софтвера који омогућава једноставно и интуитивно извођење SPARQL упита.

4.4 Прототип софтвера за претрагу CPCTAS базе података

Прототип софтвера *Experiment Search*[36] развијен је са циљем да олакша претрагу CPCTAS базе и омогући праћење свих резултата рада Лабораторије. Он је замишљен као веб страница на којој корисници могу одабрати одговарајуће критеријуме претраге (дефинисане класама и предикатима), а који одговарају инстанцама базе. На самом почетку, када није селектован ниједан критеријум претраге, приказује се комплетна листа експеримената (PDF/Word извештаја). Они се даље филтрирају задавањем одговарајућих критеријума претраге. Критеријуми су подељени у следеће групе: *Institution/personal data*, *Required analysis*, *Dates*, *Analysis method*, *Active substance* и *Model system*. Свако ограничење из било које групе се може комбиновати са било којим другим ограничењем. Нека ограничења су логички зависна: одабрана институција аутоматски ограничава инстанце *User/UserRepresentative* и *User/RequiredAnalysis*; ставка *Active substance type* (*Chemical* или *Plant*) ограничава тип активне супстанце; ставка *Model System* ограничава тип модел система; ставка *Analysis method* ограничава тип аналитичке методе. Таква зависност помаже корисницима да боље искомбинују потребне услове претраге. Слика 4.6 представља пример интерфејса са селектованим критеријумима који одговарају условима упита (Слика 4.5). Након

[36] *http://cpctas-lcmb.pmf.kg.ac.rs/LCMB/expSearch.html*

постављених ограничења и клика на дугме за претрагу (*Experiment search*) у позадини се креира и извршава одговарајући SELECT SPARQL упит, а листа експеримената се филтрира на основу задатих критеријума. Експерименти су линковани и када се кликне на дати линк, преузима се комплетан PDF/Word извештај датог експеримента.



Слика 4.6 Кориснички интерфејс прототипа софтвера за претрагу CPCTAS базе података

Сврха имплементираног прототипа софтвера јесте преузимање комплетних извештаја појединачних експеримената. Из датих извештаја корисници могу открити детаљније информације о биолошким системима, методама, типовима есеја, ћелијских линија, IC_{50} вредностима итд. који су коришћени или остварени у актуелним експериментима. На основу резултата претраге корисници могу доћи до значајних информација и сазнања који могу утицати на планирање и извођење будућих експеримената, онако како је то дефинисано у одељку 2.2.2.

На основу спроведених истраживања можемо закључити да су технологије семантичког веба одиграле значајну улогу у процесу представљања и претраге података за потребе Лабораторије. Коришћењем OWL синтаксе омогућено је једноставно креирање онтологија и декомпозиција базе. Декомпозиција омогућава лако проширење саме базе, што је са друге стране могуће с обзиром на карактеристику проширивости (могуће екстензије) PIBAS онтологије. Ово има важну улогу у процесу интеграције CPCTAS базе података. Развијена база података поред епитета „онтолошке“ има и епитет „базе знања“ која би требало да обезбеди знање за планирање будућих експеримената. Упитни језик SPARQL је од велике важности јер омогућава претрагу базе на једноставан и лак начин, пружајући корисницима информације од интереса. База података би требало да има карактеристику транспарентности, што би значило да подаци морају да буду доступни заинтересованим истраживачима, како би били подложни

реалним критикама, као и коришћењу у другим истраживањима. Подаци који су коришћени у процесу интеграције (одељак 4.5) искључиво су тест подаци. Научни доприноси након фазе развоја PIBAS онтологије и CPCTAS базе података су истраживања [36,73,74,75,76,77].

4.5 Интеграција CPCTAS базе података

У овом одељку је најпре извршен преглед актуелних биоинформатичких репозиторијума и њихових онтолошких база података како би се указало на значај семантичких технологија у представљању и интеграцији података. Након тога се описује процес интеграције CPCTAS базе података са неким од представљених репозиторијума. На крају се анализирају и циљеви интеграције, који су у комбинацији са претходно презентованим прототипом софтвера иницирали развој Платформе представљене у дисертацији.

4.5.1 Биоинформатички репозиторијуми

У домену биоинформатике тренутно постоји релативно велики број јавно доступних база података са свакодневном тенденцијом раста [78]. Актуелне базе се развијају аутономно или су делови великих репозиторијума. Свака база је моделована за специфичне потребе, различите је величине и намене и као таква од велике је важности за истраживачку заједницу. Међутим, многе базе у домену биоинформатике суочавају су са различитим проблемима. Најчешћи заједнички проблеми су редундантност података, константне промене података, некомплетне информације или некоректни линкови које оне садрже, а проблем хетерогености постоји у многим аспектима, укључујући форматирање података, конвенцију и значење [78]. Због ових фактора традиционални приступи за претрагу података често су производили незадовољавајуће резултате. Технологије семантичког веба уклањају актуелне проблеме дефинишући стандарде који олакшавају процесе имплементације, интеграције и откривања (претраге) биоинформатичких података. Два главна семантичка стандарда у биоинформатици су RDF и OWL (онтологије) и многе организације и научно-истраживачке институције их примењују у циљу решавања наведених проблема. У наставку су представљени неки актуелни семантички репозиторијуми.

Bio2RDF [79] је *open-sourc*e пројекат који користи технологије семантичког веба за изградњу LOD (*Linked Open Data*) података [80]. LOD подаци су структурирани подаци, који се могу међусобно повезати и на тај начин постати доступни биоинформатичкој заједници. Bio2RDF пројекат је у развоју и тренутно је доступна његова верзија 3 (*Bio2RDF Release 3 (July 2014) Release Notes*[37]). Он конвертује примарне базе података из домена природних наука (међу којима су најзначајније NCBI Gene[38], KEGG[39], PharmgKB[40], PDB[41] и DrugBank[42]) користећи RDF као модел података. Процес конвертовања база се реализује дефинишући скуп једноставних конвенција за креирање компатибилних RDF докумената у форми нормализованих URI спецификација. Нормализовани URI има форму облика *http://bio2rdf.org/namespace:id*, где *namespace* одговара називу оригиналне базе података, док *id* представља идентификатор концепта. Рецимо, *http://bio2rdf.org/drugbank:DB00544* (*drugbank:DB00544*) је пример нормализоване URI спецификације, који одговара запису *https://www.drugbank.ca/drugs/DB00544* (лек *Fluorouracil*) DrugBank базе података. Користећи дати процес конвертовања Bio2RDF се може интегрисати са било којом базом података чиме се омогућава раст LOD података. Ово је свакако од велике важности како за семантички веб, тако и за домен биоинформатике. Bio2RDF тренутно садржи преко 2,3 милијарди триплета организованих у 35 база података. Подацима се може приступити извођењем SPARQL упита на одговарајућим *remote endpoint*[43]-има.

---

[37] *http://bio2rdf.org/*

[38] *https://www.ncbi.nlm.nih.gov/gene*

[39] *http://www.genome.jp/kegg/kegg1.html*

[40] *https://www.pharmgkb.org/*

[41] *http://www.rcsb.org/pdb/home/home.do*

[42] *https://www.drugbank.ca/*

[43] Листа свих endpoint-а Bio2RDF пројекта доступна је на *http://download.bio2rdf.org/files/release/3/release.html*.

LODD (*Linking Open Drug Data*) [81] је пројекат организације HCLS IG[44] (*W3C Semantic Web Health Care and Life Sciences Interest Group*), који повезује RDF податке из LCT[45] (*Linked Clinical Trials*) базе података са многим другим изворима података. Ова база садржи податке који се односе на развој лекова и клиничко истраживање. LODD репозиторијум има више од 8,4 милиона RDF триплета и 388000 линкова ка екстерним скуповима података. Слични су и пројекти LinkHub [82] и BioGateway [83].

Chem2Bio2RDF [39] је јединствени репозиторијум који је настао агрегацијом података из хемогеномских база укључујући PubChem Bioassay[46], DrugBank, KEGG Ligand, CTD[47], BindingDB[48], PharmGKB, MATADOR[49] као и великог броја малих QSAR[50] (*Quantitative structure-activity relationship*) база података доступних на вебу. Базе података Chem2Bio2RDF репозиторијума су организоване у шест категорија одређених на основу врста биолошких и хемијских концепата које они садрже: хемикалије и лекови, протеини и гени, хемогеномика, системи (ћелијске линије), фенотипови (болести и нуспојаве) и литература. Chem2Bio2RDF је развио сопствену шему[51] за класификацију концепата и RDF ресурса, и његови подаци су представљени у форми триплета. Chem2Bio2RDF је додатно интегрисао податке из LODD и Bio2RDF репозиторијума користећи *owl:sameAs* синтаксу. Тако је нпр. лек *Fluorouracil* (*drugbank_drug:DB000544*) у DrugBank бази преко *owl:SameAs* синтаксе интегрисан са лековима у LODD (*drugbank_drug:DB00544 owl:sameAs lodd:DB00544*) и Bio2RDF (*drugbank_drug:DB00001 owl:sameAs drugbank:DB00544*) репозиторијумима. Chem2Bio2RDF тренутно има 27 база података и преко 78 милиона триплета. Претрага података је могућа извршавањем SPARQL упита над одговарајућим *remote endpoint* [52]-има.

EMBL-EBI [84] је репозиторијум који пружа приступ свеобухватним и јавно доступним молекуларним базама података покривајући при томе читав спектар биоинформатичког знања (биолошке податке, супстанце, различите хемијске структуре итд.). Најзначајније базе података овог репозиторијума су Reactome[53], UniProt[54] и ChEMBL[55]. UniProt садржи високо квалитетне ресурсе о секвенцама протеина. RDF репрезентација ове базе базирана је на UCV[56] (*UniProt Core Vocabulary*) речнику. Reactome је мануелно креирана база података људских ћелијских путева и реакција (енгл. *pathways*). RDF репрезентација ове базе је заснована на *BioPAX Level 3 Reactome*[57] верзији. ChEMBL база података се може описати као хемијска база података биоактивних молекула са особинама сличним лековима. RDF репрезентација ове базе користи интерну CCO онтологију (*ChEMBL Core Ontology*) [85]. Ова онтологија је интегрисана са екстерним онтологијама, као што су: BAO[58] (*BioAssay Ontology*), ChEBI Ontology[59], CHEMINF[60] (*Chemical Information Ontology*), *Bibliographic Ontology*[61], UO[62] (*Unit Ontology*), QUDT

44 *https://www.w3.org/2001/sw/hcls/*

45 *https://clinicaltrials.gov/*

46 *https://pubchem.ncbi.nlm.nih.gov/help.html#PubChem_BioAssay_Database*

47 *http://www.ctsdatabase.com/*

48 *http://www.bindingdb.org/bind/index.jsp*

49 *http://matador.embl.de/*

50 *https://en.wikipedia.org/wiki/Quantitative_structure%E2%80%93activity_relationship*

51 *http://cheminfov.informatics.indiana.edu:8080//download.html*

52 *http://chem2bio2rdf.wikispaces.com/SPARQL+Endpoints*

53 *http://www.reactome.org/*

54 *https://www.uniprot.org/*

55 *https://www.ebi.ac.uk/chembl/*

56 *http://lov.okfn.org/dataset/lov/vocabs/uniprot*

57 *http://www.biopax.org/release/biopax-level3-documentation.pdf*

58 *http://bioassayontology.org/*

59 *http://www.ebi.ac.uk/chebi/*

60 *https://bioportal.bioontology.org/ontologies/CHEMINF*

61 *http://bibliontology.com*

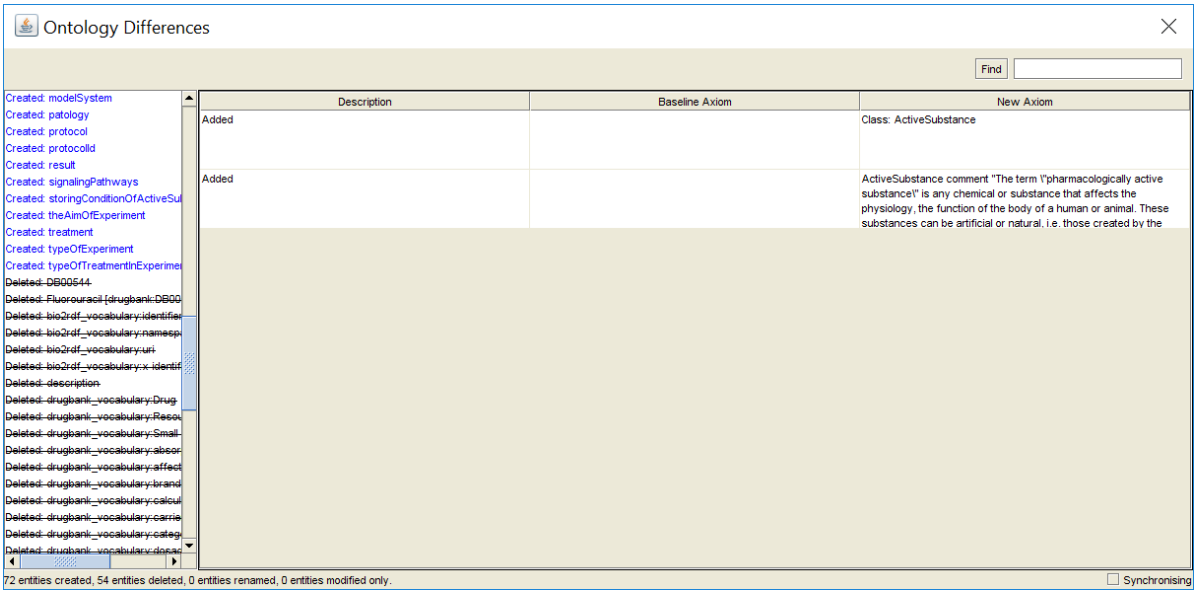
62 *https://code.google.com/archive/p/unit-ontology/*

Ontology[63] и SIO[64] (*Semantic Science Ontology*). Сваки лек у ChEMBL бази података представљен је преко одговарајућих атрибута, који се односе на IUPAC номенклатуру[65] (текстуалне идентификаторе једињења као што су *SMILES*, *InChi* и *InChiKey*), механизме деловања, физичко-хемијске особине једињења итд. EMBL-EBI представља своје базе података као повезане RDF скупове података, а корисницима омогућава да извођењем SPARQL упита кроз *remote endpoint*[66] истраже и открију релевантне податке неопходне за даља истраживања. EMBL-EBI платформа нуди и различите сервисе[67], који се односе на претрагу (*EBI Search*), анализу (*Pratt*) и статистику података (*SAPS*).

4.5.2 Поступак интеграције

Многи истраживачи су мотивисани да податке добијене из својих истраживања поделе са остатком истраживачке заједнице и да их интегришу са другим доступним онтолошким подацима, показујући на тај начин да су њихова истраживања актуелна на биоинформатичкој сцени. Анализом претходног одељка може се закључити да постоји велика повезаност у тематици и примени семантичких технологија између популарних репозиторијума и CPCTAS базе због чега је постојала посебна мотивација за интеграцијом.

У циљу интеграције је најпре извршена компарација онтологија CPCTAS базе података са онтолошким базама података које припадају Bio2RDF, Chem2Bio2RDF и EMBL-EBI репозиторијумима. За овај приступ коришћен је Protégé алат OWLDiff[68]. Дати алат је генерално намењен за компарацију верзија исте онтологије, али се може искористи и за поређење онтологија које припадају различитим именским просторима. Слика 4.7 представља резултат компарације онтолошких база PIBAS/CPCTAS и Drugbank/Bio2RDF. На основу резултата компарације уочено је да обе онтологије садрже класу *ActiveSubstance*. На исти начин се могла уочити аналогија између класа које се односе на биолошке системе.



Слика 4.7 Резултат поређења онтолошких база података Drugbank/Bio2RDF и PIBAS/CPCTAS применом *OWLDiff* алата *Protégé* едитора

Даља намера компарације била је утврђивање односа (сличности) између инстанци одговарајућих класа, које су послужиле као добар повод за процес интеграције. Процес интеграције је започео екстензијом

63 *http://www.qudt.org/*

64 *https://github.com/micheldumontier/semanticscience*

65 *https://en.wikipedia.org/wiki/International_Union_of_Pure_and_Applied_Chemistry*

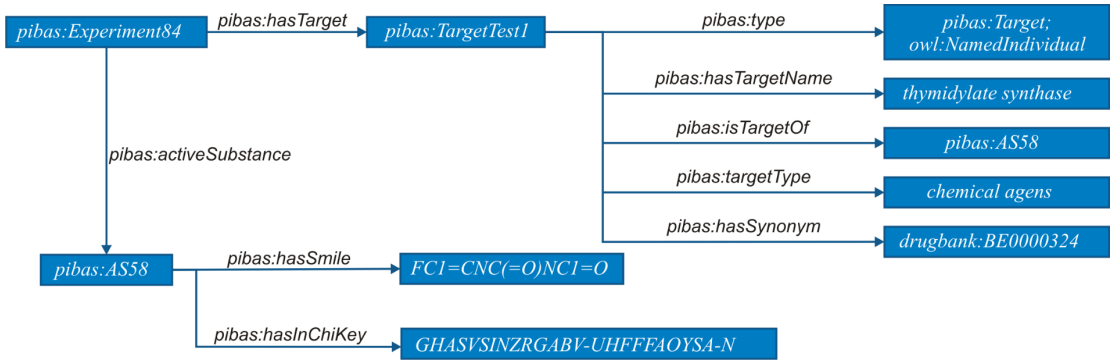
66 *https://www.ebi.ac.uk/rdf/services/sparql*

67 *https://www.ebi.ac.uk/services/all*

68 *https://protegewiki.stanford.edu/wiki/OWLDiff*

PIBAS онтологије. Класа *pibas:ActiveSubstance* је проширена новим својствима типа података *pibas:hasSmile* и *pibas:hasInChiKey*, с обзиром на чињеницу да су корисници Лабораторије користили IUPAC номенклатуру (*InChiKey* и *SMILES* параметре) за интерно текстуално идентификовање активних супстанци. Потом је дефинисана класа *pibas:Target* која представља биолошке мете који се користе у Лабораторији. Ова класа има дефинисано својство типа података *pibas:hasTargetName* која означава име биолошке мете, својство типа података *pibas:targetType* које означава тип биолошке мете, објектно својство *pibas:isTargetOf* које за кодомен има класу *pibas:ActiveSubstance* и својство типа података *pibas:hasSynonym* које указује на потенцијалне синониме - инстанце других база података. Ово својство обезбеђује интеграцију са више инстанци истовремено. Треба напоменути да је класа *pibas:Target* дефинисана као еквивалент класе *pibas:modelSystem*, иако је класа *pibas:modelSystem* више оријентисана ка имортализованим ћелијским линијама канцера, а не и ка другим потенцијалним типовима биолошких мета. Класа *pibas:Experiment* проширена је објектним својством *pibas:hasTarget* које за кодомен има класу *pibas:Target.* У процесу интеграције нису разматране уочене сличности између инстанци које се односе на ћелијске линије и есеје. Током компаративног процеса нису уочени заједнички концепти који припадају уграђеним именским просторима.

Слика 4.8 представља пример интеграције: инстанца *pibas:Experiment84* (класе *pibas:Experiment*) је преко објектног својства *pibas:activeSubstance* повезана са инстанцом *pibas:AS58* (класе *pibas:ActiveSubstance*); инстанца *pibas:AS58* је додатно описана својствима *pibas:hasInchiKey* и *pibas:hasSmile*; инстанца *pibas:Experiment84* је преко објектног својства *pibas:hasTarget* повезана са инстанцом *pibas:TestTarget1* (класе *pibas:Target*); инстанца *pibas:TargetTest1* је преко својства *pibas:hasSynonym* повезана са инстанцом класе DrugBank/Bio2RDF (*drugbank:BE0000324*) и типа је *chemical agens* (*pibas:targetType*). Овакав приступ је донекле паралелан са идејом интеграције Chem2Bio2RDF репозиторијума. На овај начин се CPCTAS база података може интегрисати са било којом другом онтолошком базом. Подаци који су овом случају коришћени за процес интеграције искључиво су тест подаци, како би се заштитила ауторска права CPCTAS базе података.



Слика 4.8 Графички приказ интеграције PIBAS/CPCTAS базе података са DrugBank/Bio2RDF базом података

Поступком интеграције, CPCTAS база података се придружила визији LOD података, чиме је њено „знање“ потенцијално постало доступно биоинформатичкој заједници. Над скупом овако повезаних података могуће је изводити различите претраге релевантних података. Употреба Federated SPARQL упита допринела би откривању чињеница које могу утицати на процес планирања рационалног дизајна лекова, односно преклиничког тестирања супстанци.

4.5.3 Циљеви претрага биоинформатичких база података

Подаци доступни у биоинформатичким онтолошким базама података могу у великој мери олакшати читаву студија развоја лекова [86]. За кориснике Лабораторије (или било које друге лабораторије која се бави сличном тематиком) током истраживачког процеса кључне су претраге података које би омогућиле:

a) Откривање биолошких мета који су у интеракцији са одговарајућим лековима
b) Откривање есеја који су у интеракцији са одговарајућим лековима

c) Откривање ћелијских линија који су у интеракцији са одговарајућим лековима
d) Откривање информација о лековима
e) Откривање публикација

Ове захтеве називамо шаблонима (енгл. *template*) и користимо ознаку (*) уколико их наводимо у целини. У наставку је представљен њихов значај за процес преклиничког тестирања лекова.

Шаблон (*a*) је типична полазна тачка савременог процеса откривања лека. Пуно труда је уложено у процес идентификације биолошких мета и овај проблем је био нарочито присутан пре постојања онтолошких база података и технологија семантичког веба. Валидација биолошких мета у традиционалним приступима подразумевала је примену одговарајућих стратегија, као што су генски нокаут или директна инхибиција мета уз помоћ малих молекула, пептида, антитела или биле које друге класе инхибитора [87]. Ипак, ови традиционални приступи могу бити прилично дуготрајни и довести до негативних резултата, што за консеквенцу може имати велики утрошак ресурса. Други начин је да се одабир биолошких мета врши на основу различитих критеријума који су представљени у одељку 2.2.2. Подаци онтолошких база података могу се користити као доказ провере ових критеријума [86]. На пример, онтолошке базе података се могу искористити да би се утврдило да ли су одговарајуће биолошке мете повезане са болестима, односно да ли су коришћене за одговарајућа експериментална истраживања. Лековитост се може проценити кроз интегративну анализу функционалне класе протеина, у смислу да ли постоји везујућа страна у протеину (енгл. *binding site*) која се може користити за везивање малих молекула [88,86]. „*Помоћу ових података могу се пронаћи структуре нових молекула које ће се јаче везивати и боље модулирати активност биолошког мете*“ [89]. Јавно доступне онтолошке базе података омогућавају откривање биолошких мета које се користе у „успешно“ спроведеним експериментима, односно у оним експериментима код којих је IC_{50} вредност нижа до 30 μM (вредност профилисана на основу искустава корисника Лабораторије).

Шаблон (*b*) проналази есеје који су у интеракцији са одговарајућим леком. Идентификација есеја који су у интеракцији са неким леком може сугерисати на избор есеја у експерименталним истраживањима за побољшану биолошку мету или за сличне лекове. Корисник на основу тих података може побољшати одређене компоненте, које би омогућиле још ефикасније експерименте.

Шаблон (*c*) пронализи ћелијске линије које су у интеракцији са одговарајућим леком. Избор ћелијске линије захтева успостављање баланса између избора одговарајућег модела и одабира ћелијске линије са којом се може радити [90]. Фактори као што су услови културе, стопа раста, морфологија и подћелијска структура као и сами трошкови могу имати велики утицај на експерименте и њихов крајњи резултат [90]. Традиционално, истраживачи доносе одлуке о ћелијским линијама које се користе у експерименталним приступима. Међутим, иницијални план можда не произведе очекиване резултате и стога се често мора приступити алтернативним ћелијским линијама. Кроз процес анализе информација о ћелијским линијама, доступним кроз одређене онтолошке базе, корисник има квалитетнији увид у циљу планирања својих експеримената. Корисник на основу тих података може побољшати неку компоненту ћелијске линије, а на тај начин и само истраживање.

Шаблон (*d*) омогућава откривање информација о лековима. Онтолошке базе података садрже различите информације повезане са структуром и активношћу активних супстанци (лекова), као што су молекулска маса, физичко-хемијски параметри, подаци које се односе на биолошку активност, афинитет, токсичност, фармакокинетику[69], спектралне податке итд. [89]. „*Молекулске особине једињења могу се довести у везу са биолошким дејством и активношћу, а на основу њих се, за дату структуру, може предвидети сличност леку, растворљивост, токсичност, фармакокинетика итд. Прикупљени подаци могу се употребити у виду анализе података, корелације и дизајна модела. Оваквим поступком се сужава простор за тестирање молекула, те се фокусира и убрзава експериментални део потраге за што ефикаснијом активном супстанцом*“ [89]. Један од најпознатијих начина за одређивање потенцијално

[69] *https://en.wikipedia.org/wiki/Pharmacokinetics*

добрих супстанци јесте покушај да се утврди да ли супстанца задовољава правило 5 Липинског (енгл. *rule of 5*), које представља сет емпиријских правила која сугеришу да ли супстанца може имати проблем са биорасположивошћу[70] (удео примењене дозе лека који у неизмењеном облику доспева у системску циркулацију) [89]. Супстанца ће вероватно имати мању биорасположивост ако су два или више од ових правила прекршена [89]: (1) логаритам дистрибуционог односа октанол-вода ($logP$) ≤ 5 ($logP$ је мера хидрофилности или хидрофобности хемијске супстанце; партициони коефицијенти су корисни за процену дистрибуције лека у телу); (2) молекулска маса ≤ 500, (3) број акцептора водоничних веза $(HBA) \leq 10$, (4) број донора водоничнх веза $(HBD) \leq 5$.

Шаблон (*e*) је од изузетне важности за кориснике, јер сваки озбиљан научно-истраживачки рад подразумева и адекватан преглед литературе која може утицати на планирање будућих истраживања. Подаци онтолошких база података које се односе на публикације сугеришу актуелна (новија) истраживања на основу датума објаве публикација, сугеришу часописе у којима се могу публиковати нови приступи, омогућавају увид у садржај публикације на основу апстракта и утичу на успостављање евентуалне сарадње између истраживача (аутора), што доприноси популаризацији података мање познатих лабораторија или истраживачких организација. Имајући у виду значај литературних података, CPCTAS база је проширена онтологијама *Reference* базе података [91]. Ова база такође представља један од научних доприноса ове дисертације, и директна је последица истраживања примене технологија семантичког веба у домену управљања научним референцама. Она садржи податке о референцама научног особља запосленог на Природно-математичком факултету, Универзитета у Крагујевцу. Дата база користи *bibo*[71] (*Bibliographic Ontology Specification*) речник за представљање својих података, на сличан начин као и EMBL-EBI [39] репозиторијум.

Осим што се користе као извор референци, онтолошке базе података се у домену биоинформатике интензивно користе за претрагу и имплементацију разних метода и алгоритама, који могу побољшати квалитет биоинформатичких истраживања. С обзиром на тематику и типове података које садрже, CPCTAS база података и базе података које су делови великих актуелних репозиторијума, представљених у одељку 4.5.1, имају велики капацитет за пружање одговора за шаблоне (*). Због тога је од круцијалног значаја њихова ефикасна претрага. Да би се извршила претрага више онтолошких извора података неопходно је применити одговарајуће Federated SPARQL упите. Међутим, коришћење SPARQL синтаксе није нимало охрабрујуће за кориснике који немају искуство са технологијама семантичког веба. Такође, ни само познавање SPARQL синтаксе често није довољан предуслов за креирање упита, јер се од корисника очекује како доменско знање, тако и познавање онтолошке структуре. Како би се извршили упити над одговарајућим базама података неопходно је уложити пуно труда и открити ентитете, релевантне везе између ентитета, интегрисане и редундантне податке, који утичу на формирање коначних упита. Такође, упите је потребно на неки начин повезати са корисницима (слично идеји прототипа софтвера представљеног у одељку 4.4) пружањем једноставног корисничког интерфејса. Са том идејом развијено је софтверско решење PIBAS FedSPARQL (Платформа) [7], које представља главни допринос ове дисертације. Наредна поглавља су посвећена искључиво Платформи - опису архитектуре, начину примене и значају њених метода за шаблоне (*).

70 *https://en.wikipedia.org/wiki/Bioavailability*

71 *http://bibliontology.com/*

5 Биоинформатичка платформа - PIBAS FedSPARQL

Главни резултат дисертације јесте софтверско решење које представља оквир за извршавање Federated SPARQL упита над биоинформатичким базама података као и детекцију сличних података над резултатима (инстанцама онтолошких база података) који су добијени извршавањем упита. Платформа је представљена у раду [7] и настала је као последица истраживања у домену семантичког веба. У овом тренутку она подржава неколико иницијалних упита, али с обзиром на скалабилност могућа је и екстензија додатним упитима, који би омогућили откривање нових знања. Методе Платформе - извршавање предефинисаних Federated SPARQL упита над иницијалним и кориснички селектованим базама података, динамичко филтрирање резултата упита, као и детекција сличних података - могу се применити и у било ком другом домену. У овом поглављу су представљени принципи развоја Платформе, њена архитектура са приказом рада основних компоненти и генералним начином функционисања.

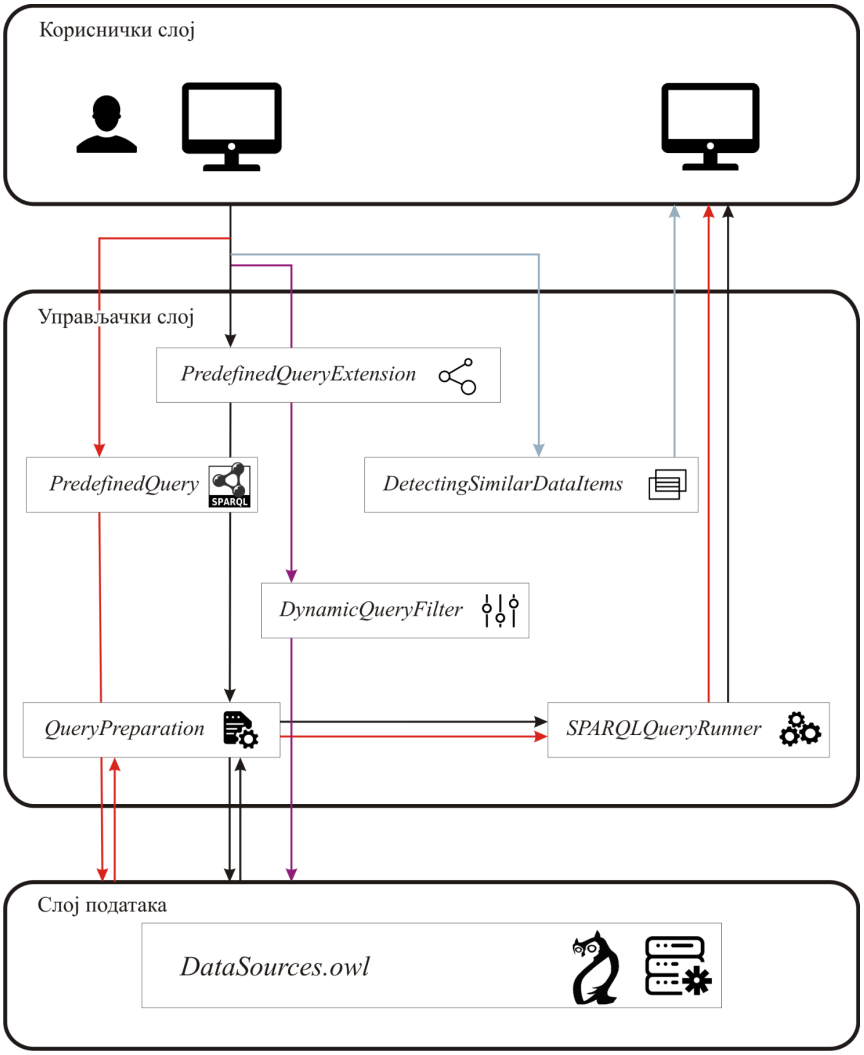
5.1 Принципи развоја Платформе

Одговоре на разна биоинформатичка питања, укључујући и одговоре на шаблоне (*) дефинисане у одељку 4.5.3, пружају многе индустријске или научно-истраживачке институције које поседују сопствене апликације за претрагу интерних база података. Такође, постоје многе јавне апликације (представљене у одељку 6.4) које делимично задовољавају шаблоне (*) и пружају компетентне одговоре комбиновањем технологија семантичког веба и алата за визуелизацију података. Међутим, ове апликације су често ограничене: не дозвољавају дефинисање нових упита, не постоји могућност проширења упита новим базама података или је визуелизација података лимитирана одређеним командама. Како би се делимично отклонили наведени проблеми, развијена је Платформа која је представљена у дисертацији. Платформа је имплементирана са циљем да задовољи следеће:

- Омогући откривање релевантних и комплементарних података који задовољавају шаблоне (*) извођењем предефинисаних Federated SPARQL упита над онтолошким базама података;
- Омогући додавање кориснички селектованих база података у предефинисане упите;
- Омогући динамичко филтрирање резултата на основу RDF структуре база података који учествују у упитима;
- Омогући детекцију сличних података над резултатима упита;
- Прилагоди рад система хетерогеној природи ресурса, односно онтолошких база података које имају велики број триплета у својим RDF графовима;
- Омогући једноставан кориснички интерфејс приликом примене одговарајућих метода и интерпретације резултата;
- Омогући једноставно проширење новим методама и компонентама.

5.2 Архитектура програмског решења

Платформа има трослојну архитектуру која обухвата: управљачки слој, слој корисничког интерфејса и слој података. Слика 5.1 представља архитектуру Платформе. Управљачки слој садржи одговарајуће компоненте које представљају спону између слоја корисничког интерфејса и слоја података. Управљачки слој заправо има улогу посредника између два преостала слоја. Управљачки слој извршава процесе свих метода које су доступне на Платформи. Тренутно су на Платформи актуелне методе извршавања предефинисаних упита, додавања кориснички селектоване базе података, динамичког филтрирања резултата упита и детекције сличних података. Слој података покрива *DataSources* онтологију, која садржи предефинисане упите и информације о базама података који се користе за њихово дефинисање. Слој корисничког интерфејса има задатак да омогући комуникацију корисника са управљачким слојем и да обезбеди приказ резултата.



Слика 5.1 Архитектура Платформе

5.2.1 Управљачки слој

Управљачки слој садржи компоненте[72] које имају задатак да спроводе процесе током извршавања одговарајућих метода на Платформи.

Компонента *PredefinedQuery* преузима корисничке захтеве за извршавање предефинисаних упита и реализацијом одговарајућих SELECT SPARQL упита „скенира“ *DataSources* онтологију у циљу преузимања неопходних података који се прослеђују *QueryPreparation* компоненти.

Компонента *PredefinedQueryExtension* преузима податке које корисник прослеђује приликом додавања кориснички селектованог скупа података у предефинисани упит. Ова компонента затим обрађује и процесира податке, а потом их припрема за ажурирање *DataSources* онтологије. Ажурирање се постиже извршавањем UPDATE SPARQL упита.

Компонента *DynamicQueryFilter* на основу захтева корисника и података преузетих из *DataSources* онтологије, омогућава динамичку пројекцију скупова података, приказ одговарајућих предиката, као и извођење *star-shaped*[73] *SPARQL* упита (одељак 6.2.1), који омогућавају побољшање релевантности резултата. Ово се постиже у комбинацији са *QueryPreparation* компонентом.

[72] Називи компоненти су делимично измењени у односу на оригинално истраживање [6].

[73] У наставку ће се користити термин звездасти.

Компонента *QueryPreparation* на основу података преузетих из *DataSources* онтологије припрема упите за њихово коначно извршавање. Резултат се, у комбинацији са *SPARQLQueryRunner* компонентом, приказује кориснику.

Компонента *SPARQLQueryRunner* извршава SELECT SPARQL упите на Платформи. Упити се изводе на CPCTAS *endpoint*[74]-у. Иначе, због велике количине података који се преузимају из онтолошких база података преко *remote endpoint*-а, критичан ресурс је брзина којом се спроводе одређене функционалности. Упити који су развијени за потребе Платформе користе SERVICE SILENT клаузулу за преузимање података са *remote endpoint*-а.

Улога компоненте *DetectingSimilarDataItems* јесте да на захтев корисника изврши детекцију сличних података добијених након извршавања предефинисаних упита и упита са кориснички селектованим базама података. За ове потребе развијен је посебан алгоритам који је детаљно размотрен у Седмом поглављу.

5.2.2 Слој података

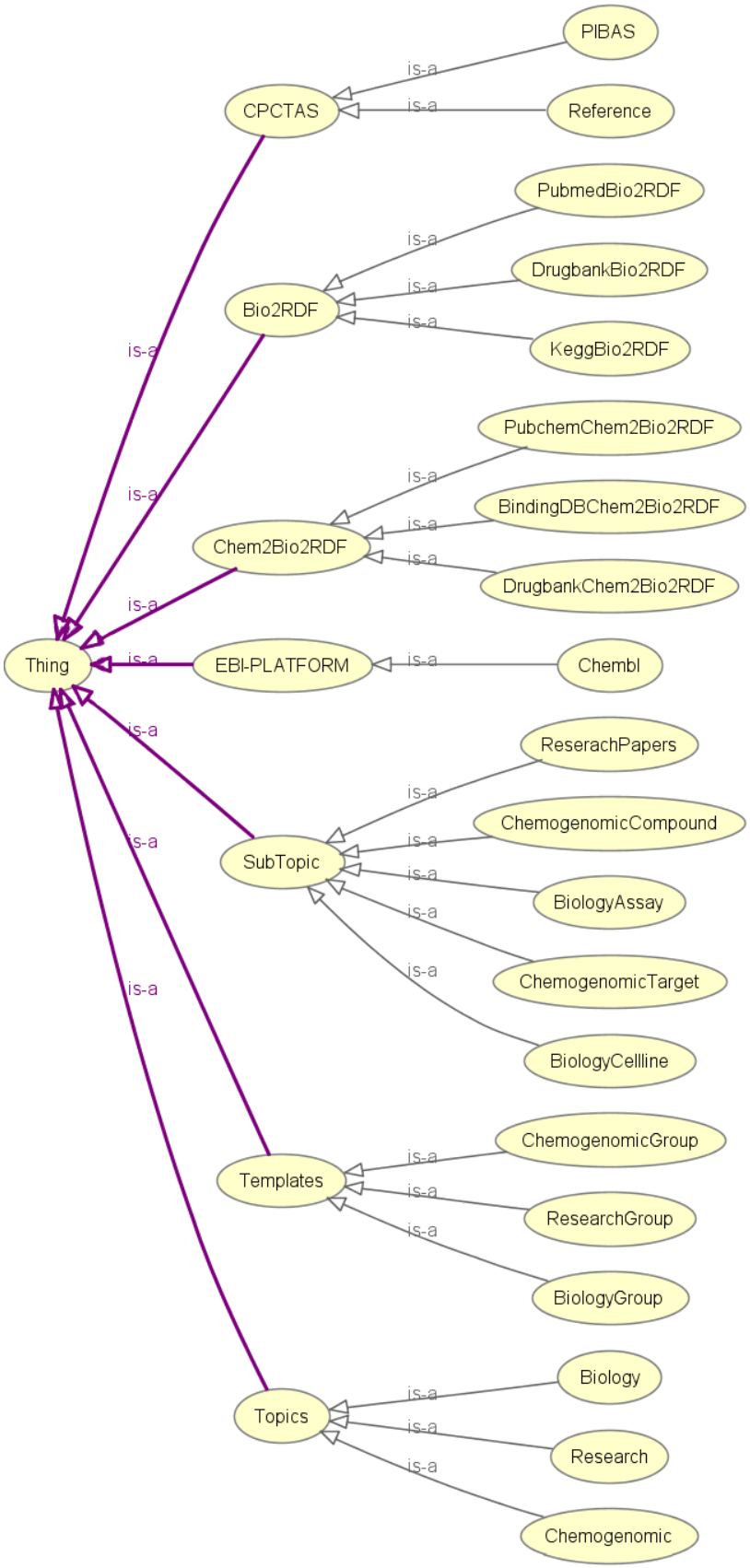
У слоју података налази се *DataSources*[75] онтологија, која садржи предефинисане упите, као и информације о скуповима података који су укључени у упите. Слика 5.2 приказује основну таксономију ове онтологије.

Све класе у онтологији су потклase од *owl:Thing* и припадају одговарајућем хијерархијском нивоу. Међу класама првог хијерархијског нивоа издвојићемо класе *pibas:Topics, pibas:SubTopic* и *pibas:Templates*. Ове класе се респективно користе за дефинисање тема, подтема и шаблона, који су кроз GUI доступне кориснику. Класа *pibas:Topics* има дефинисане потклase *pibas:Biology, pibas:Chemogenomics* и *pibas:Research.* Термини *Chemogenomic*[76]*, Biology* и *Reserach* одабрани су у складу са тематиком која се односи на шаблоне (*). Прва потклаcа служи за представљање шаблона (*b*) и (*c*), друга за представљање шаблона (*a*) и (*d*), а трећа за представљање шаблона (*e*). Свака од ових класа има своје инстанце које имају објектна и својства типа података. Конкретно, све потклase класе *pibas:Topics* имају објектна својства *pibas:hasSubTopic*, којим се захтева да инстанце ових класа имају вредност дефинисану инстанцом класе *pibas:SubTopic*. Својство *pibas:hasName* се користи да опише назив теме, која је кроз GUI евидентна кориснику. Класа *pibas:SubTopic* се користи за дефинисање подтема. Њене потклase су *pibas:ChemogenomicTarget*, *pibas:ChemogenomicCompound, pibas:BiologyAssay*, *pibas:BiologyCellLine* и *pibas:ReserachPapers.* Свака од ових класа, које припадају класама другог хијерархијског нивоа, има истоимене инстанце које су преко објектних својства *pibas:hasTemplate* повезане са инстанцама класа *pibas:Templates*. Својство *pibas:hasName* се користи да опише назив подтеме и ова вредност је доступна кроз GUI. Класа *pibas:Templates* се користи да опише шаблоне који се користе за извођење упита. Њене потклase су *pibas:BiologyGroup*, *pibas:ChemogenomicGroup* и *pibas:ReserachGroup*. Свака од ових класа тренутно садржи по једну или више инстанци које представљају шаблоне. Атрибут *pibas:hasName* означава назив шаблона, који је кроз GUI такође доступан кориснику. Табела 5.1 представља актуелне релације између тема, подтема, шаблона и кључних речи у *DataSources* онтологији.

74 *http://cpctas-lcmb.pmf.kg.ac.rs:3030/PIBAS/sparql*

75 *http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS/DataSources.owl*

76 *https://en.wikipedia.org/wiki/Chemogenomics*



Слика 5.2 Основна таксономија концепата *DataSources* онтологије[77]

[77] За визуелизацију је коришћен Protégé додатак OWLViz.

Табела 5.1 Релације између класа *pibas:Topics*, *pibas:SubTopic* и *pibas:Templates* у *DataSources* онтологији

Класе	**Поткласе**	**Назив поткласе (*pibas:subTopicName*)**
pibas:Topics	*pibas:Chemogenomic*	*Chemogenomic*
	pibas:Biology	*Biology*
	pibas:Research	*Research*
pibas:SubTopic	*pibas:ChemogenomicTarget*	*Targets*
	pibas:ChemogenomicCompund	*Compounds*
	pibas:BiologyAssay	*Assay*
	pibas:BiologyCellLine	*Cell line*
	pibas:ResearchPapers	*Papers*
pibas:Tepmplates	*pibas:BiologyGroup*	
	pibas:ChemogenomicGroup	
	pibas:ResearchGroup	
Класе	**Инстанце**	
pibas:BiologyGroup	*pibas:Template1*	
	pibas:Template3	
pibas:ChemogenomicGroup	*pibas:Template2*	
	pibas:Template5	
pibas:ResearchGroup	*pibas:Template4*	
Инстанце	**Назив шаблона (*pibas:TemplateName*)**	**Кључна реч (*pibas:hasInput*)**
pibas:Template1	*Find assays for the drug*	*SMILES*
pibas:Template3	*Find cell lines for the drug*	*InchiKey*
pibas:Template2	*Find targtes for the drug*	*InchiKey*
pibas:Template5	*Find info about drug*	*InchiKey*
pibas:Template4	*Find papers with a tittle for the keyword*	*Keyword*
Објектно својство	**Домен**	**Кодомен**
pibas:hasSubTopic	*pibas:Topics*	*pibas:SubTopic*
pibas:hasTemplate	*pibas:SubTopic*	*pibas:Tepmplates*
Својство типа података	**Домен**	**Кодомен**
pibas:hasName	*pibas:Topics, pibas:SubTopic, pibas:Tepmplates*	*xsd:string*

На првом хијерархијском нивоу су и класе које одговарају репозиторијумима представљеним у одељку 4.5.1. Потклассе ових класа садрже описе база података који се користе у предефинисаним упитима. Класа *pibas:Bio2RDF* садржи потклассе *pibas:DrugbankBio2RDF* и *pibas:PubmedBio2RDF*; класа *pibas:Chem2Bio2RDF* садржи потклассе *pibas:DrugBankChem2Bio2RDF*, *pibas:BindingDBChem2Bio2RDF* и *pibas:PubchemChem2Bio2RDF*; класа *pibas:EBI-PLATFORM* садржи потклассу *pibas:ChemblEMBLEBI*; а класа *pibas:CPCTAS* садржи потклассе *pibas:PIBAS* и *pibas:Reference*. Све наведене потклассе припадају другом хијерархијском нивоу и свака од њих садржи инстанцу која се односи на дату базу, а која је представљена предикатима *pibas:comment*, *pibas:hasName*, *pibas:endpoint*, *pibas:fromDataSource* и *pibas:link*. Табела 5.2 представља преглед инстанци свих класа, које се користе у предефинисаним упитима са приказом предиката.

Табела 5.2 Преглед инстанци свих класа за предефинисне упите у *DataSources* онтологији

Инстанца	**Предикати**
pibas:PIBAS/CPCTASInstance	***pibas:hasName*** *PIBAS* ***pibas:fromDataSource*** *CPCTAS;* ***pibas:endpoint*** *http://cpctas-lcmb.pmf.kg.ac.rs:3030/PIBAS/sparql* ***pibas:comment*** *The subject of the analysis that are carried out at the Laboratory include monitoring of in vitro effects of active substances in the cell lines of different origin (primarily cancer cell lines) and primary cells isolated from different tissues.Tests include cytotoxic active substances in human cancer cell lines, while monitoring the type of cell death, the mechanisms of apoptosis, migration and angiogenesis and prooxidant-antioxidant mechanisms which underlie the regulation of these processes.* ***pibas:link*** *http://cpctas-lcmb.pmf.kg.ac.rs/lcmb/index.php?jezik=english*
pibas:Reference/CPCTASInstance	***pibas:hasName*** *Reference* ***pibas:fromDataSource*** *CPCTAS;*

	pibas:endpoint *http://cpctas-lcmb.pmf.kg.ac.rs:3030/PIBAS/sparql* ***pibas:comment*** *Reference dataset contains information about publictions for researchers at the Faculty of Science, University of Kraguejvac, Serbia.* ***pibas:link*** *http://physics.kg.ac.rs/references/*
pibas:Chembl/EMBL-EBIInstance	***pibas:hasName*** *Chembl* ***pibas:fromDataSource*** *EMBL-EBI;* ***pibas:endpoint*** *https://www.ebi.ac.uk/rdf/services/sparql* ***pibas:comment*** *The ChEMBL database contains compound bioactivity data against drug targets. Bioactivity is reported in Ki, Kd, IC_{50}, and EC_{50}.* ***pibas:link*** *https://www.ebi.ac.uk/chembl/*
pibas:Drugbank/Bio2RDFInstance	***pibas:hasName*** *Drugbank* ***pibas:fromDataSource*** *Bio2RDF;* ***pibas:endpoint*** *http://drugbank.bio2rdf.org/sparql* ***pibas:comment*** *The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information....* ***pibas:link*** *http://www.drugbank.ca/*
pibas:Kegg/Bio2RDFInstance	***pibas:hasName*** *Kegg* ***pibas:fromDataSource*** *Bio2RDF;* ***pibas:endpoint*** *http://kegg.bio2rdf.org/sparql* ***pibas:comment*** *KEGG is an integrated database resource consisting of 16 main databases, broadly categorized into biological systems information, genomic information, and chemical information.* ***pibas:link*** *http://kegg.bio2rdf.org/sparql*
pibas:Pubmed/Bio2RDFInstance	***pibas:hasName*** *Pubmed* ***pibas:fromDataSource*** *Bio2RDF;* ***pibas:endpoint*** *http://lod.openlinksw.com/sparql* ***pibas:comment*** *PubMed comprises more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.* ***pibas:link*** *https://www.ncbi.nlm.nih.gov/pubmed/*
pibas:BindingDB/Chem2Bio2RDF Instance	***pibas:hasName*** *BindingDB* ***pibas:fromDataSource*** *Chem2Bio2RDF;* ***pibas:endpoint*** *http://cheminfov.informatics.indiana.edu:8080/bindingdb/sparql* ***pibas:comment*** *BindingDB is a publicly accessible database currently containing ~20 000 experimentally determined binding affinities of protein–ligand complexes, for 110 protein targets including isoforms and mutational variants, and ~11 000 small molecule ligands.* ***pibas:link*** *http://www.bindingdb.org/bind/index.jsp*
pibas:Pubchem/Chem2Bio2RDFInstance	***pibas:hasName*** *Pubchem* ***pibas:fromDataSource*** *Chem2Bio2RDF;* ***pibas:endpoint*** *http://cheminfov.informatics.indiana.edu:8080/pubchem/sparql* ***pibas:comment*** *The PubChem BioAssay Database contains bioactivity screens of chemical substances described in PubChem Substance. It provides searchable descriptions of each bioassay, including descriptions of the conditions and readouts specific to that screening procedure.* ***pibas:link*** *https://www.ncbi.nlm.nih.gov/pcassay*

Инстанце које представља Табела 5.2 су у релацијама са инстанцама класе *pibas:Templates* преко објектног својства *pibas:connectedWith*. Табела 5.3 даје преглед инстанци класе *pibas:Templates* са приказом најрелевантнијих предиката. Својство *pibas:hasInitalQuery* класе *pibas:Templates* дефинише предефинисани упит који се извршава на Платформи. За сваки шаблон везује се и атрибут *pibas:hasInput* који одређује тип кључне речи коју корисник уноси на Платформи. Вредност својства типа података *pibas:hasSimilar* дефинише да ли је за дати шаблон омогућена метода детекција сличних података.

Табела 5.3 Преглед инстанци класе *pibas:Templates* са приказом релевантних предиката у *DataSources* онтологији

Инстанца (*pibas:hasName*)	***pibas:topicName***	***pibas:connectedWith***	***pibas:hasInitialQuery***	***pibas:hasSimilar***
Find assays for the drug	*Assay*	*pibas:PIBAS/CPCTASInstance*	Слика 6.6 б)	Yes

		pibas:PubChem/Chem2Bio2RDFInstance *pibas:Chembl/EMBL-EBIInstance*		
Find cell lines for the drug	*Cell line*	*pibas:PIBAS/CPCTASInstance* *pibas:Chembl/EMBL-EBIInstance*	Слика 6.6 ц)	Yes
Find targtes for the drug	*Target*	*pibas:PIBAS/CPCTASInstance* *pibas:BindingDB/Chem2Bio2RDFInstance* *pibas:Chembl/EMBL-EBIInstance* *pibas:Drugbank/Bio2RDFInstance* *pibas:Kegg/Bio2RDFInstance*	Слика 6.6 а)	Yes
Find info about drug	*Compound*	*pibas:*PIBAS/CPCTASInstance *pibas:BindingDB/Chem2Bio2RDFInstance* *pibas:Drugbank/Bio2RDFInstance* *pibas:Kegg/Bio2RDFInstance* *pibas:Chembl/EMBL-EBIInstance*	Слика 6.6 д)	No
Find papers with a tittle for the keyword	*Paper*	*pibas:Reference/CPCTASInstance* *pibas:Pubmed/Bio2RDFInstance* *pibas:Chembl/EMBL-EBIInstance*	Слика 6.6 е)	No

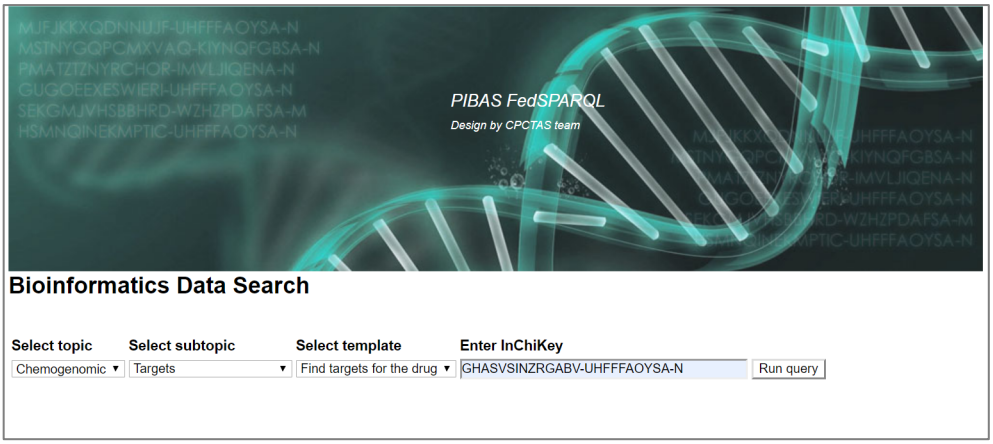
Онтологија *DataSources* је креирана тако да се може лако проширити новим класама, инстанцама и предикатима и то без нарушавања њене таксономије. Сви ентитети се такође могу елегантно модификовати. На пример, инстанца *pibas:Template2* која захтева *InChiKey* вредност као кључну реч, може се једноставно трансформисати у шаблон чија је кључна реч *SMILES* параметар. Такође, шаблони се могу проширити новим суповима података или се неки могу изопштити из њих. Ово промене тренутно мануелно реализује администратор Платформе.

5.2.3 Слој корисничког интерфејса

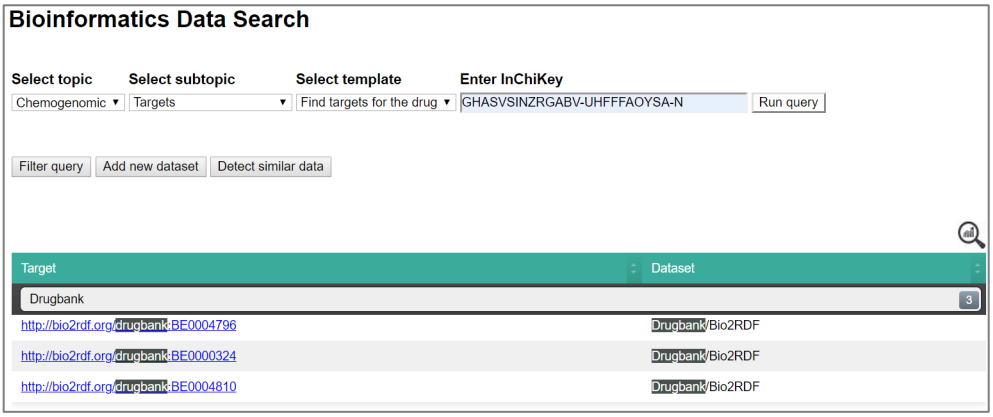
Слој корисничког интерфејса има за циљ да корисницима омогући:

- Интуитивну навигацију кроз елементе;
- Приказ резултата у форми табела;
- Сортирање и претрагу табеларних приказа;
- Квантитативни приказ резултата упита;
- Унос кориснички селектованих база података;
- Динамичко филтрирање резултата упита;
- Детекцију сличних података.

Иницијални кориснички интерфејс омогућава интуитивну селекцију одговарајућих тема, подтема, шаблона и уноса кључне речи (Слика 5.3). Ове процесе спроводи *PredefinedQuery* компонента, која управља одговарајућим деловима програмског кода. Рад дате компоненте се заправо своди на извршавање одговарајућих SPARQL упита, који на основу селектованих критеријума претраге преузимају податке који су смештени у *DataSources* онтологији. Када су селектовани (односно попуњени) сви елементи на иницијалном корисничком интерфејсу, омогућено је извршавање предефинисаних упита кликом на дугме *Run query*.

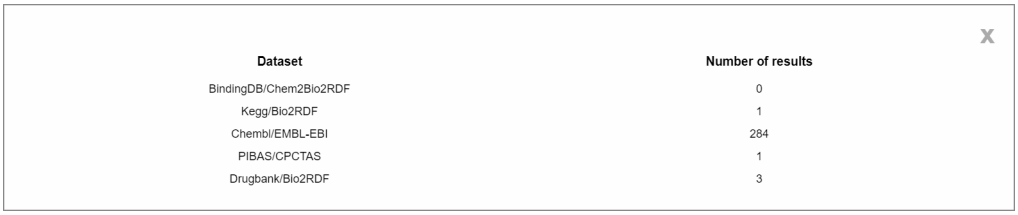


Слика 5.3 Иницијални кориснички интерфејс за извршавање предефинисаних упита на Платформи



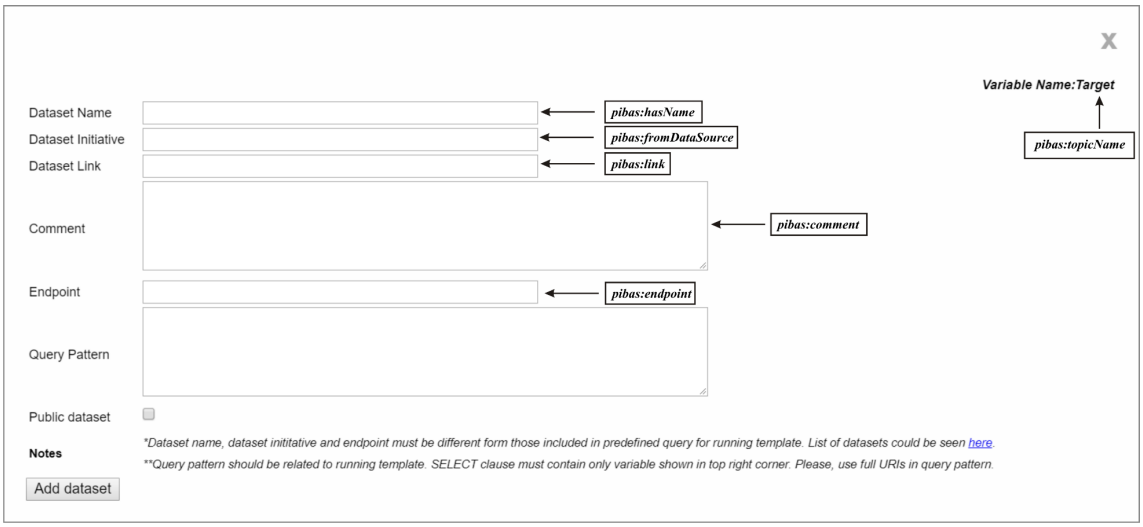
Слика 5.4 Табеларни приказ резултата извршавања предефинисаног упита на Платформи

Реализацијом предефинисаних упита преузимају се резултати (инстанце) одговарајућих база. Ови резултати могу бити многобројни (више стотина или хиљада инстанци) и због тога је неопходно представити их на адекватан начин. Прикладан приказ података окосница је за даље процесирање и анализу информација, која у домену биоинформатике није тако једноличан процес и често је у тесној корелацији са репрезентацијом самих података. Због тога је табеларни приказ одабран као најпогоднији начин за репрезентацију података на Платформи. Табеле имају две колоне: прва колона приказује резултат упита у форми URI спецификације (евентуално стринга), док друга колона служи за приказ назива базе података и репозиторијума из које резултат произилази (Слика 5.4). Табеле се могу сортирати и претраживати по тексту. Након извршавања предефинисаних упита постоји могућност и квантитативног приказа резултата упита у *pop-up* форми (Слика 5.5), која је имплементирана користећи *jQuery* и *HTML*. Овај приказ је доступан кликом на иконицу и уз помоћ њега могуће је проверити базе података које су укључене у предефинисани упит, без обзира на то да ли имају повратних вредности или не. Компоненте управљачког слоја *PredefinedQuery*, *QueryPreparation* и *SPARQLQueryRunner* задужене су за спровођење ових функционалности.



Слика 5.5 Квантитативни приказ резултата извршавања предефинисаног упита на Платформи

Након реализације методе извршавања предефинисаних упита доступне су и методе додавања кориснички селектоване базе података, методе динамичког филтрирања резултата упита и метода селекције сличних података. Свака од ових метода се одликује специфичним корисничким интерфејсом. За методу додавања кориснички селектоване базе података, имплементирана је *pop-up* форма (Слика 5.6) која је омогућена кликом на дугме *Add new dataset*.

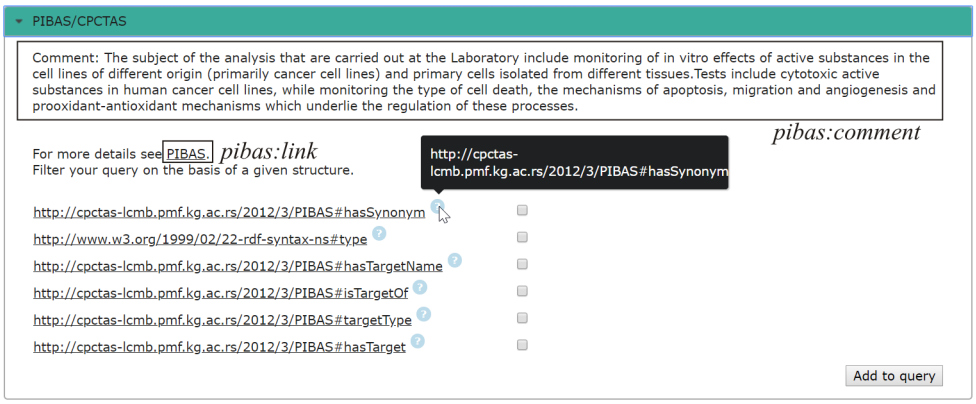


Слика 5.6 Форма за додавање кориснички селектоване базе података на Платформи

Форма је имплементирана користећи *jQuery* и *HTML.* Типови података који се у овом случају попуњавају на форми објашњени су у одељку 6.3. Променљива која се појављује у горњем десном углу форме одговара објектној вредности предиката *pibas:topicName* који се везује за селектовану инстанцу класе *pibas:Templates*. Спровођење *PHP* функција за извршавање ових функционалности поверено је *PredefinedQueryExtension* компоненти која је у спрези са *QueryPreparation* и *SPARQLQueryRunner* компонентама.

Након извршавања предефинисаних упита доступна је и метода динамичког филтрирања резултата упита. Кликом на дугме *Filter query* корисницима се приказују елементи (панели) који се на основу резултата одговарајућих упита динамички попуњавају RDF предикатима, који одговарају актуелној структури селектоване базе података (Слика 5.7). За имплементацију овог елемента коришћен је *jQuery accordion widget*[78], који обезбеђује панеле за приказ информација у ограниченом садржају. Садржај панела се одређује динамички и ово је детаљније објашњено у одељку 6.2. Спровођење *PHP* функција за извршавање датих функционалности поверено је *DynamicQueryFilter* компоненти која је у спрези са *QueryPreparation* и *SPARQLQueryRunner* компонентама.

[78] *https://jqueryui.com/accordion/*



Слика 5.7 Пример панела PIBAS/CPCTAS базе података за селектовање предиката за примену методе динамичког филтрирања резултата упита

Метода детекције сличних података омогућена је након извршавања предефинисаних упита (кликом на дугме *Detect similar data*). Њена примена је ограничена увођењем предиката *pibas:hasSimilar*, који подржавају инстанце класе *pibas:Templates*. Ограниченост је уведена с обзиром на то да над неким резултатима (инстанцама одговарајућих база података) није од важности применити процес утврђивања сличности. Конкретно, над резултатима шаблона (*d*) ова опција није од превеликог значаја, јер је резултат извршавања шаблона *Find info about drug* скуп идентичних података, који припадају различитим базама. Слично важи и за шаблон (*e*) јер нема превеликог смисла поредити инстанце које се односе на публикације. Спровођење програмског кода за извршавање ове функционалности поверено је *DetectingSimilarDataItems* компоненти. Алгоритам ове методе је детаљније разматран у Седмом поглављу.

5.3 Ток рада

Платформа омогућава корисницима да на интуитиван начин - селекцијом теме, подтеме, шаблона и уноса кључне речи - извршавају предефинисане Federated SPARQL упите. У предефинисане упите могу се додати и кориснички селектоване базе података. На основу добијених резултата корисници имају могућност да одаберу скупове података у складу са њиховим интересовањима и да кроз интуитивни преглед RDF структура тих скупова изаберу одговарајућа својства и спроведу додатне упите у циљу побољшања релевантности резултата. Платформа омогућава детекцију и сличних података, који су добијени извршавањем предефинисаних упита, користећи посебно имплементиран алгоритам. Значај ових метода је презентован у наредним поглављима.

5.4 Детаљи имплементације програмског решења

Код развијеног софтвера написан је у програмским језицима *PHP* (верзија 5.3.28) и *Python* (верзија 2.7.12[79]). Као програмско окружење коришћен је *NetBeans*, верзија 8.2. Комплетан код доступан је на GitHub[80]-у. Укупна количина изворног кода износи око 11 MB. Током имплементације коришћено је и неколико спољних библиотека, сервера и алата, и то:

- За имплементацију *DataSources*, *PIBAS* онтологије и осталих онтологија у CPCTAS бази података коришћен је Protégé едитор, верзија 4.0.2. Додатно, коришћени су Protégé алати и додаци: OWLDiff[81] за утврђивање компарације између онтологија, OWLViz[82] за преглед таксономије и OntoGraph[83] за приказ релација између концепата у онтологији;
- За креирање (Federated, SELECT и UPDATE) SPARQL упита коришћена је SPARQL синтакса

[79] Прелазак на *Python 3* је могућ уз ажурирање синтаксе постојећег кода и верзија коришћених пакета и модула.

[80] *https://github.com/marijadjokic/PIBASFedSPARQL*

[81] *https://protegewiki.stanford.edu/wiki/OWLDiff*

[82] *https://protegewiki.stanford.edu/wiki/OWLViz*

[83] *http://semanticweb.org/wiki/Ontograph.html*

верзија 1.1;

- За складиштење CPCTAS базе података коришћен је Fuseki SPARQL сервер, верзија 2.4.1, *endpoint*: *http://cpctas-lcmb.pmf.kg.ac.rs:3030/PIBAS/sparql*;
- За складиштење тестних података коришћен је Fuseki SPARQL сервер, верзија 2.4.1, *endpoint*: *http://cpctas-lcmb.pmf.kg.ac.rs:3030/mytestdataset/sparql*;
- За извршавање SPARQL упита коришћена је PHP SPARQL Lib библиотека, LGPL License, *https://github.com/cgutteridge/PHP-SPARQL-Lib*;
- За имплементацију корисничког интерфејса коришћен су *HTML*, *CSS*, *javaScript* и *jQuery*;
- За имплементацију алгоритма за детекцију сличних података коришћен је програмски језик *Python*. Вештине развоја софтвера и искуства која су стечена у раду [92], а који такође представља допринос овог истраживања и дисертације, утицала су на избор *Python* програмског језика за имплементацију алгоритма.

6 Основне методе Платформе

У овом поглављу су представљене основне методе Платформе: извршавање предефинисаних Federated SPARQL упита, извршавање предефинисаних Federated SPARQL упита са кориснички селектованим базама података као и динамичко филтрирање резултата упита. Генерално, описана је функционалност датих метода и интерпретација резултата. За сваку методу су наведене теоријске основе, технике или помоћни алати који су коришћени за њену имплементацију. У складу са тим користе се описи неких делова корисничког интерфејса као и делови програмског кода. Посебна пажња је усмерена на опис методологије (процес селекције извора података) која је у овом истраживању примењена за креирање предефинисаних упита. Акценат је усмерен на потенцијалне проблеме, који могу отежати креирање упита.

6.1 Креирање и извршавање предефинисаних упита

Све већа иницијатива корисника за објављивањем података и стварањем јавних SPARQL *endpoint*-а чине LOD добром платформом за истраживање и откривање знања [93]. Међутим, кључни изазов је управо комбиновање података из различитих база како би се открили релевантне информације. Ово је нарочито евидентно у процесу рационалног дизајна лекова, где је неопходно повезати хемијске информације лека (хемијске базе података) са биолошким подацима (биолошким базама података), како би истраживачи могли да утврде евентуални утицај лека на биолошке системе. Са информатичког аспекта то заправо подразумева да је за извођење SPARQL упита неопходан приступ подацима различитих база, односно неопходно је извршити креирање Federated SPARQL упита. Да би се овакви упити креирали потребно је извршити селекцију извора података (онтолошких база података), утврдити све релације између ентитета и дефинисати одговарајуће обрасце (подупите), који су саставни део Federated SPARQL упита.

Да би се превазишао проблем упита над вишеструким изворима података, познавање SPARQL синтаксе често је есенцијално, али обично не и довољно за успешно креирање коначних упита. Одабир погрешних извора података, осим што може утицати на повећање интернет саобраћаја и време извршавања упита, може довести и до нерелевантних резултата, а то је круцијалан проблем са којим се истраживачи највише суочавају [94]. Додатно, кључан проблем није само број расположивих база података, већ и варијабилност њихових речника. У суштини, како се мења сама наука, мења се и модел података, што се аутоматски одражава на синтаксу SPARQL упита. На пример, термин *Target* је уопштено дефинисан као одређени протеин, али тренутно се развијају и лекови који делују на колекције протеина, што се аутоматски одражава на RDF репрезентацију овог термина у одговарајућим базама података. Даље, у многим природним језицима постоје хомоними и синоними. Као пример хомонима може се навести термин *Paper,* који може означавати *папир* или *документ*, док термини *Paper* и *Article* могу представљати синониме [95]. Дакле, како базе података могу да варирају у својим RDF репрезентацијама, неопходно је константно одржавање упита у циљу њихових валидности. Такође, корисници се могу суочити и са проблемом константног напуштања извора података из јавног домена, у смислу да одређене институције или истраживачки центри могу заштити властите базе података или их трајно склонити из јавног домена. Осим тога, сваки домен је одговоран за сопствено представљање и управљање знањем, јер не постоји *a-priori* споразум у вези онтолошког језика нити грануларности [95].

Истраживањем литературе дошло се до сазнања да не постоје аутоматски генератори Federated SPARQL упита који раде униформно, јер је готово немогуће аутоматски претражити све релевантне ентитете у базама података (које имају више милиона триплета), издвојити релевантне релације између ентитета и аутоматски формирати упите који задовољавају корисничке потребе. Према сазнањима, постоје само полу-аутоматски генератори Federated SPARQL упита [96], који на основу селектованих извора података генеришу финалне упите коришћењем унапред дефинисаних образаца селектованих извора података. Често се користе обрнути процеси, који полазе од унапред генерисаних SPARQL упита и који покушавају да утврде да ли делови упита одговарају одређеним изворима података (*endpoint*-има), а затим се на основу те провере врши прегруписавање и креирање коначних Federated SPARQL упита [97]. Због тога се на Платформи користе предефинисани упити, јер је акценат усмерен на указивању

комплексних односа између ентитета у онтолошким базама података, а не на аутоматском генерисању упита. У наставку је описан процес креирања предефинисаних упита на Платформи, који су развијени са циљем да задовоље шаблоне (*) наведене у одељку 4.5.3.

6.1.1 Методологије за креирање упита

Креирање SPARQL упита изузетно је комплексан процес, имајући у виду велики број триплета које једна онтолошка база података може поседовати. Подсећања ради, DrugBank/Bio2RDF база података има 3.672.531 триплета и 316.950 ентитета[84]. Како познавање SPARQL синтаксе често није довољно да би се процес креирања упита успешно спровео, неопходно је користити додатне технике и приступе, како би се овај процес унапредио. Процес креирања SPARQL упита често се може поистоветити са процесом откривања извора података (енгл. *source discovery*) [98]. Процес откривања извора података је процес лоцирања извора података, откривање његових типова података и утврђивање способности [98] у смислу провере да ли одређени обрасци могу бити подржани од стране датог извора. Генерално, да би се овај процес спровео често је неопходно проверити таксономију базе података, искористити методе онтолошког поравнања [99] и спровести селекцију извора података за дефинисане упите (енгл. *source selection*) [98]. Ови приступи су неретко међусобно зависни и обично је један приступ предуслов да се испуни други.

Процес утврђивања таксономије базе података обично се своди на утврђивање постојећих ентитета (концепата) онтологије и релација које постоје између њих. У овом случају се утврђују односи између класа (*rdfs:subClassOf*), класа и својстава, класа и инстанци. На овај начин се стварају предуслови за онтолошко поравнање. „*Поравнање онтологија подразумева процес упоређивања онтологија који се односи на проналажење семантичких односа или подударања између ентитета различитих онтологија*“ [100]. То је поступак у којем се за сваки елемент из једне онтологије проналази одговарајући идентичан или сличан елемент у другој онтологији, уколико такав постоји [100,99]. Онтологије морају у одређеној мери показивати и сличност, односно морају постојати и нека преклапања међу њима, како би се успешно могли успоставити односи подударности међу појединим ентитетима (класама, својствима и инстанцама) [100]. Због тога је кључни корак у поступку поравнавања управо израчунавање сличности између ентитета, што имплицира да је неопходно дефинисати меру сличности [100]. Ово је искоришћено као базна идеја за процес утврђивања сличних података, који је описан у Седмом поглављу, где су детаљно образложене неке од основних техника поравнања онтологија и њихов утицај на методу детекције сличних података која је представљена у овом истраживању. На основу ентитета који се пореде, поравнање се генерално дели на поравнање на нивоу концепта или класе (енгл. *class alignment, or class matching*), поравнање на нивоу својства (енгл. *property alignment, or property matching*) и поравнање на нивоу инстанце (енгл. *instance alignment, or instance matching*) [93]. Поравнање на нивоу класе обично подразумева претпоставку да су одређене класе сличне уколико су њихове надређене или подређене класе сличне [100]. Поравнање на нивоу својства представља важан и сложен изазов, јер својства обухватају комплексне структуре и значење инстанци на нивоу података, док класе имају апстрактније значење [93]. Поравнање на нивоу инстанце је јако важно јер се утврђивањем истих ентитета који припадају различитим базама података повећава интероперабилност података и ови типови поравнања се обично базирају на примени сличности између стрингова [93]. Линкови који се у овом случају узимају у обзир углавном су типа *owl:sameAs*, *rdfs:seeAlso*, *skos:closeMatch*, *skos:relatedMatch* итд. [93]. Основни циљ поравнавања онтологија јесте надвладати њихову семантичку хетерогеност и на тај начин редуковати обрасце како би се евентуално избегли дуплирани резултати [100]. Утврђивање онтолошког поравнања предуслов је за селекцију извора података.

Селекција извора података најчешће је заснована на традиционалној „захтев/одговор“ парадигми, где корисник има улогу клијента, а база података улогу сервера [95]. Међутим, ово је остваривo ако су домени база података познати самим корисницима. Када то није случај, откривање извора података може бити прилично комплексно и прави изазов у домену LOD-а [101]. На основу истраживања литературе, закључено је да постоје различите методе за откривање и класификацију извора података. Ladwig и др.

[84] *http://download.bio2rdf.org/files/release/3/drugbank/drugbank.html*

[102] уводе три категорије за откривање релевантних извора података: одозго на доле (енгл. *top-down*), одоздо на горе (енгл. *bottom-up*) и комбиновану стратегију (енгл. *mixed strategy*). Прва стратегија се ослања на различите врсте механизама индексирања како би се пронашли релевантни извори података. SPLENDID [103] сакупља статистичке податке из VoID [104] описа и креира локални индекс, који мапира предикате и типове на скупове података и друге статистичке информације. Приликом извршавања упита SPLENDID додељује одређени скуп података датом триплету на основу мапирања ограничених предиката (енгл. *bounded predicates*) и типова информација у упиту са локалним индексом. FedX [105] такође подржава дату стратегију за селекцију извора података. Он извршава SPARQL ASK упите за сваки триплет у обрасцу на *endpoint*-у, за сваки претходно одабран скуп података. Резултат ASK упита се одражава на све друге сличне триплете у упиту. Међутим, проблем са оваквим приступом може утицати на прецењеност неког скупа података уколико се у оквиру ASK упита користи генерички триплет, као што је *?s rdf:type ?o*, који је подржан у скоро свим онтолошким базама података. Друга стратегија се фокусира на проналажење релевантних извора података у ходу (енгл. *on-the-fly*). Hartig и др. [101] проналазе одговарајуће скупове података користећи технике преносних линкова. Овом методом се извршавају делови SPARQL-а упита тако што се најпре проналазе одговарајуће URI спецификације у упиту, а затим се даље користе друге URI спецификације, које се добијају из парцијалних резултата. Дакле, како би се иницирало извршавање упита, морају постојати URI спецификације, а истовремено је могуће да тај приступ не успе да преузме комплетни резултат на крају. Овакво решење може довести до бесконачног откривања веза, где систем не може да испуни услове престанака и наставља тражење веза. Систем Feedback [106] је предложио још један приступ за селекцију извора података, који почиње са иницијалним URI спецификацијама у упиту. На основу њих врши се прикупљање скупова података као базних ресурса, а затим се даље траже нове URI спецификације користећи предикате као што су *rdfs:seeAlso*, *owl:sameAs* и *owl:equivalentClass*. Након идентификације свих скупова података, врши се њихово рангирање анализом повратних информација корисника. Nikolov и др. [107] предлажу приступ који се базира на сличности између термина који припадају различитим скуповима података. Наиме, они врше екстракцију инстанци из скупова података и пореде њихове ознаке (*rdfs:label*) и на основу те сличности врши се рангирање скупова података. Комбинована стратегија настоји да комбинацијом претходне две поменуте стратегије постигне најбоље резултате у циљу селекције релевантних скупова података и одржи баланс између новијих (свежих) резултата и брзог извршавања упита [93]. Ladwig и др. [102] користе локалне индексе заједно са триплетима из обрасца за идентификацију извора података као иницијалну листу могућих релевантних извора и даље откривају изворе засноване на садржају, који се обрађује из почетног релевантног извора и средњих резултата. Процес проналажења релевантних скупова података завршава се на основу унапред конфигурисаних вредности као што су број излазних резултата и број изворних скупова података. Затим се врши рангирање коришћењем одређених метрика које користе низ функција као што је кардиналност (број триплета у скупу података се подудара са триплетом из обрасца) и број долазних веза са релевантним ресурсима.

Стратегија „одозго на доле“ се ослања на претходно познавање скупова података и због тога се упити могу оптимизовати за брже извршавање приликом идентификације релевантних извора података [93]. Међутим, ова стратегија можда не препозна ажуриране резултате, јер се идентификовани скупови података прикупљају у време индексирања, а резултати могу бити другачији у времену извршавања упита [93]. Насупрот томе стратегија „одоздо на горе“ проналази релевантне скупове података у ходу, што омогућава идентификацију ажурираних резултата, иако то може довести до проблема попут бесконачног откривања веза између ентитета, што може имати утицај на брзину одговора [93]. Заједничко свим наведеним стратегијама је да оне захтевају претходно делимично познавање структуре базе података и према сазнањима не постоје апликације које могу аутоматски да детектују релевантне изворе података, нити да на лак начин прилагоде постојеће обрасце одговарајућим захтевима.

6.1.2 Процес креирања предефинисаних упита

Преглед и анализа актуелне литературе у процесу селекције извора података, могу довести до закључка да још увек има простора за унапређење постојећих техника, као и за дефинисање нових приступа који би допринели олакшаном креирању SPARQL упита. Због тога, SPARQL стручњаци могу да бирају

између постојећих или да примене сопствене методе. За потребе Платформе било је неопходно креирати Federated SPARQL упите који би омогућили преузимање релевантних података са различитих извора (CPCTAS базе и база података представљених у одељку 4.5.1), а који би задовољили критеријуме за шаблоне (*) представљене у одељку 4.5.3. На основу анализе литературних података предложена је методологија за креирање упита која се састоји из следећих корака:

1) Селекција иницијалних упита;
2) Утврђивање таксономије концепата;
3) Онтолошко поравнање;
4) Селекција извора података;
5) Модификација иницијалних упита - креирање коначних Federated SPARQL упита;
6) Процена и вредновање упита.

У првом кораку се врши мануелна селекција иницијалних SPARQL упита истраживањем литературе и репозиторијума представљених у одељку 4.5.1. У другом кораку спроведено је полу-аутоматско утврђивање таксономије концепата база података укључених у иницијалне упите. У трећем и четвртом кораку коришћени су мануелни принципи, у смислу креирања и извршавања одговарајућих SELECT SPARQL упита за селекцију извора података и примену онтолошког поравнања. Након овог корака, а на основу усвојеног приступа који нуди систем Feedback [106], откривени су нови обрасци који одговарају захтевима шаблона (*). Сврха претходних корака је модификација иницијалних упита и креирање коначних Federated SPARQL упита. Последњи корак методологије укључује примену људских капацитета у смислу провере квалитета резултата упита. У наставку су наведени кораци детаљније образложени.

6.1.2.1 Селекција иницијалних упита

Прва фаза предложене методологије подразумевала је преглед литературе која се односи на репозиторијуме EMBL-EBI [84], Chem2Bio2RDF [39] и Bio2RDF [79], као и проверу постојећих SPARQL упита доступних у оквиру ових репозиторијума. У овом кораку методологије циљ је издвајање иницијалних упита, који би се касније модификовали и искористили за дефинисање коначних Federated SPARQL упита који се чувају у *DataSources* онтологији. Након исцрпног прегледа литературе дошло се до следећих сазнања:

a) За проналажење биолошких мета, есеја и ћелијских линија на EMBL-EBI репозиторијуму откривени су упити доступни на *https://www.ebi.ac.uk/rdf/services/sparql* (Слика 6.1);
b) За проналажење биолошких мета у Chem2Bio2RDF репозиторијуму откривен је упит на веб локацији *http://chem2bio2rdf.wikispaces.com/multiple+sources* (Слика 6.2).
c) Слика 6.3 представља упите за проналажење биолошких мета и публикација у Bio2RDF репозиторијуму. Упити су доступни на веб локацији *https://github.com/bio2rdf/bio2rdf-scripts/wiki/Query-repository*.

```
PREFIX cco:<http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX chembl_molecule:<http://rdf.ebi.ac.uk/resource/chembl/molecule/>
SELECT ?activity ?assay ?target ?targetcmpt ?uniprot
WHERE {
   ?activity a cco:Activity;
              cco:hasMolecule chembl_molecule:chemblID;
              cco:hasAssay ?assay.
   ?assay cco:hasTarget ?target.
   ?target cco:hasTargetComponent ?targetcmpt.
   ?targetcmpt cco:targetCmptXref ?uniprot.
   ?uniprot a cco:UniprotRef.
}
```

а)

```
PREFIX cco:<http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX chembl_molecule:<http://rdf.ebi.ac.uk/resource/chembl/molecule/>
SELECT ?assay ?cellline ?IC50value
WHERE {
   ?activity cco:hasMolecule chembl_molecule:chemblID;
             cco:standardUnits "nM";
             cco:standardType "IC50";
             cco:standardValue ?IC50value.
   ?assay cco:hasActivity ?activity.
   ?assay cco:hasCellLine ?cellline.
}
```

б)

Слика 6.1 SPARQL упити за претрагу биолошких мета (а), есеја и ћелијских линија (б) у Chembl/EMBL-EBI бази података. Променљива *chemblID* представља ID лека (активне супстанце)

```
PREFIX compound:<http://chem2bio2rdf.org/pubchem/resource/>
PREFIX drugbank:<http://chem2bio2rdf.org/drugbank/resource/>
PREFIX uniprot:<http://chem2bio2rdf.org/uniprot/resource/>

SELECT ?compound_cid ?target ?geneSymbol
FROM <http://chem2bio2rdf.org/pubchem>
FROM <http://chem2bio2rdf.org/drugbank>
FROM <http://chem2bio2rdf.org/uniprot>
WHERE
{
    ?compound compound:CID ?cid.
    FILTER (?compound_cid= 123631).
    ?chemical bindingdb:cid ?compound.
    ?target bindingdb:Monomerid ?chemical.
}
```

Слика 6.2 SPARQL упит за претрагу биолошких мета у BindingDB/Chem2Bio2RDF бази података. Променљива *?cid* представља ID лека (активне супстанце)

```
PREFIX drugbank:<http://bio2rdf.org/drugbank_vocabulary>
SELECT DISTINCT ?target
WHERE {
    ?drug drugbank:target ?target.
    FILTER(?drug = <?compound_uri>).
}
```

а)

```
PREFIX v:<http://bio2rdf.org/pubmed_vocabulary:>
SELECT ?article ?article_title
 {
   ?article ?p ?author;
            rdfs:label ?article_title;
            a v:PubMedRecord.
   ?author a v:Author;
           v:last_name ?ln;
           v:initials ?in.
   ?ln bif:contains ?author_name;
       bif:contains "m".
 }
```

б)

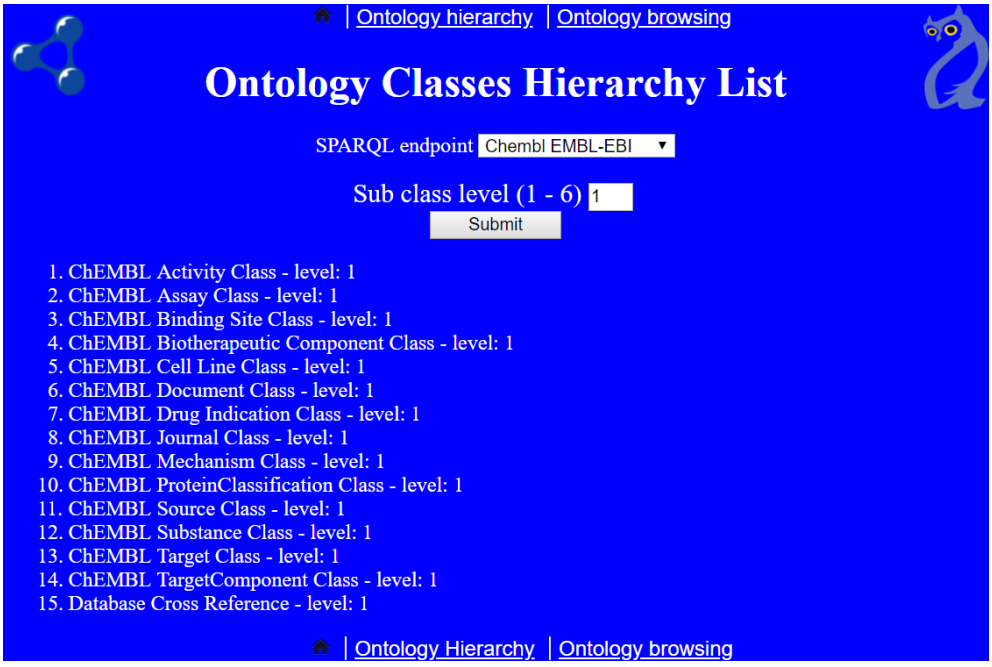
Слика 6.3 SPARQL упити за претрагу биолошких мета у Drugbank/Bio2RDF бази података (а) и публикација у PubMed/Bio2RDF бази података (б). Променљива *?compound_uri* представља URI спецификацију лека (активне супстанце). Променљива *?author_name* представља име аутора

Након селекције иницијалних упита приступа се следећем кораку методологије - утврђивању таксономије база података над којима се ови упити извршавају.

6.1.2.2 Утврђивање таксономије

Други корак методологије подразумевао је проверу структура онтолошких база података, над којима су се извршавали иницијални упити представљени у претходном кораку. Циљ ове фазе је утврдити релације између одговарајућих класа, својстава и инстанци, јер су ови односи од важности за наредне кораке предложене методологије.

Метода провере таксономије базе података често подразумева и примену одговарајуће апликације у циљу аутоматизације или полу-аутоматизације процеса, јер се на тај начин повећава могућност прикупљања знања [98]. Елементарни разлог због кога се користе помоћне апликације, или се дефинишу и примењују SPARQL упити специјалних намена, јесте тај што дијаграми шема база података често не карактеришу актуелну таксономију базе, па су релације између ентитета у бази другачије од оних које представљају шеме. Међутим, проблем примене помоћних апликација базира се на томе што је пре свега неопходно открити функционалности апликације, а затим је интегрисати у постојећи процес. Тежећи да други корак предложене методологије учинимо што продуктивнијим, одлучили смо се за полу-аутоматски приступ. За реализацију овог корака коришћено је софтверско решење представљено у [108], које такође представља један од резултата овог истраживања. Поменуто софтверско решење нуди две кључне функционалности: приказ хијерархијске структуре базе података са корисничким дефинисаним нивоом дубинских поткласа као целокупан увид у онтолошку структуру и претрагу онтолошке структуре бирањем произвољног пута кроз онтологију. Ово скалабилно софтверско решење је првобитно развијено за проверу таксономије DBpedia[85]-е (онтолошка база Wikipedia-е), али је локално модификовано и искоришћено за утврђивање таксономије онтолошких база података иницијалних упита. Примена помоћног софтвера иницијално је имала за циљ детекцију класа које се односе на биолошке мете, есеје, ћелијске линије, лекове и публикације који су од важности за шаблоне (*). Слика 6.4 представља приказ класа првог хијерархијског нивоа Chembl/EMBL-EBI базе података, који је постигнут применом помоћног софтверског решења.



Слика 6.4 Приказ класа првог хијерархијског нивоа Chembl/EMBL-EBI базе података добијен применом софтверског решења представљеног у *[108]*

[85] *https://wiki.dbpedia.org/*

У првом хијерархијском нивоу су уочене класе *chembl:Traget*, *chembl:CellLine*, *chembl:Assay*, *chembl:Journal*, као и класа *chembl:SmallMolecule,* која се односи на лекове. На сличан начин у BindingDB/Chem2Bio2RDF бази података откривена је класа *bindingdb:bindingdb_interaction*, која означава интеракцију протеина, а која је по значењу слична класи која представља биолошке мете. У бази података Drugbank/Bio2RDF класе откривене су класе *drugbank:Traget* и *drugbank:Drug*, а у PubMed/Bio2RDF бази класа *pubmed_vocabulary:journal*. Помоћно софтверско решење омогућава приказ хијерархијског нивоа до дубине 6. Табела 6.1 представља број класа првог хијерархијског нивоа база података иницијалних упита (резултати укључују и CPCTAS базу). Постојање релативно великог броја класа одређених база јасан је показатељ комплексности, која може имати ефеката на креирање коначних упита.

Табела 6.1 Број класа првог хијерархијског нивоа потенцијалних кандидата (база података) за креирање предефинисаних Federated SPARQL упита за шаблоне (*)

База података/ Репозиторијум	**Број класа на првом хијерархијском нивоу**	**Детектоване класе**
PIBAS/CPCTAS	132	*Target/Assay/ActiveSubstance/CellLine*
Reference/CPCTAS	1	*AcademicArticle*
Drugbank/Bio2RDF	2123	*Target/Drug*
PubMed/Bio2RDF	80	*Journal*
Chembl/EMBL-EBI	15	*Target/Assay/CellLine/SmallMolecule/Journal*
BindingDB/Chem2Bio2RDF	2	*BindingInteraction*

Иако је селекција извора података на основу класа иницијално добар приступ, постојање класа без присуства одговарајућих инстанци је безначајно за генерисање упита. Зато се приступило откривању инстанци. За ту сврху је коришћена опција за пребројавање инстанци помоћног софтверског решења, која на основу селектоване класе и *endpoint*-а базе којој класа припада генерише одређене SPARQL упите који имају за циљ утврђивање броја инстанци дате класе. Табела 6.2 представља резултат примене ове опције.

Табела 6.2 Број инстанци класа које припадају потенцијалним кандидатима (базама података) за креирање предефинисаних упита за шаблоне (*)

База података/Репозиторијум	**Класа**	**Број инстанци**
PIBAS/CPCTAS	*pibas:Target*	4
	pibas:ModelSystem	7
	pibas:Experiment	29
	pibas:Assay	5
	pibas:ActiveSubstance	45
Reference/CPCTAS	*bibo:AcademicArticle*	1688
Drugbank/Bio2RDF	*drugbank:Target*	4026
PubMed/Bio2RDF	*pubmed_vocabulary:journal*	19210000
Chembl/EMBL-EBI	*chembl:SingleProtein*	7031
	chembl:CellLineTarget	1068
	chembl:ProteinComplex	384
	chembl:Assay	1060283
	chembl:CellLine	1667
	chembl:Compound	1701313
	chembl:Journal	233
BindingDB/Chem2Bio2RDF	*bindingdb:bindingdb_interaction*	57144

Међутим, на овај начин могу изостати адекватне повратне информације. На пример, у Chembl/EMBL-EBI бази података егзистирају класе *chembl:SingleProtein* и *chembl:CellLineTarget* (из другог и трећег хијерархијског нивоа) које су потклассе класе *chembl:Traget*. Класа *chembl:SingleProtein* има 7031 инстанцу, а класа *chembl:CellLineTarget* 1068 инстанци, што би значило да је број инстанци класе *chembl:Traget* једнак 8103 (7031+1068+4). Ово подразумева да је неопходно одређено знање о домену, односно онтолошкој бази података чија таксономија се разматра. Постојање великог броја инстанци, показатељ је да су дате базе података погодне за финалне упите, јер представљају значајан ресурс за откривање релевантних информација.

Како је за поједине шаблоне (*) било неопходно проширити иницијалне упите новим обрасцима који би омогућили примену *InChiKey* или *SMILES* параметара, од кључног интереса је било утврдити и својства одговарајућих класа. Циљне класе у овом случају биле су оне које се односе на лекове (активне супстанце), с обзиром да се дати идентификатори користе за њихову идентификацију. За ту сврху делимично је коришћена опција помоћног софтвера, која омогућава проверу својстава. Уз помоћ ове опције могуће је утврдити својства која припадају класама, али не и она која припадају потклasама или надкалсама. Упити за преузимање ових података су зависили од таксономије базе, тако да није било могуће дефинисати униформне упите. Слика 6.5 представља упите за откривање предиката који се односе на *InChiKey* и *SMILES* параметре за Chembl/EMBL-EBI и BindingDB/Chem2Bio2RDF базе података. Упити су креирани за насумично одабране инстанце типа *chembl:smallMolecule* и *bindingdb:bindingdb_ligand*. Њиховим извршавањем откривени су атрибути *sio:SIO_000008* и *bindingdb:inchikey* из чијих даљих описа је утврђено да означавају везу ка *InChiKey*/*SMILES* параметрима. На сличан начин може се утврдити присутност својстава која покривају ове параметре и у другим базама података.

```
SELECT DISTINCT *
    WHERE {
      <http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL100177> ?hasProperty ?property.
      ?property ?hasInichiKey ?InichiKey.
      FILTER (contains (str(?property), "inchi") || contains (str(?property), "smiles")).
}
```

а)

```
SELECT DISTINCT *
       WHERE{
       <http://chem2bio2rdf.org/bindingdb/resource/bindingdb_ligand/23850> ?hasProperty ?property.
       FILTER (contains(str(?hasProperty), "inchi") || contains(str(?hasProperty), "smiles")).
}
```

б)

Слика 6.5 Примери SPARQL упита за откривање *InChiKey* и *SMILES* параметара. Упит (а) се извршава над инстанцом типа *chembl:SmallMolecule*, а упит (б) над инстанцом типа *bindingdb:bindingdb_ligand*

Од значаја за поједине шаблоне (*) је и присуство својстава која референцирају ка IC_{50} вредностима. На основу таксономије откривене су циљне класе које располажу атрибутима овог типа. На пример, код Chembl/EMBL-EBI базе података од важности је класа *chembl:Activity*, чије инстанце поседују својство *cco:standardType* које представља конекцију ка IC_{50} вредностима. У BindingDB/Chem2Bio2RDF бази података, конекција ка овом параметру откривена је преко класе *bindingdb:Target*. Инстанце овог типа користе *bindingdb:ic50_value* за представљање IC_{50} вредности. Код других класа нису откривени подаци овог типа.

На основу истраживања која су обављена у овом кораку методологије може се закључити да су селектовани иницијални упити одлична основа за креирање коначних Federated SPARQL упита. Међутим, пре финалног корака је спроведен процес онтолошког поравнања и селекције извора података у циљу укључивања додатних образаца (база података) у предефинисане упите.

6.1.2.3 Онтолошко поравнање и селекција извора података

Имајући у виду дефиницију онтолошког поравнања, приступа се процесу поређења инстанци, како би се открили међусобни односи и утврдило да ли на неки начин они могу утицати на финалне упите, јер се откривањем истих или сличних ентитета који припадају различитим базама података повећава интероперабилност података. Интероперабилност је способност хетерогених система да раде заједно како би информације биле доступне кориснику, а да при томе нису потребне додатне операције за споразумевање два система [109]. Онтолошке базе података (и репозиторијуми којима они припадају) имају особину хетерогених система. У циљу савладавања хетерогености приступило се стратегији која је описана у наставку.

Иницијални корак онтолошког поравнања подразумева примену SELECT SPARQL упита, који

омогућавају преузимање свих предиката и објектних вредности инстанци. У овом случају кључни су предикати који означавају везу ка другим базама података. На пример, од значајнијих предиката инстанце *drugbank:DB00554* може се издвојити предикат *drugbank_vocabulary:x-kegg* и његова објектна вредност - инстанца *kegg:D02778*, која припада Kegg/Bio2RDF бази података. Поређењем ових инстанци утврђено је да оне представљају идентичне ентитете. Напоменућемо да су за поређење одговарајућих објектних вредности коришћени само мануелни приступи, односно људске процене, које су биле саставни део истраживања представљеног у раду [7]. Појава идентичних и сличних инстанци у овом кораку била је главна мотивација за развој алгоритма за детектовање сличних података, који је предложен у Седмом поглављу. Појава сличних (идентичних) инстанци је утицала да се узму обзир и скупови података који садрже такве типове података, јер се на тај начин могу преузети комплементарне информације. Како се тежило ка идентификацији нових скупова података који би могли да задовоље критеријуме шаблона (*), методологија за креирање упита предложена у овом истраживању усвојила је делимичну примену стратегије „одоздо на горе“ за селекцију извора података. Конкретно, URI спецификација *kegg:D02778*, искоришћена је за идентификацију нових веза и откривања даљих информација у Kegg/Bio2RDF бази података. На тај начин је утврђена егзистенција предиката *kegg_vocabulary:drug-target* који означава релацију ка биолошким метама. До сличних чињеница се може доћи захваљујући интеграцији PIBAS/CPCTAS базе података са Drugbank/Bio2RDF базом података. Дата стратегија је имала учинка и у откривању веза између лекова и публикација у ChemBL/EMBL-EBI бази података, што је од важности за шаблон (*e*): инстанце типа *chembl:Substance* су преко објектног својства *chembl:hasDocument* повезане са инстанцама типа *chembl:Document*, односно инстанцама класе *chembl:Journal*, која презентује класу публикација.

Након идентификације скупова података који могу допринети комплементарним подацима, приступа се модификацији иницијалних и креирању коначних Federated SPARQL упита.

6.1.2.4 Модификација иницијалних упита - креирање Federated SPARQL упита

На основу претходних корака може се закључити да су односи између ентитета у оквиру, али и између онтолошких база података прилично комплексни и да је несумњиво неизоставно одређено знање о домену пре него што се приступи креирању упита. Слика 6.6 представља коначне Federated SPARQL упите за шаблоне (*). Може се уочити да је суштинска разлика између иницијалних и модификованих образаца у укључивању *InChiKey* параметара који се користе за идентификацију лекова. У овом случају избор параметара који јединствено идентификују активну супстанцу вршен је произвољно. То заправо значи да се упит може модификовати према потреби коришћењем *SMILES* уместо *InChiKey* параметра. Код образаца који се односе на есеје коришћени су *SMILES* параметри. Такође, на основу онтолошких релација откривених у претходним корацима методологије, неки обрасци из иницијалних упита су уклоњени јер је процењено да нису од важности за шаблон. На пример, уклоњени су обрасци *?target cco:hasTargetComponent ?targetcmpt*; *?targetcmpt cco:targetCmptXref ?uniprot*; *?uniprot a cco:UniprotRef* који се користе у иницијалном упиту над Chembl/EMBL-EBI базом података. Такође, може су увидети да упити садрже и опцију филтрирања за IC_{50} вредност, јер је од важности било укључити експерименте позитивних исхода. Над неким обрасцима није било могуће применити филтрирање по IC_{50} вредностима, јер овакви атрибути нису били доступни у структури онтологије. Финални упити подржавају преузимање података и из CPCTAS базе. Такође, поред модификованих упита, одговарајући шаблони садрже и обрасце других база података код којих су откривени потенцијално корисни ентитети. Упит за шаблон (*a*) додатно укључује обрасце Kegg/Bio2RDF базе података, упит за шаблон (*b*) обрасце PubChem/Chem2Bio2RDF базе података, а упит за шаблон (*e*) обрасце Chembl/EMBL-EBI базе података. Такође, креиран је и упит за шаблон (*d*) који омогућава откривање информација о лековима (активним супстанцама).

```
PREFIX  pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX  drugbank:<http://bio2rdf.org/drugbank_vocabulary:>
PREFIX  bindingdb:<http://chem2bio2rdf.org/bindingdb/resource/>
PREFIX  kegg:<http://bio2rdf.org/kegg_vocabulary:>
PREFIX  sio:<http://semanticscience.org/resource/>
PREFIX  cco:<http://rdf.ebi.ac.uk/terms/chembl#>

SELECT DISTINCT ?Target ?Dataset
WHERE
  {   { ?activeSubstance pibas:hasInChiKey "%s".
        ?Experiment pibas:activeSubstance ?activeSubstance.
        ?Experiment pibas:hasTarget ?Target;
                    pibas:IC50 ?ic50value.
        FILTER(?ic50value<300000.0).
        BIND("PIBAS/CPCTAS" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <http://drugbank.bio2rdf.org/sparql>
          {
            ?calculated_properties drugbank:value "InChIKey=%s"^^<http://www.w3.org/2001/XMLSchema#string>.
            ?drugabank_drug drugbank:calculated-properties ?calculated_properties;
                            drugbank:target ?Target.
          }
        BIND("Drugbank/Bio2RDF" AS ?Dataset).
      }

       UNION
      { SERVICE SILENT <http://kegg.bio2rdf.org/sparql>
          { ?calucalted_properties drugbank:value "InChIKey=%s"^^<http://www.w3.org/2001/XMLSchema#string>.
            ?drugbank_drug drugbank:calculated-properties ?calucalted_properties;
                           drugbank:x-kegg ?kegg_drug.
            ?kegg_drug kegg:target ?Target.
          }
        BIND("Kegg/Bio2RDF" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/sparql>
          { ?chembl sio:SIO_000008 ?hasInChiKey.
            ?hasInChiKey sio:SIO_000300 "%s".
            ?activity cco:hasMolecule ?chembl.
            ?activity cco:hasAssay ?assay;
                      cco:standardType ?ic50;
                      cco:standardValue ?ic50value.
            ?assay cco:hasTarget ?Target.
            FILTER(?ic50value<300000.0).
          }
        BIND("Chembl/EMBL-EBI" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <http://cheminfov.informatics.indiana.edu:8080/bindingdb/sparql>
          { ?bindingdb_drug bindingdb:inchikey "%s".
            ?Target bindingdb:Monomerid ?bindingdb_drug;
                    bindingdb:ic50_value ?ic50.
            FILTER(?ic50<300000.0).
          }
        BIND("BindingDB/Chem2Bio2RDF" AS ?Dataset).
      }
      %s
   }
```

a)

```
PREFIX  pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX  pubchem:<http://chem2bio2rdf.org/pubchem/resource/>
PREFIX  cco:<http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX  sio:<http://semanticscience.org/resource/>
PREFIX  rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?Assay ?Dataset
WHERE
  {   { ?activeSubstance pibas:hasSmile "%s”.
        ?Experiment pibas:activeSubstance ?activeSubstance.
        ?Experiment pibas:experimentalMethod ?Assay;
                    pibas:IC50 ?ic50value.
        FILTER(?ic50value<300000.0).
        BIND("PIBAS/CPCTAS" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <http://cheminfov.informatics.indiana.edu:8080/pubchem/sparql>
          { ?assay_interaction pubchem:CID ?pubchem_compound.
            ?pubchem_compound pubchem:openeye_iso_smiles "%s".
            ?assay_interaction pubchem:AID ?Assay.
          }
        BIND("Pubchem/Chem2Bio2RDF" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/sparql>
          { ?hasSmile sio:SIO_000300 "%s".
            ?chembl_compound sio:SIO_000008 ?hasSmile.
            ?activity rdf:type cco:Activity;
                      cco:hasMolecule ?chembl_compound;
                      cco:hasAssay ?Assay;
                      cco:standardValue ?ic50value;
                      cco:standardType ?ic50.
            FILTER(?ic50value<300000.0).
          }
        BIND("Chembl/EMBL-EBI" AS ?Dataset).
      }
      %s
  }
```

б)

```
PREFIX  pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX  cco:<http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX  sio:<http://semanticscience.org/resource/>

SELECT DISTINCT ?CellLine ?Dataset
WHERE
  {   { ?activeSubstance pibas:hasInChiKey "%s"^^<http://www.w3.org/2001/XMLSchema#string>.
        ?Experiment pibas:activeSubstance ?activeSubstance.
        ?Experiment pibas:experimentalMethod ?Assay;
                    pibas:modelSystem ?CellLine;
                    pibas:IC50 ?ic50value.
         FILTER(?ic50value<300000.0).
         BIND("PIBAS/CPCTAS" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/sparql>
          { ?chembl_drug sio:SIO_000008 ?hasInChiKey.
            ?hasInChiKey sio:SIO_000300 "%s".
            ?activity cco:hasMolecule ?chembl_drug;
                      cco:hasAssay ?assay;
                      cco:standardType ?ic50;
                      cco:standardValue ?ic50value.
            ?assay cco:hasCellLine ?CellLine.
            FILTER(?ic50value<300000.0).

        BIND("Chembl/EMBL-EBI" AS ?Dataset).
        }
      }
    %s
  }
```

ц)

```
PREFIX  pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX  drugbank:<http://bio2rdf.org/drugbank_vocabulary:>
PREFIX  bindingdb:<http://chem2bio2rdf.org/bindingdb/resource/>
PREFIX  rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  dcterms:<http://purl.org/dc/terms/>
PREFIX  cco:<http://rdf.ebi.ac.uk/terms/chembl#>

SELECT DISTINCT ?Drug ?Dataset
WHERE
  {   { ?Target pibas:hasTargetName "%s".
        ?Drug pibas:hasTarget ?Target.
        BIND("PIBAS/CPCTAS" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <http://drugbank.bio2rdf.org/sparql>
          { ?target dcterms:title "%s"@en.
            ?Drug drugbank:target ?Target;
                  rdf:type drugbank:Drug.
          }
        BIND("Drugbank/Bio2RDF" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/sparql>
          { ?target dcterms:title "%s".
            ?assay cco:hasTarget ?Target.
            ?activity cco:hasAssay ?assay;
                      cco:hasMolecule ?Drug.
          }
        BIND("Chembl/EMBL-EBI" AS ?Dataset).
      }
    UNION
      { SERVICE SILENT <http://cheminfov.informatics.indiana.edu:8080/bindingdb/sparql>
          { ?binding_interaction bindingdb:TARGET "%s";
                                 bindingdb:Monomerid ?Drug.
          }
        BIND("BindingDB/Chem2Bio2RDF" AS ?Dataset).
      }
    %s
  }
```

д)

```
PREFIX  bibo:<http://purl.org/ontology/bibo/>
PREFIX  pubmed:<http://bio2rdf.org/pubmed_vocabulary:>
PREFIX  cco:<http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX  dcterms:<http://purl.org/dc/terms/>
PREFIX  rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

SELECT DISTINCT ?Title ?Dataset
WHERE
  {
    {
        ?Paper rdf:type bibo:AcademicArticle;
               dcterms:title ?Title.
        FILTER regex(?Title, "%s", "i").
        BIND("Reference/CPCTAS" AS ?Dataset).
      }
      UNION
      { SERVICE SILENT <http://lod.openlinksw.com/sparql>
          { ?Paper pubmed:journal ?journalRef;
                   dcterms:title ?Title.
            FILTER regex(?Title, "%s", "i").
          }
        BIND("Pubmed/Bio2RDF" AS ?Dataset).
      }
      UNION
      { SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/sparql>
          { ?document cco:documentType "PUBLICATION";
                      dcterms:title ?Title.
            FILTER regex(?Title, "%s", "i").
          }
        BIND("Chembl/EMBL-EBI" AS ?Dataset).
      }
      %s
  }
```

е)

Слика 6.6 Предефинисани Federated SPARQL упити који се користе за откривање биолошких мета (а), есеја (б), ћелијских линија (ц), лекова (д) и публикација (е) на Платформи

Сви упити садрже карактере „%s“ који ће бити замењени одговарајућом кључном речју коју корисник уноси на Платформи. Последњи карактер сваког упита потенцијално је резервисан за додавање кориснички селектованог скупа података (новог елемента уније) што је описано у одељку 6.3. Пре него што се упити складиште у *DataSources* онтологију, приступа се последњем кораку предложене методологије - процени и вредновању њихових резултата.

6.1.2.5 Процена и вредновање упита

Биоинформатичка истраживања захтевају строгу контролу квалитета података, јер потенцијалне грешке могу директно утицати на резултат истраживања [98]. Конкретно, у домену преклиничког тестирања лекова то би значило негативан ефекат на здравље људи. Квалитет података у биоинформатици је јако важан и укључује питања поверења у одређене изворе података који се користе током истраживања [98]. Зато је јако битно извршити процену и вредновати резултате упита пре него што се они директно користе у процесу самог истраживања. У овом случају, процену и вредновање упита вршило је особље Лабораторије. Сваки предефинисани упит је тестиран за одређене улазне вредности (*InChiKey*, *SMILES* или текст вредности) и вршена је:

1. **Провера тачности резултата** - да ли резултати упита (URI спецификације) заиста припадају одговарајућим базама података које се користе за изградњу Federated SPARQL упита;
2. **Провера квалитета резултата** - да ли резултати упита (URI спецификације) заиста одговарају критеријумима претраге који су дефинисани за изградњу шаблона (*) представљених у одељку 4.5.3.

Валидација претходних критеријума вршена је у оквиру истраживања [7] и доступна је на адреси *http://cpctas-lcmb.pmf.kg.ac.rs/fed/evaluation/*. За потребе дисертације извршена је и додатна валидација упита[86]. Табела 6.3 представља резултат валидације. Прва колона је резервисана за ознаку шаблона, док друга колона одговара кључним речима - *InChiKey/SMILES* параметрима или текстуалним вредностима. Имајући у виду област рада Лабораторије, већина улазних параметара се односи на активне супстанце који се користе у процесу лечења канцера. Упити су извршени над *remote endpoint*-има, што заправо означава да резултати могу варирати у реалном времену. Последње две колоне су резервисане за валидацију резултата. Напоменућемо да су упити за биолошке мете, ћелијске линије и есеје додатно валидирани кроз систем SpecINT [96], који такође представља један од резултата овог истраживања у домену семантичког веба.

Табела 6.3 Валидација резултата извршених предефинисаних упита

		Валидација резултата	
Шаблон	**Кључна реч**	**Провера тачности**	**Провера квалитета**
a/c	GHASVSINZRGABV-UHFFFAOYSA-N	Да	Да
a/c	DQLATGHUWYMOKM-UHFFFAOYSA-L	Да	Да
a/c	RCINICONZNJXQF-MZXODVADSA-N	Да	Да
a/c	PTOAARAWEBMLNO-KVQBGUIXSA-N	Да	Да
a/c	KTUFNOKKBVMGRW-UHFFFAOYSA-N	Да	Да
a/c	BIIVYFLTOXDAOV-YVEFUNNKSA-N	Да	Да
a/c	FOCVUCIESVLUNU-UHFFFAOYSA-N	Да	Да
b	FC1=CNC(=O)NC1=O	Да	Да
b	[NH2-].[NH2-].Cl[Pt+2]Cl	Да	Да
d	GHASVSINZRGABV-UHFFFAOYSA-N	Да	Да
d	DQLATGHUWYMOKM-UHFFFAOYSA-L	Да	Да
e	Cisplatin	Да	Да
e	Fluorouracil	Да	Да

С обзиром да су задовољени услови и последњег корака предложене методологије, упити су складиштени у *DataSources* онтологију, као вредности предиката *pibas:hasInitalQuery* одговарајућих

[86] Резултати упита који су коришћени за додатну валидацију доступни су на *https://figshare.com/articles/Rezultati_izvrsavanja_predefinisanih_upita_za_proces_validacije/7667387*

инстанци класе *pibas:Templates*. Овакви упити су применљиви на Платформи и може се приступити методи извршавања предефинисаних упита.

6.1.3 Извршавање предефинисаних упита

Процес извршавања предефинисаних упита на Платформи могућ је након селекције одговарајућих критеријума (теме, подтеме и шаблона) и уноса кључне речи на иницијалном корисничком интерфејсу. Слика 5.3 представља кориснички интерфејс за селектовану тему *Chemogenomic*, подтему *Targets* и шаблон *Find targets for the drug*. Избор теме врши се селекцијом из падајуће листе *Select Topic*. Вредности - *Biology*, *Chemogenomic* и *Research* - које су доступне кроз ову падајућу листу одговарају објектним вредностима својства *pibas:topicName* који се везује за инстанце класе *pibas:Topics* у *DataSources* онтологији. Избор теме аутоматски ограничава избор подтеме због објектног предиката *pibas:hasSubTopic*, који за домен има класу *pibas:Topics*, а за кодомен класу *pibas:SubTopic*. Избор подтеме аутоматски ограничава избор шаблона због објектног својства *pibas:hasTemplate*, који за домен има класу *pibas:SubTopic*, а за кодомен класу *pibas:Templates.* Такође, избор шаблона одређује и тип кључне речи која се очекује од корисника, а који је одређен својством *pibas:hasInput.* Претрага по кључним речима (енгл. *keyword-based web searches*) је кориснички једноставна (енгл. *user-friendly*) и чест је пример претраге за разна биоинформатичка истраживања [110,111]. Претпоставимо да је корисник заинтересован за откривање биолошких мета који су у интеракцији са леком *Fluorouracil* [112]. Како се у овом случају као кључна реч очекује *InChiKey* параметар лека тако корисник у поље *Enter InChiKey* уноси вредност *GHASVSINZRGABV-UHFFFAOYSA-N*. Кликом на дугме *Run query* компонента *PredefinedQuery* „скенира“ *DataSources* онтологију и проналази „поклапајући“ упит. Процес скенирања се у овом случају своди на екстракцију објектне вредности својства *pibas:hasInitialQuery*, што се постиже извршавањем одговарајућег SELECT SPARQL упита (Слика 6.7). Промењљива *$templateid* биће замењена идентификационим бројем селектованог шаблона. Повратна вредност датог упита еквивалентна је предефинисаном упиту који представља Слика 6.6 а).

```
PREFIX pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
SELECT ?initilaquery
WHERE
{
  ?template pibas:id '" . $templateid . "'^^xsd:int.
  ?template pibas:hasInitialQuery ?initilaquery.
}
```

Слика 6.7 SELECT SPARQL упит за преузимање објектне вредности предиката *pibas:hasInitialQuery* из *DataSources* онтологије за селектовани шаблон на Платформи. Промењљива *$templateid* означава *ID* шаблона

Компонента *QueryPreparation* врши процесирање предефинисаног упита и замену $n-1$ карактера „%s“ унетом кључном речју, након чега компонента *SPARQLQueryRunner* извршава упит. Упити у *PHP* коду се извршавају применом одговарајућих метода PHP SPARQL Lib библиотеке. Резултат извршеног предефинисаног упита се парсира и креира се JSON објекат, који идентификује скупове података над којима је упит извршен, као и излазне параметре (у овом случају биолошке мете). Слика 6.8 представља пример структуре JSON објекта који се користи за презентацију резултата, о чему је детаљније дискутовано у наредном одељку.

```
{
  "Variable1": "Target",
  "Variable2": "Dataset",
  "children": [
     {
     "Target": "http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#TestTarget1",
     "Dataset": "PIBAS/CPCTAS"
     },
     {
       "Target": "http://bio2rdf.org/drugbank:BE0004796",
       "Dataset": "Drugbank/Bio2RDF"
     },
     {
       "Target": "http://bio2rdf.org/drugbank:BE000324",
       "Dataset": "Drugbank/Bio2RDF
     },
     {
       "Target": "http://bio2r0f.org/drugbank:BE0004810",
       "Dataset": "Drugbank/Bio2RDF"
     },
     {
       "Target": "http://bio2rdf.org/kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71”,
       "Dataset": "Kegg/Bio2RDF"
     },
     {
       "Target": "http://rdf.ebi.ac.ok/resource/chembl/target/CHEMBL612518",
       "Dataset": "Chembl/EMBL-EBI"
     }
             .
             .
             .
   ]
}
```


Слика 6.8 Делимични приказ JSON објекта добијеног извршавањем предефинисаног упита за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* на Платформи

6.1.3.1 Презентација резултата

Презентација резултата извршавања предефинисаних упита на Платформи има табеларну форму (Слика 6.9), која настаје процесирањем JSON објекта (Слика 6.8).

Target	Dataset
Search...	289
http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#TestTarget1	PIBAS/CPCTAS
http://bio2rdf.org/drugbank:BE0004796	Drugbank/Bio2RDF
http://bio2rdf.org/drugbank:BE0000324	Drugbank/Bio2RDF
http://bio2rdf.org/drugbank:BE0004810	Drugbank/Bio2RDF
http://bio2rdf.org/kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71	Kegg/Bio2RDF
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL612518	Chembl/EMBL-EBI
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL613490	Chembl/EMBL-EBI

Слика 6.9 Делимични приказ резултата извршавања предефинисаног упита за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* на Платформи

Прва колона табеле садржи URI спецификације (односно инстанце одговарајућих база података) које у овом примеру представљају биолошке мете. Вредности дате колоне одговарају промењивој *Variable1* из JSON објекта. Приказ URI спецификација је погоднији него да је у обзир узета само ознака (објектна вредност предиката *rdfs:label* или *rdfs:name*) инстанце. Главни разлог оваквог начина представљања података је у томе што су многе инстанце јавно доступне, а то заправо означава да се може приступити њиховим веб дескрипцијама (енгл. *description*), који садрже њихове предикате и објектне вредности (Слика 6.10). Чест је случај да дескрипције не садрже све триплете, што може бити одређено правилима репозиторијума или саме базе података. Такође, неке инстанце иако су линковане немају јавно доступне дескрипције, као што је то случај са CPCTAS базом. У оваквим ситуацијама, без обзира на то што дескрипције нису јавне, не значи да подаци дате инстанце нису доступни кроз одговарајуће *remote endpoint*-е. Друга колона табеле служи за приказ назива базе података и репозиторијума коме резултујућа инстанца припада. Ова колона одговара промењивој *Variable2* из JSON објекта. Резултујућа табела се може сортирати и претраживати по тексту. Овакав вид представљања резултата предефинисаних упита

важи за све шаблоне. Треба напоменути да резултати упита могу варирати јер се изводе над *remote endpoint*-има, који могу бити недоступни услед велике фреквенције коришћења или блокираности.

About: Thymidylate synthase [drugbank:BE0000324] Goto Sponge NotDistinct Permalink
An Entity of Type : http://bio2rdf.org/drugbank_vocabulary:Target, within Data Space : bio2rdf.org associated with source document(s)
Type: Target [drugbank_vocabulary:Target] New Facet based on Instances of this Class

Attributes	Values
rdf:type	drugbank resource [drugbank_vocabulary:Resource] Target [drugbank_vocabulary:Target]
rdfs:label	Thymidylate synthase [drugbank:BE0000324]
dcterms:title	Thymidylate synthase
dcterms:identifier	drugbank:BE0000324
void:inDataset	http://bio2rdf.org/drugbank_resource:bio2rdf.dataset.drugbank.R3
Bio2RDF identifier	BE0000324
Bio2RDF namespace	drugbank
Bio2RDF uri	http://bio2rdf.org/drugbank:BE0000324
identifiers.org URI	http://identifiers.org/drugbank/BE0000324
amino acid sequenc...no-acid-sequence]	>Thymidylate synthase PVAGSELPRRPLPPAAQERDAEPRPPHGELQYLGQIQHILRCGVRKDDRTGTGTLSVFGM...
gene name [drugban...bulary:gene-name]	TYMS
gene sequence [dru...ry:gene-sequence]	>942 bp ATGCCTGTGGCCGGCTCGGAGCTGCCGCGCCGGCCCTTGCCCCCCGCCGCACAGGAGCGG...
general function [...general-function]	Nucleotide transport and metabolism
go function [drugb...lary:go-function]	catalytic activity transferase activity methyltransferase activity transferase activity, transferring one-carbon groups 5,10-methylenetetrahydrofolate-dependent methyltransferase activity »more»
go process [drugba...ulary:go-process]	metabolism physiological process cellular metabolism nucleotide metabolism pyrimidine nucleotide biosynthesis »more»
locus [drugbank_vocabulary:locus]	18p11.32
molecular weight [...molecular-weight]	35585.000000(xsd:float)
name [drugbank_vocabulary:name]	Thymidylate synthase
organism [drugbank_vocabulary:organism]	Human
synonym [drugbank_vocabulary:synonym]	TS TSase EC 2.1.1.45
theoretical pi [dr...y:theoretical-pi]	7.000000(xsd:float)
transmembrane regi...membrane-regions]	None
x genatlas [drugba...ulary:x-genatlas]	http://bio2rdf.org/genatlas:TYMS
x genbank [drugban...bulary:x-genbank]	http://bio2rdf.org/genbank:X02308
x genecards [drugb...lary:x-genecards]	http://bio2rdf.org/genecards:TYMS
x gi [drugbank_vocabulary:x-gi]	http://bio2rdf.org/gi:37479
x hgnc [drugbank_vocabulary:x-hgnc]	Gene Symbol for TYMS [hgnc:12441]
x pfam [drugbank_vocabulary:x-pfam]	http://bio2rdf.org/pfam:PF00303
x taxonomy [drugba...ulary:x-taxonomy]	http://bio2rdf.org/taxonomy:9606
x uniprot [drugban...bulary:x-uniprot]	Thymidylate synthase [uniprot:P04818] http://bio2rdf.org/uniprot:TYSY_HUMAN
is enzyme [drugbank_vocabulary:enzyme] of	Methotrexate [drugbank:DB00563] Fluorouracil [drugbank:DB00544] drugbank:DB00544 to drugbank:BE0000324 relation [drugbank_resource:DB00544_BE0000324] drugbank:DB00563 to drugbank:BE0000324 relation [drugbank_resource:DB00563_BE0000324]
is target [drugbank_vocabulary:target] of	Raltitrexed [drugbank:DB00293] drugbank:DB00322 to drugbank:BE0000324 relation [drugbank_resource:DB00322_BE0000324] Trifluridine [drugbank:DB00432] drugbank:DB00432 to drugbank:BE0000324 relation [drugbank_resource:DB00432_BE0000324] drugbank:DB00440 to drugbank:BE0000324 relation [drugbank_resource:DB00440_BE0000324] »more»

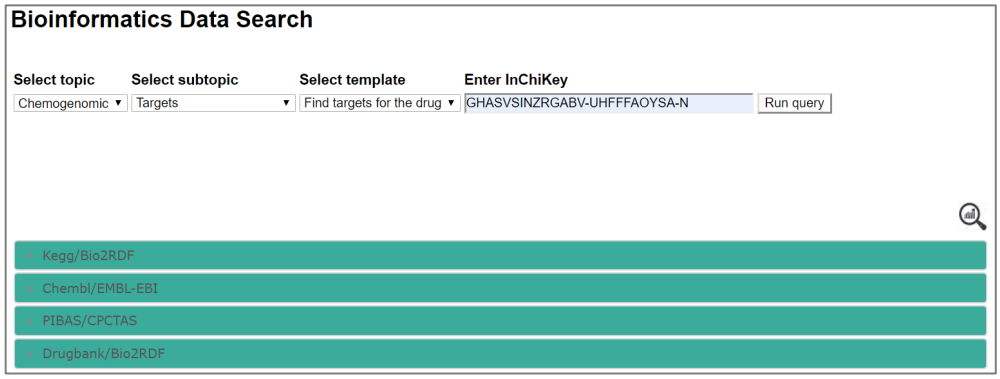
Слика 6.10 Делимични приказ дескрипције *drugbank:BE0000324* инстанце

Додатно, Платформа омогућава и опцију статистичког приказа резултата (Слика 5.5). На *pop-up* прозору су представљене базе података које се користе у предефинисаном упиту, као и број резултата (инстанци) који одговара свакој бази. На овај начин корисници могу имати представу о томе колико су неке базе „богате“ одређеним подацима. Дати резултат може утицати на даље опредељење корисника и његову евентуалну намеру да детаљније истражи инстанце које припадају „богатијим“ базама.

Функционалност извршавања предефинисаних упита има за циљ да заинтригира кориснике са релативно мало искуства у претрази семантичких база података, и укаже им на то колико су иницијални подаци важни. Осмо поглавље је резервисано за опсежну дискусију резултата предефинисаних упита и кроз различите тестне сценарије је објашњено на који начин су резултати од значаја за даљи ток истраживања. Такође, уз напредније функционалности Платформе указује се на који начин се иницијални резултати могу даље искористити, и како то утиче на повећање потенцијала да се открије знање које је од важности за планирање будућих истраживања.

6.2 Динамичко филтрирање резултата упита

Иницијални резултати предефинисаних упита довољни су да заинтересују корисника и упуте га на базе података које могу бити од важности за његов истраживачки рад. Приступ појединачним инстанцама (дескрипцијама) од велике је важности јер се могу открити релевантне информације, као и везе ка другим базама података. Међутим, проблем са оваквим приступом је у томе што су резултати предефинисаних упита често бројни (по неколико стотина или хиљада инстанци), што би у том случају значило да корисници морају приступати свакој инстанци појединачно. Како би се овај приступ олакшао, имплементирана је метода динамичког филтрирања резултата упита, која има за циљ убрзање анализе и побољшање релевантности резултата, уз елиминацију потребе да корисник поседује вештине о технологијама семантичког веба. Примена ове методе могућа је након извршавања предефинисаних упита. Корисник има могућност да кликом на дугме *Filter query* отвори нове елементе корисничког интерфејса - *accordion* елементе - панеле (Слика 6.11).



Слика 6.11 Кориснички панели методе динамичког филтрирања резултата предефинисаног упита за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* на Платформи

Сваки панел се односи на једну базу података која је коришћена у предефинисаном упиту. Базе које немају повратних вредности (као што је у овом случају BindingDB/Chem2Bio2RDF) нису приказане међу панелима. Кликом на одговарајући панел компонента *DynamicQueryFilter* (у комуникацији са *DataSources* онтологијом) најпре попуњава панел подацима који се односе на опис базе. Овај процес се заснива на извођењу SPARQL упита, који преузима објектне вредности предиката *pibas:comment* и *pibas:link*. Затим се генерише упит који има за циљ да на основу селектоване базе и инстанци (које су добијене као резултат извршавања предефинисаних упита) преузме предикате који одговарају инстанцама те базе. Слика 6.12 представља пример таквог упита за PIBAS/CPCTAS панел. Добијени предикати, који одговарају актуелној RDF структури дате базе, користе се за попуњавање панела. С обзиром да се упити креирају динамички садржај панела може варирати. Слика 5.7 представља пример PIBAS/CPCTAS панела.

```
PREFIX pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?predicate ?description
WHERE
 { { <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#TestTarget1> ?predicate ?object.
      OPTIONAL
        { ?predicate rdfs:label ?description. }
   }
  UNION
   { ?subject ?predicate <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#TestTarget1>.
      OPTIONAL
        { ?predicate rdfs:label ?description. }
   }
}
```

Слика 6.12 Пример SPARQL упита за добијање предиката за попуњавање PIBAS/CPCTAS панела за примену методе динамичког филтрирања резултата на Платформи

Панел PIBAS/CPCTAS базе попуњен је предикатима: *rdf:type*, *pibas:hasTargetName, pibas:targeType, pibas:isTargetOf* и *pibas:hasSynonim*. Треба напоменути да динамички упити преузимају искључиво предикате које се односе на инстанце, а не на класе у целини. Разлог томе је што неке инстанце не поседују предикате који припадају директним надкласама. На пример, инстанца *drugbank:BE0000324* поседује предикат *drugbank_vocabulary:synonym*, који није дефинисан као предикат директне надкласе *drugbank:Target*, већ надкласе *drugbank_vocabulary:Resource*. Сви предикати у оквиру једног панела садрже описе који одговарају ознакама, односно објектним вредностима *rdfs:label* предиката. Ови описи су доступни преласком преко иконице . У случају да вредности *rdfs:label* предиката не постоје, описи ће одговарати линковима предиката. Сваки линк омогућава приступ веб дескрипцијама датих предиката (слично дескрипцијама инстанци). На основу ових дескрипција корисник може одабрати предикате који су у складу са интересовањима и селектовати их. У оквиру сваког панела налази се дугме *Add to query*, које уз помоћ компоненти *DynamicQueryFilter* и *QueryPreparation*, преузима селектоване предикате. У случају да је селектован макар један предикат из неког панела, дугме *Run query* се ажурира у дугме *Run new query*, наговештавајући кориснику да постоји могућност извођења звездастих SPARQL упита.

6.2.1 Звездасти SPARQL упити

Током истраживања која су обављена за потребе динамичког филтрирања резултата упита, звездасти SPARQL упити [113] су се показали као најпогоднији, јер омогућавају преузимање различитих информација који се односе на једну инстанцу. У наставку следи дефиниција.

Дефиниција 1. Триплети облика $\{?s\ ?p_i\ ?o_i\}$, $1 \leq i \leq n$, такви да је $?s \neq ?p_i, s \neq ?o_i$ и $?p_i \neq ?o_i$ дефинишу звездасте SPARQL упите.

Претпоставимо да је корисник селектовао предикате *pibas:targetType* и *chembl:targetType* у PIBAS/CPCTAS и Chembl/EMBL-EBI панелима, респективно. Слика 6.13 представља примере звездастог SPARQL упита за селектоване предикате. Код првог упита од значаја је триплет *?Target pibas:targetType ?targetType,* а код другог упита триплет *?Target cco:targetType ? targetType*. У овом случају променљива *?Target* има улогу субјекта и важи да је:

- *?Target ≠ pibas:targetType*
- *?Target ≠ cco:targetType*
- *?Target ≠ ?targetType*

Примена звездастих SPARQL упита знатно олакшава процес извођења самих упита јер се *endpoint*-у приступа само једном, без обзира на број селектованих предиката. На тај начин се скраћује време добијања резултата.

```
PREFIX pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
SELECT DISTINCT ?Target ?Dataset ?targetType
WHERE
  { ?activeSubstance pibas:hasInChiKey "GHASVSINZRGABV-UHFFFAOYSA-N".
    ?Experiment  pibas:activeSubstance  ?activeSubstance;
                 pibas:hasTarget ?Target;
                 pibas:IC50 ?ic50value.
    FILTER (?ic50value < 300000.0).
    ?Target <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#targetType> ?targetType.
    BIND("PIBAS/CPCTAS" AS ?Dataset).
  }
```

а)

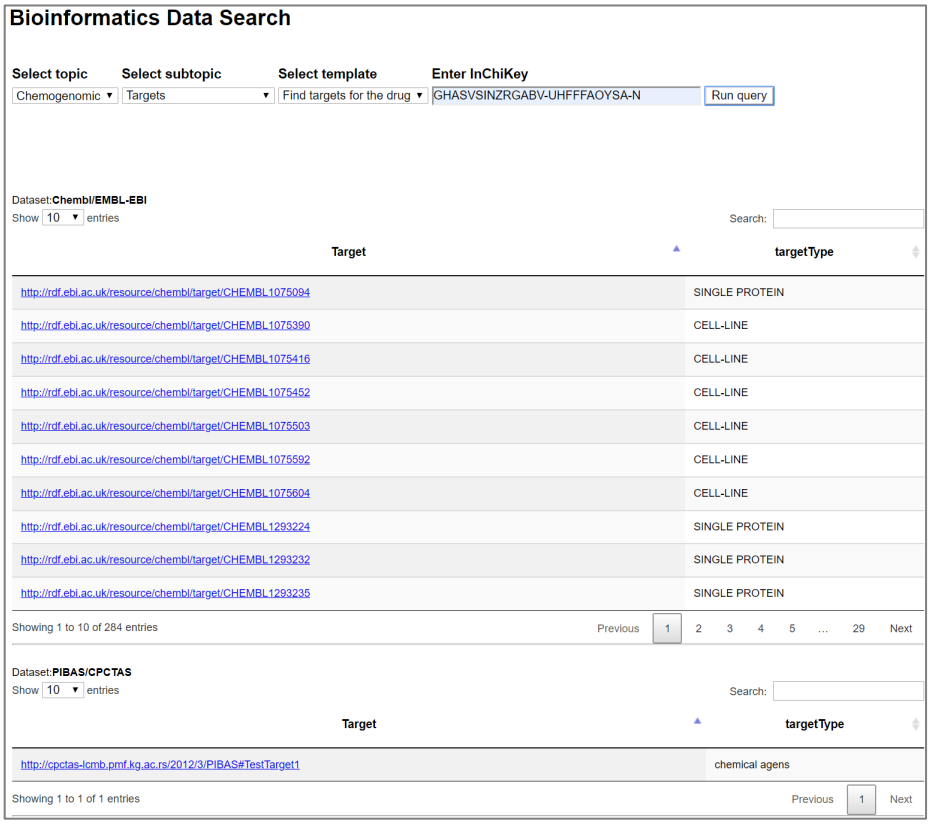
```
PREFIX sio:<http://semanticscience.org/resource/>
PREFIX cco:<http://rdf.ebi.ac.uk/terms/chembl#>
SELECT DISTINCT ?Target ?Dataset ?targetType
WHERE
  { SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/sparql>
      { ?chembl sio:SIO_000008 ?hasInChiKey.
        ?hasInChiKey sio:SIO_000300 "GHASVSINZRGABV-UHFFFAOYSA-N".
        ?activity cco:hasMolecule ?chembl.
        ?activity cco:hasAssay ?assay;
                  cco:standardType ?ic50;
                  cco:standardValue ?ic50value.
        ?assay cco:hasTarget ?Target.
        ?Target cco:targetType ?targetType.
        FILTER(?ic50value<300000.0).
      }
    BIND("Chembl/EMBL-EBI" AS ?Dataset).
  }
```

б)

Слика 6.13 Примери звездастих SPARQL упита за примену методе динамичког филтрирања резултата предефинисаних упита за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* на Платформи

6.2.2 Презентација резултата

Резултат извршавања претходно генерисаних звездастих SPARQL упита представљен је у форми подељених табела (енгл. *pagination table*), које су имплементиране применом *jQuery-a* (Слика 6.14). Овакав вид представљања података је јако погодан када звездасти SPARQL упити имају велики број повратних резултата.

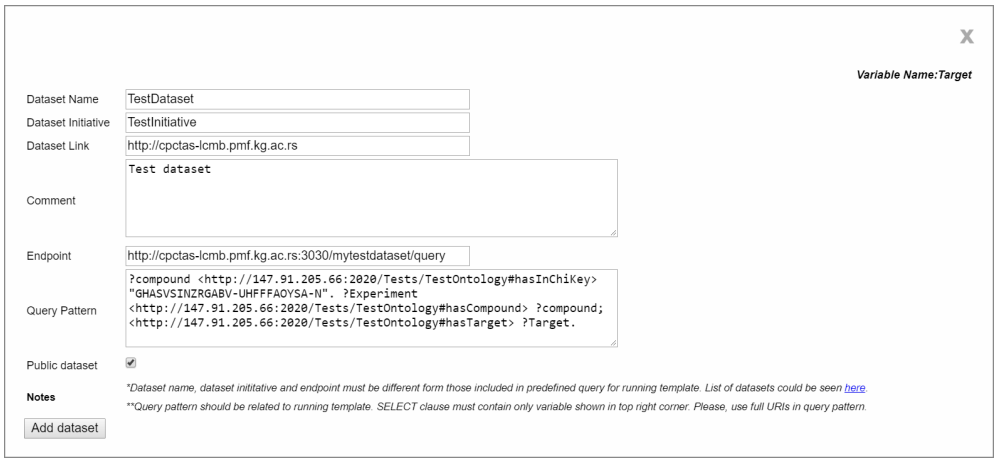


Слика 6.14 Резултат примене методе динамичког филтрирања резултата за селектовани предикат *pibas:targetType* из PIBAS/CPATAS панела и *chembl:targetType* из Chembl/EMBL-EBI панела

Резултат подељених табела је увек организован по извору података, али и по селектованим предикатима. Корисник уз помоћ опције *Show* може изабрати број ентитета (10, 25, 50 или 100) који ће бити приказан по извору података, вршити сортирање и претрагу колона на основу вредности унетих у *Search* поље. На овај начин се побољшава презентација резултата у смислу приказа најрелевантнијих података. Такође, корисник може вршити и мануелно поређење резултата у оквиру једне или више база, што може сугерисати правце његових даљих истраживања.

6.3 Додавање кориснички селектованог скупа података

Имплементација методе која омогућава додавање кориснички селектованог скупа података у предефинисане упите, настала је као подршка мањим научно-истраживачким институцијама и организацијама да своје податке представе широј јавности и да присуством на биоинформатичкој сцени скрену пажњу других истраживача. Метода додавања кориснички селектованог скупа података омогућена је за сваки предефинисани упит на Платформи. Корисник кликом на дугме *Add new dataset*, отвара *pop-up* форму која омогућава унос података (Слика 6.15).



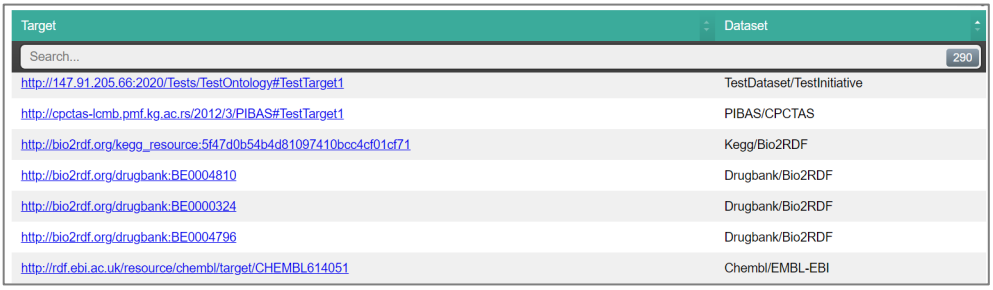
Слика 6.15 Форма за додавање кориснички селектоване базе података на Платформи попуњена тест подацима

У циљу постизања униформности подаци које корисник уноси на форми су:

- *Dataset Name* - јединствен назив базе података који је доступан на GUI-у;
- *Dataset Initiative* - јединствен назив репозиторијума који је доступан на GUI-у (ако назив репозиторијума није унет, користи се вредност поља *Dataset Name*);
- *Dataset Link* - URL базе података (вредност овог поља користи се за методу динамичког филтрирања резултата упита);
- *Comment* - крађи опис базе података (вредност овог поља користи се за методу динамичког филтрирања резултата упита);
- *Endpoint* - *endpoint* базе података коју корисник уноси;
- *Query Pattern* - образац (подупит) који ће постати део предефинисаног упита. У белешкама (*Notes*) је наглашено која променљива мора бити коришћена у оквиру обрасца. Ово поље захтева познавање SPARQL синтаксе;
- *Public dataset* - *checkbox* поље које омогућава кориснику да своју базу података представи јавности.

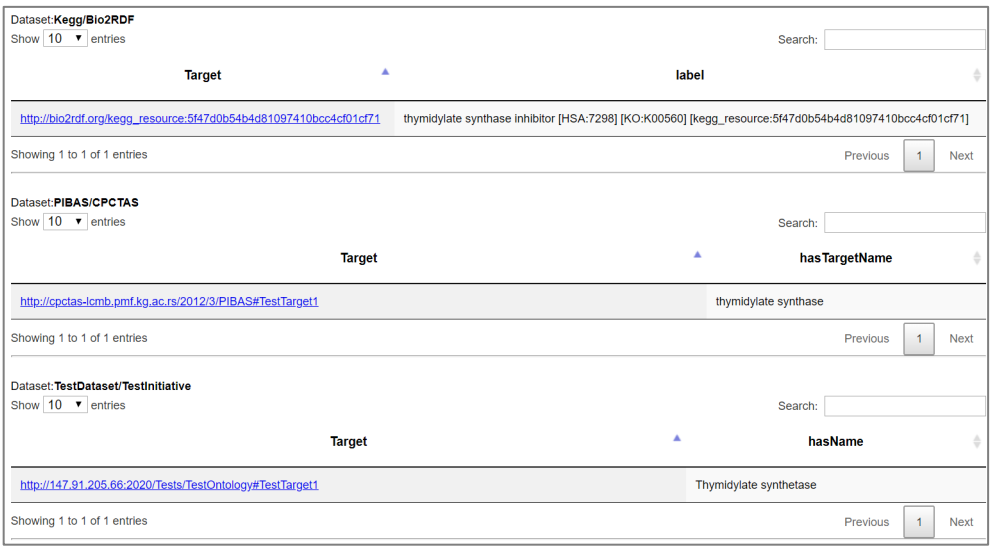
Кликом на дугме *Add dataset* компонента *PredefinedQueryExtension* преузима податке које је корисник унео кроз форму. Подаци се затим обрађују, процесирају и припремају за складиштење у *DataSources* онтологију уколико је корисник прихватио опцију *Public dataset*. У том случају, процес ажурирања *DataSources* онтологије врши се извођењем UPDATE SPARQL упита и онтологија се проширује подацима са форме. Да би новододата база података била јавна на Платформи, потребно је одобрење администратора. Након процеса додавања кориснички селектоване базе, може се извршити „проширени“ предефинисани упит (Слика 6.16), а затим применити и метода динамичког филтрирања резултата

(Слика 6.17). Уз помоћ ове методе корисник може упоредити податке своје базе са подацима осталих база у упиту. Тиме се проширује потенцијал откривања нових, релевантних, комплементарних и сличних (идентичних) информација.



Target	Dataset
Search...	290
http://147.91.205.66:2020/Tests/TestOntology#TestTarget1	TestDataset/TestInitiative
http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#TestTarget1	PIBAS/CPCTAS
http://bio2rdf.org/kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71	Kegg/Bio2RDF
http://bio2rdf.org/drugbank:BE0004810	Drugbank/Bio2RDF
http://bio2rdf.org/drugbank:BE0000324	Drugbank/Bio2RDF
http://bio2rdf.org/drugbank:BE0004796	Drugbank/Bio2RDF
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL614051	Chembl/EMBL-EBI

Слика 6.16 Резултат извршавања методе предефинисаног упита са кориснички селектованом базом података за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* на Платформи



Dataset:**Kegg/Bio2RDF**
Show 10 entries Search:

Target	label
http://bio2rdf.org/kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71	thymidylate synthase inhibitor [HSA:7298] [KO:K00560] [kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71]

Showing 1 to 1 of 1 entries Previous 1 Next

Dataset:**PIBAS/CPCTAS**
Show 10 entries Search:

Target	hasTargetName
http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#TestTarget1	thymidylate synthase

Showing 1 to 1 of 1 entries Previous 1 Next

Dataset:**TestDataset/TestInitiative**
Show 10 entries Search:

Target	hasName
http://147.91.205.66:2020/Tests/TestOntology#TestTarget1	Thymidylate synthetase

Showing 1 to 1 of 1 entries Previous 1 Next

Слика 6.17 Резултат извршавања методе динамичког филтрирања резултата са кориснички селектованом базом података за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* Платформи

Може се закључити да су презентоване методе Платформе имплементиране тако да омогућавају корисницима, који немају превеликог искуства у претрази и анализи семантичких података, да релативно једноставно врше претрагу биоинформатичких база података. Циљ ових метода је да укажу на велику количину доступних података и њихову релевантност за даља истраживања. Значаји резултата ових метода су систематичније приказани кроз тестне сценарије у Осмом поглављу.

6.4 Компаративни преглед литературе

Решења која би допринела откривању нових информација и знања, која могу утицати на квалитетније процесе истраживања, од великог су значаја за биоинформатичку заједницу. Да би се овај успех остварио неопходно је на неки начин представити, интегрисати и претражити податке који имају тенденцију раста. Како технологије семантичког веба нуде могућност решавања наведених проблема, тако су многе организације и институције усвојиле семантичке стандарде и допринеле својим конструктивним решењима. Као последица тога, развијен је велики број апликација које обезбеђују разумљив кориснички интерфејс (за претрагу, визуелизацију и анализу података) и које притом подржавају актуелне базе података. Платформа, која представља главни предмет истраживања ове дисертације, усвојила је стандарде семантичких технологија и својим методама допринела истраживањима у области рационалног дизајна лекова. У наставку је извршен компаративни преглед Платформе са софтверским решењима у сличним биоинформатичким доменима. У Осмом поглављу извршена је и компарација

резултата метода одговарајућих софтверских решења са резултатима основних метода Платформе.

SPARQLGraph [114] је платформа која дозвољава откривање знања у домену биоинформатике тако што је корисницима омогућено да визуелно креирају графове, који се затим конвертују у SPARQL упите и изводе над јавним *endpoint*-ом. Развој графова се заснива на *step-by-step* приступу, у смислу да корисник креира граф селекцијом класа и својстава одговарајућих база. SPARQLGraph подржава неколико база података из EMBL-EBI (Atlas, Chembl, Reactome и Uniprot) и Bio2RDF (Entrez Gene, DrugBank, KEGG, PharmGKB) репозиторијума. Корисницима је омогућено да изводе упите над једном или више база података. SPARQLGraph нуди и неколико предефинисаних шаблонских графова (упита): *Search for a compound in Chembl, Find all associated documents in a Chembl to a compound; Find a specific protein in the database; Find all proteins which are annotated with a specific disease*. Ови графови представљају одличну почетну тачку за откривање релевантних информација и генерално се преклапају са доменима предефинисаних упита на Платформи. SPARQLGraph омогућава корисницима да на основу шаблонских графова креирају нове графове (упите) и поделе их са другим истраживачима. Ово са једне стране представља предност у односу на Платформу, али може се сматрати и озбиљним недостатком, јер је изградња упита помоћу визуелних приступа погоднија када се креирају једноставни упити, али комплекснији упити могу иницијално обесхрабрити кориснике. У таквим околностима визуелизација је ограничена одређеним командама, а кориснички интерфејс може бити преокупирати елементима. У таквим ситуацијама корисници могу закључити да је лакше овладати SPARQL синтаксом, него уложити време у визуелну конструкцију упита. За разлику од SPARQLGraph система, Платформа представљена у дисертацији има за циљ да корисницима најпре укаже на једноставност приступа, тако што им обезбеђује једноставан кориснички интерфејс, а значај иницијалних резултата би требало да их заинтригира да примене и друге методе. SPARQLGraph не омогућава додавање нових база података, односно ажурирање иницијалних графова (упита), што је супротно од приступа који нуди Платформа. Обе платформе представљају резултат у табеларној форми и нуде добар потенцијал за откривање знања.

Софтверско решење QueryMed [110] послужило је као основа за развој Платформе представљене у дисертацији. Циљ QueryMed приступа јесте омогућити корисницима, који немају довољно искуства са SPARQL синтаксом, лак и интуитиван приступ за откривање знања у домену биоинформатике. QueryMed нуди могућност претраге по кључним речима. На пример, корисник може унети назив одговарајућег лека или болести (*key_word*), а у позадини се извршава једноставан SPARQL упит, који преузима све триплете облика $?s\,?p\,?o$, где је избор променљиве $?o$ условљен изразом облика *FILTER regex(?return_value,"key_word","i")*. Овај иницијални приступ се разликује од Платформе, јер су предефинисани упити на Платформи доменски специјализованији, односно преузимају специфичније податке из база. Корисник QueryMed платформе може утврдити над којим скуповима података је упит извршен, с тим што није наглашено да ли постоје базе над којима није откривен резултат. На Платформи је ова опција подржана статистичким приказом резултата, који укључује и базе података које немају повратних вредности. Ово свакако може бити од значаја јер дате базе могу бити потенцијално погодне за откривање неких других типова података. QueryMed омогућава и додавање кориснички одабраних скупова података, што је подржано и на Платформи, али без могућности да овај скуп података буде доступан другим истраживачима. На QueryMed платформи овај корак не захтева познавање SPARQL синтаксе, што није случај на Платформи с обзиром на специфичност предефинисаних упита. Такође, кориснику QueryMed платформе је омогућено да на основу RDF структуре одговарајуће базе података изврши селекцију предиката и спроведе додатне упите који му могу помоћи у откривању релевантнијих информација. У овом кораку се од корисника очекује познавање SPARQL синтаксе. Сличан приступ омогућен је и на Платформи, али без потребе за SPARQL рутином. Обе платформе представљају резултат у табеларној форми.

Платформа GFBio представљена у раду [115] омогућава семантичку претрагу над преко 4,6 милиона скупова података, који су део немачке федерације биолошких података - GFBio (*The German Federation for Biological Data*) специјализованих за нуклеотидне и податке биолошке средине (PANGAEA), као и податке за природне науке (BGBM). Претрага података на GFBio платформи подразумева претрагу по кључној речи, што је слично претрази која се одвија на Платформи. Међутим, разлика је у томе што

Платформа нуди предефинисане упите, а GFBio као семантичку методу претраге користи експанзију упита. Наиме, GFBio за задату кључну реч проналази синониме користећи *Terminology Service* компоненту, која садржи речнике и онтологије које се односе на биолошке податке, а затим их користи за експанзију упита. На пример, за кључну реч *t=Apis mellifera* (пчелињи мед) компонента *Terminology Service* враћа скуп синонима облика [Honey bee, Honeybee, European Honey bee,Western Honey bee, Bee] - $S_t = \{s_1, s_2, \dots, s_n\}$. Свако s у S_t се додаје кључној речи t помоћу логичког оператора OR (*Honey bee OR Honeybee OR European Honey bee OR Western Honey bee OR Bee*). Дати израз се затим конвертује у упит који је приближно облика *?returnvalue :label Apis mellifera; :synonims [Honey bee, Honeybee, European Honey bee,Western Honey bee, Bee]*. Резултат извршавања упита јесу подаци који се преузимају из NCBItaxon[87] онтологије, која описује класификацију и номенклатуру живих врста. У овом случају резултат претраге јесу публикације или инстанце дате базе података. Овакав вид извршавања упита је добар за уопштену претрагу, али у случају да корисник жели опипљивије информације, морао би накнадно да врши анализу добијених резултата. Платформа GFBio не нуди опцију динамичког филтрирања резултата, која би могла допринети откривању релевантнијих информација, као што је то случај на Платформи.

BioSearch [111] је платформа имплементирана за претрагу Bio2RDF репозиторијума, тачније његових база података: Drugbank, InterPro, KEGG, MeSH, NCBI Gene, OMIM, Orphanet и PharmGKD. Платформа BioSearch има за циљ да побољша конструкцију Federated SPARQL упита за кориснике који немају искуства са SPARQL синтаксом, али и да побољша начин организовања резултата претраге. Ово софтверско решење има поједностављени кориснички интерфејс, који омогућава претрагу база на основу кључне речи, као што је то случај на Платформи, а резултат се даље филтрира на основу класе (*C:name_of_class*), својстава (*P:name_of_property*) и имена базе података (*S:name_of_database*). Сви ови елементи су доступни на интерфејсу и приступа им се *step-by-step* селекцијом, што заправо подразумева да корисник има одређено доменско знање. На пример, могуће је извршити претрагу лека *Fluorouracil* над Drugbank базом података (*S:drugbank*), а потом узети у обзир својства *P:drugbank calculated-properties* и *P:drugbank target*. BioSearch платформа користи *AND* логички оператор да комбинује кључне речи које припадају претходно наведеним нивоима претраге, док се логички оператор *OR* користи за комбиновање кључних речи из исте категорије. Семантика упита *Fluorouracil S:drugbank P:drugbank calculated-properties P:drugbank target* је облика (*Fluorouracil*) AND *S:drugbank* AND (*P:drugbank* OR *calculated-properties P:drugbank target)*. Коришћење логичког оператора *OR* може проузроковати много резултата које садрже само једну кључну реч. Ово је рецимо супротно од приступа који нуди Платформа, јер сваки подупит у предефинисаном упиту користи одговарајућу кључну реч и тежи да условно речно, комбинацијом AND и OR оператора преузме адекватни резултат. Резултат претходно дефинисаног упита на BioSearch платформи јесу све инстанце (лекови) Drugbank базе података, које су на неки начин повезане са кључном речи *Fluorouracil*. Претрагом појединачних инстанци, откривају се биолошки системи који су повезани са датим лековима. На сличан начин се извршава претрага и KEGG базе података. Обе платформе заправо производе идентичне резултате у случају коришћена Bio2RDF репозиторијума, само што је приступ на Платформи једноставнији јер се од корисника не захтева доменско знање за извршавање предефинисаних упита.

BioCarian [116] је платформа која омогућава претрагу база података које се баве DNK проблематиком - dbSNP, GWAS, Ensembl, UniProt, KEGG и Reactome базе података, као и база које чувају податке о болестима - OMIM и DisGeNET. Претрага података је омогућена на основу кључне речи (као и на Платформи), на основу SPARQL упита или на основу претраге засноване на аспектима (енгл. *facet based searches*). Први тип претраге подразумева креирање одговарајућих SPARQL упита, који се извршавају на одређеном *endpoint*-у. На пример, корисник може вршити претрагу по кључној речи *Thymidylate synthase* и у том случају се креира SPARQL упит који садржи образац облика *(?subject ?score) text:query ('Thymidylate synthase' 100)*. Дуги тип претраге је прилагођен корисницима који располажу искуством са SPARQL синтаксом. У овом случају они користе едитор који омогућава директни унос упита. На овај начин корисницима је дозвољено да креирају и Federated SPARQL упите. Последњи тип претраге може

[87] *http://www.obofoundry.org/ontology/ncbitaxon.html*

комбиновати претходна два типа. Резултати претраге су груписани по базама података и додатно се могу филтрирати на основу аспеката (енгл. *facets*). Сваки аспект представља скуп одговарајућих ентитета, који су доступни кроз дате базе података. На пример за дати биолошки систем, могу се открити ентитети попут *Gene Symbols*, *Comments* итд. Слични подаци се добијају и на Платформи, користећи опцију динамичког филтрирања резултата. BioCarian не дозвољава унос кориснички селектоване базе података, што је супротно приступу на Платформи.

BioQueries [117] платформа је дизајнирана са циљем да пружи подршку истраживањима у домену биоинформатике и биомедицине. BioQueries тренутно нуди преко 400 упита груписаних по категоријама и базама податка (и овај број упита константно расте). Регистровани корисници имају опцију да креирају упите и да их деле са другим истраживачима. Претрага упита се може вршити по кључној речи или по селектованој бази. Упити BioQueries система искоришћени су да би се додатно проверили обрасци који су делови предефинисаних упита на Платформи. Конкретно, за кључну реч *target* могу се открити упити типа *Get ChEMBL targets*; *Get ChEMBL activities, assays and targets for the drug* који се изводе над Chembl/EMBL-EBI или упит *All targets and their functions stored in Kegg database* који се изводи над Kegg/Bio2RDF бази података. За кључну реч *assay* откривен је упит *Display information on assays according to an introduced term from ChEMBL* који се изводи над Chembl/EMBL-EBI базом података. Претрага по кључној речи *compound* издвојила је упите типа *All ChEBI compounds stored in Bio2RDF Atlas* и *Samples treated with a compound of or a more specific* који се изводе над Atlas/Bio2RDF базом података и упит *Pahtways related with a given compound* који се изводи над Kegg/Bio2RDF базом података итд. Примери ових упита су мање-више идентични иницијалним упитима представљеним у одељку 6.1.2.1, а који су искоришћени за креирање предефинисаних упита на Платформи. Примери упита који се користе за претрагу публикација изводе се над Atlas/Bio2RDF и WikiPathways/EMBL-EBI базом података и они се разликују од упита овог типа на Платформи. Резултати упита на BioQueries платформи могу поред табеларне форме, као и на Платформи, бити и у форми графа.

Open PHACTS [118] платформа је развијена са циљем да интегрише и обезбеди приступ фармаколошким базама података, користећи семантичке веб стандарде и технологије. Интеграција различитих база података изведена је користећи VoID речник [104] и нанопубликације. Open PHACTS нуди различите алате за претрагу и анализу података, а један од значајнијих је Open PHACTS Explorer. Он омогућава интуитивни начин претраге свих фармаколошких и физичко-хемијских извора података, као што су Chebi, Chembl/EMBL-EBI, SureChEMBL/EMBL-EBI, UniProt/EMBL-EBI ChemSpider, ConceptWiki, DisGeNET, Drugbank, WikiPathways, ENZYME, FDA Adverse Events (FAERS), Gene Ontology, Gene Ontology Annotations, neXTProt и WikiPathways, који су интегрисани на Open PHACTS платформи. Open PHACTS Explorer је направљен на основу специфичних потреба истраживача, дизајнираних да помогну у одговору на критична фармаколошка питања, као што су: *For a given compound, give me the interaction profile with targets; Give me all oxidoreductase inhibitors active < 100 nM in human and mouse; For a target, give me all active compounds with the relevant assay data. For this target which active compounds have been reported in the literature?*. Примери ових упита јасан су показатељ колике су ове теме актуелне у домену биоинформатике, што је подједнако важно и за Платформу. Претрага података на Open PHACTS платформи врши се задавањем кључне речи, као и на Платформи, а затим се може вршити филтрирање по једињењу или биолошким системима. Даље, резултати се могу филтрирати на основу RDF структуре, тачније одговарајућих својстава. Поред оваквог приступа могу се користити и REST позиви. Open PHACTS извршава упите над базама података које се чувају локално. На овај начин се убрзава време извођења упита, али се повећава ризик да резултати нису *up-to-date*, што је супротно од приступа које нуди Платформа. Иначе, Open PHACTS пројекат је развијен 2012-те године, а интензивно се развија и данас и има значајну улогу у домену многих биоинформатичких истраживања. У периоду између априла 2013-те и марта 2014-те године ова платформа бележи преко 13,5 милиона посета што говори о њеној популарности.

Постојање и константни развој великог броја апликација које користе технологије семантичког веба, експлицитан је показатељ да биоинформатичка заједница користи његове предности и доприноси квалитетнијем истраживачком раду. Платформа представљена у дисертацији претендује да искористи

потенцијал технологија семантичког веба и да својим методама у потпуности парира актуелним софтверским решењима. На основу прегледа литературе може се закључити да су базе података, које су иницијално подржане на Платформи, прилично актуелне у савременом биоинформатичком истраживању. Платформа обезбеђује једноставан начин претраге биоинформатичких података у циљу откривања знања. Иако тренутно пружа подршку истраживању у домену рационалног дизајна лекова, Платформа има потпуне услове да своје деловање прошири и на неку другу истраживачку област. На тај начин би допринос Платформе у домену истраживања био подигнут на једну вишу лествицу.

Иако су биоинформатичка истраживања применом семантичких технологија постала квалитетнија, ипак постоји довољно простора за побољшање и решавање потенцијалних проблема. Један од проблема који је често присутан међу онтолошким базама података јесте проблем сличних (идентичних) података [119]. Технологије семантичког веба, у комбинацији са одређеним информатичким приступима и математичким алатима, помажу у решавању овог проблема. У наредном поглављу је детаљније дискутовано о наведеном проблему, као и предложеном начину његовог решавања на Платформи.

7 Метода детекције сличних података на Платформи

У овом поглављу је образложен проблем детекције сличних података у домену биоинформатике, као и његов утицај на планирање будућих експеримената. У складу са тим је извршен преглед актуелне литературе и техника онтолошког поравнања, као једног од најчешћих приступа за откривање сличних података у домену биоинформатике. У овом сегменту је представљена и конекција појмова онтолошког поравнања и рударења текстуалних података. С тим у вези је термин рударења текстуалних података посебно разматран. Нарочита пажња је усмерена на разграничавање појмова семантичке сличности и семантичке повезаности података, који су у тесној вези са рударењем текстуалних података. Централни део поглавља је усмерен на репрезентацију алгоритама за детекцију сличних података који је имплементиран на Платформи, а који се генерално заснива на примени екстензијских техника онтолошког поравнања, методама рударења текстуалних података, конвертовању текстуалних података у векторске величине и примени мере косинусе сличности у циљу коначне детекције сличности. У овом сегменту су прво представљени циљеви алгоритма, а затим детаљно образложене компоненте модела које спроводе одговарајуће кораке алгоритма. Представљени су и одговарајући математички алати (модел векторског простора и мера косинусе сличности) коришћени за имплементацију алгоритма.

7.1 Детекција сличних података у домену биоинформатике

Откривање потенцијално сличних података у домену биоинформатике, има значајан утицај на планирање будућих експеримента. Ако се утврди да биолошка мета у интеракцији са одређеном супстанцом (леком) припада групи експеримената са позитивним учинком, онда егзистира велика вероватноћа да ће исход експеримента бити позитиван и за неку другу сличну биолошку мету. Такође, ако се утврди да једињење из неке групе комплекса учествује у успешно спроведеним експериментима, онда постоји могућност да ће неко друго слично једињење из исте групе комплекса, поновити успех приликом реализације експеримента. Ови принципи се заснивају на запажању да сличне хемикалије често деле сличне физичко-хемијске особине и биолошке активности [120]. Дата хипотеза може бити јако корисна када нека лабораторија синтетише одговарајући лек и када се утврди да постоји сличност са неким другим леком, или када се намерава коришћење сличне биолошке мете или ћелијске линије. На тај начин се могу уштедети ресурси и избећи експериментисање са структурама, који могу довести до потенцијално сличних резултата. Због важности овог принципа у домену биоинформатике, предложене су различите методе сличности како би се прецизно одредила сличност између података.

Многе методе у домену биоинформатике фокусирају се на мерење структурних сличности између хемијских структура [121]. Међу традиционалним техникама за утврђивање сличности истиче се метода подструктуре и надструктуре (енгл. *substructure and superstructure relationships*). Ова метода спада међу најчешће коришћеним мерама сличности и може се дефинисати на следећи начин: „*С обзиром на две хемијске структуре А и Б, ако је структура А у потпуности садржана у структури Б, онда је А подструктура од Б, а Б је надструктура од А. Према овоме, структуре А и Б могу делити својства која су у вези са њиховом заједничком подструктуром. Због тога се подструктура која је повезана са одређеним својствима од интереса може користити као модел у бази података за идентификацију свих једињења која деле ову подструктуру (или надструктуру) и евентуално њене активности*“ [121]. Међутим, ова традиционална мера сличности је често нефлексибилна, јер заступа приступ који је утемељен на знању, што подразумева да свака подструктура, која се користи као модел, мора бити добро дефинисана, јер би у супротном резултати претраге били лошег квалитета и врло ограничене предиктивне вредности [121]. Додатно, ова мера сличности не подразумева никакву квантитативну вредност, што отежава да се резултати претраге рангирају на смислен начин [121]. Неретко се структурна сличност рачуна и помоћу методе „отисака прстију“ (енгл. *fingerprints*) [122]. У овом контексту „отисак прста“ је бит стринг (састављен од нула и јединица), где сваки бит представља присуство или одсуство

дате функције или подструктуре, па се две структуре могу упоређивати на основу броја заједничких битова применом одговарајућих коефицијента (нпр. применом *Jaccard-Tanimoto* коефицијента) [122]. Такође, било је и покушаја да се користе алгоритми упоређивања графова, који се примењују над хемијском структуром одговарајућих молекула [122]. У овом случају, одређивање сличности се сводило на тражење максималног заједничког подграфа [123]. Иако је у пракси употреба ових метода једноставна, оне су ипак лимитиране извесним аномалијама. Пре свега, оне нису у могућности да идентификују локалне сличности између структура (посебно не оних са великим разликама у величини), а поред тога њихова вредност се значајно смањује када је у питању идентификација слабијих сличности [121].

Неке од метода за утврђивање сличности заснивају се на упоређивању физичко-хемијских особина структура. Методе засноване на структурним дескрипторима (енгл. *structural descriptor-based methods*) такође су популарне за утврђивање структурних сличности [121]. „*Структурни дескриптори представљају хемијске структуре на начин да се њихова сличност може лако квантификовати. Ово се обично постиже израчунавањем структурних поткомпоненти применом разних метода*“ [121]. У суштини, методе засноване на структурним дескрипторима представљају хемијску структуру као вектор у димензионалном простору, а генерисани коефицијенти сличности ових приступа пружају математички модел за процену структурних сличности [121]. На пример, Doniger и др. [124] представљају једињење као 9-димензионални вектор, при чему сваки елемент вектора представља неко физичко-хемијско својство молекула (молекуларна маса, запремина, укупна површина итд.). Они користе и вештачку неуронску мрежу (енгл. *artificial neural network*) и методу потпорних вектора (енгл. *support vector machine*) да би идентификовали слична једињења, која могу проћи кроз одговарајуће крвно-мождане баријере. Методе засноване на структурним дескрипторима свакако су флексибилније од претходно наведене методе подструктуре и надструктуре, јер користе потпуну структуру или подструктуру као модел за идентификацију структура у бази података које су глобално „сличне“ структури модела [121].

Међутим, са трендом пораста података у биолошким и хемијским базамам, паралелно су расли трошкови складиштења, као и трошкови утврђивања сличности између података. Проналажење сличности између протеина изолованих из различитих организама или упоређивање секвенци генома [125], неки су од примера биоинформатичких проблема који су захтевали поређење великог броја података и откривање софистицираних информација. Како претходно представљене традиционалне методе нису биле довољно агилне за систематске анализе осетно великих података, тако је настала потреба за делотворнијим методама. Технологија рударења података (енгл. *data mining*) [126] такође има обећавајућу улогу у процесу детекције сличних података. Ова технологија „*повезује традиционалне методе анализе података са алгоритмима за обраду великих количина података*“ [125]. У склопу рударења података примењују се различити алгоритми како би се постигле прописане активности, а једна од њих је и кластеровање. Технике кластеровања често се користе за идентификацију групе сличних једињења [127], а популарне су методе кластеровања гена [128]: хијерархијско кластеровање, кластеризација методом К-средњих вредности (енгл. *K*-means) и К-средине (енгл. *K*-medoids). Техника кластеровања подразумева груписање објекта у неки кластер у зависности од више атрибута. Додатно, за сваки кластер може бити прописан други скуп правила за груписање. Процес кластеровања подразумева да се слични објекти налазе унутар једног кластера. Међутим, примена кластеровања у пракси подразумева појаву многих проблема, а најважнији је тај што нема сазнања да ли је проблем кластеровања успешно решен, док се додатно често пронађе и већи број одговора на дати проблем [125]. Други важан проблем је и тумачење семантике (значења) сваког кластера, што може бити оптерећујуће јер тачно значење сваког кластера не мора бити очигледно, па је често неопходно потражити помоћ експерата [125].

Претходно наведене методе представљају битне рачунарске приступе за идентификацију сличних структура. Међутим, већина алгоритама развијених за потребе ових метода ограничена је брзином, скалабилношћу и тачношћу резултата, и не може се једноставно бавити великим количинама података који се временом рапидно гомилају. Због тога истраживачи у домену биоинформатике све више напора улажу у развој ефикаснијих метода, које омогућавају софистицираније резултате над биоинформатичким

подацима. С обзиром да је семантички веб достигао значајан степен популарности у домену биоинформатике, многа истраживања су се фокусирала на употребу његових технологија у циљу решавања проблема детекције сличних података. У наставку је овај приступ детаљније образложен.

7.1.1 Детекција сличних података над онтолошкима базама података

Модели засновани на онтолошким језицима подржавају приступ структурираним подацима, који помажу при препорукама заснованим на садржају, филтрирању релевантних информација, семантичном обогаћивању или прилагођавању веб страница одређеним корисничким захтевима, али се такође користе и за откривање веза или сличности између ентитета [129]. Како актуелни подаци у сфери биоинформатике подржавају форму структурираних података (чувајући масивне количине информација дистрибуиране кроз различите репозиторијуме), тако је и над њима могуће спровести одређене методе у циљу детекције сличности података. Већина ових метода, које се називају методама или техникама онтолошког поравнања, заправо се заснива на поређењу онтолошких елемената (ентитета) [99]. Одељак 6.1.1 уводи дефиницију онтолошког поравнања и говори о неколико основних типова поравнања. Од важности за биоинформатичке податке је несумњиво поравнање на нивоу инстанци, јер су онтолошке базе података богате таквим типовима ентитета, који најаутентичније моделују податке о одређеним биолошким метама, лековима, ћелијским линијама, есејима и итд. У наставку је дата дефиниција поравнања инстанци.

Дефиниција 2: За две улазне онтологије O_1 и O_2, **поравнање инстанци** се дефинише као процес поређења инстанци $i_1 \in O_1$ и $i_2 \in O_2$ у циљу одређивања излазног параметра - нумеричке вредности која представља меру сличности између инстанци i_1 и i_2 [130].

Генерално, поређење инстанци је поступак утврђивања степена сличности између парова инстанци који припадају различитим онтолошким базама података. Што је већа сличност између инстанци, већа је вероватноћа да оне представљају идентичне ентитете [130]. Процес поређења инстанци подразумева два основна корака: (1) скенирање базе података у циљу детекције ентитета за поређење (најчешће објектних вредности), (2) мерење сличности између парова ентитета утврђених у кораку (1) [129]. Посматрајмо инстанце *drugbank:DB00544* (*Fluorouracil*) и *chembl_molecule:CHEMBL1386* (*Transplatin*). Ове инстанце (лекови) припадају DrugBank/Bio2RDF и Chembl/EMBL-EBI базама података, респективно. Иницијални корак у процесу њиховог поређења био би прилично захтеван имајући у виду да прва инстанца има 203 објектне вредности (триплета), а друга инстанца 586 објектних вредности (триплета). Укупан број парова ентитета за поређење у овом случају износио би 203 × 586. Због тога је неопходно направити адекватну селекцију предиката [129] и то је први изазов у процесу поређења инстанци. Многа истраживања [131,132] користе само уграђене предикате, као што су *rdfs:label* или *rdfs:comment,* за процес поравнања. Ипак, процес селекције предиката се мора извршити високом прецизношћу, јер је за мање скупове података пожељно поредити што више ентитета како би резултат био прецизнији, иако са друге стране превелики број предиката може повећати комплексност самог процеса [129]. У другом кораку поређења инстанци виталан проблем је одабир метрике за одређивање сличности јер је неопходно познавати семантику ентитета који се пореде. Одређивање прага сличности (енгл. *threshold*), који се користи за одбацивање парова ентитета који покажу слабију сличност [129] такође је један од знчајних проблема.

Непосредан проблем онтолошких база је и у томе што оне обично садрже дупликате, односно идентичне ентитете што је познато и као проблем повезивања података (енгл. *data linking problem*) [133]. На пример, инстанца *drugbank:DB00554* је преко објектног својства *drugbank_vocabulary:x-kegg* повезана са инстанцама *kegg:C07649* и *kegg:D00584*, који садрже податке о истом леку, па су ови ентитети идентични. Поред кориснички дефинисаних својстава за обележавање идентичних ентитета, често се користи и уграђени предикат *owl:SameAs* [129]. Ипак, истраживања у [134] су показала да се овај атрибут понекад користи погрешно (намерно или случајно) за повезивање веома сличних, али не и идентичних ентитета. Може се рећи да је ова синтакса повремено злоупотребљена у LOD контексту јер често повезује инстанце са различитим степенима грануларности [93]. Додатно, предикати из SKOS [135] речника (*skos:closeMatch* и *skos:relatedMatch*) често се користе за представљање веома сличних ентитета.

Такође, у онтологијама се јавља и проблем онтолошке разноврсности, што може проузроковати да идентични ентитети нису чак ни повезани [129]. Такође, може се десити и да у оквиру једне онтолошке базе података постоји група сличних ентитета. На пример, Chembl/EMBL-EBI база података поседује групе сличних протеина или ћелијских линија.

Дакле, проблем проналажења сличних података (инстанци) између онтолошких база података које неретко имају неколико милиона триплета и ентитета (подсећања ради DrugBank/Bio2RDF база података има 3.672.531 триплета и 316.950 ентитета[88]) изазован је из разлога што је потребно међусобно упоређивати милионе ентитета, а то заузврат захтева добар механизам селекције оних који су релевантни за процес упоређивања [93]. Затим, потребно је проверити многе параметре на нивоу ентитета, односно открити семантику ентитета, пре него што је могуће применити одговарајућу метрику за рачунање сличности. Додатно, многи ентитети се подударају или су врло слични, што одражава потребу за што прецизнијим прагом сличности. Са циљем превазилажења наведених проблема за потребе овог истраживања развијен је алгоритам за детекцију сличних података, који је детаљно образложен у одељку 7.5. Алгоритам се генерално заснива на примени одговарајућих техника онтолошког поравнања, примени метода рударења текстуалних података и конвертовања текстуалних података у векторске величине. У складу са тим у наставку је извршен преглед основних техника онтолошког поравнања и указано је на њихову повезаност са термином рударења текстуалних података. Посебна пажња је усмерена на разумевање термина семантичке сличности и семантичке повезаности података, који су такође у тесној вези са рударењем текстуалних података.

7.2 Онтолошко поравнање

Процес упоређивања и утврђивања сличности између елемената онтологија (ентитета) зове се онтолошко поравнање [99]. С обзиром да онтологија моделује концептуализацију одређеног домена, упоређивање ентитета између онтологија практично означава више од репрезентације тих ентитета на нивоу синтаксе [100]. Осим тога, битно је имати у виду њихову семантику, односно значење. Да би се реализовало такво упоређивање ентитета, користи се модус према којем се сличност може посматрати на три нивоа: на нивоу контекста (прагматичном нивоу), на нивоу онтологије (семантичком нивоу) и на нивоу података (симболичком нивоу) [100].

На нивоу контекста „*разматра се начин на који се ентитети користе у спољном контексту. Тај процес подразумева коришћење информација које се налазе изван онтологија. Контекст се посматра само као локални модел који кодира корисников субјективни поглед на домен. Са гледишта одређивања сличности најважнији је контекст примене. Поједностављено, слични елементи имају сличне узорке коришћења, тј. слични елементи се користе у сличном контексту. Код откривања сличности то се правило разматра у оба смера: уколико се два ентитета користе у истом (сличном) контексту, онда су ти елементи слични и обрнуто, уколико се у два контекста користе исти (слични) елементи, онда су та два контекста слична*“ [100].

На нивоу онтологије „*посматрају се семантички односи међу елементима онтологије. На најнижем нивоу, онтологије се посматрају само као графови са концептима и односима. Тај ниво се надограђује дескриптивним логикама, ограничењима или правилима на највишем нивоу. Међутим, како правила у семантичком вебу још нису генерално довољно истражена, овај ниво упоређивања ентитета је слабије подржан*“ [100].

На нивоу података елементи се упоређују узимајући у обзир само вредности одређених типова података, као што текстуални подаци (стрингови). Текстуални подаци су повезани са онтологијама као главним ризницама за његово формално представљање, јер су онтологије концептуални модели који имају за циљ да подрже и недвосмислено размене знања и пруже оквир за његову интеграцију [136]. Онтологија повезује текстуалне податке са дефинисаним концептима (инстанцама или класама) користећи како стандардизоване, тако и кориснички дефинисане предикате. Рецимо, предикати који често повезују

[88] *http://download.bio2rdf.org/files/release/3/drugbank/drugbank.html*

концепте са текстуалним подацима су *rdfs:label* и *rdfs:comment*. Слично важи и за предикате *dc:title* и *dc:description* који припадају DC [137] речнику. На пример, вредност предиката *rdfs:label* класе *drugbank_vocabulary:Drug* је *Drug [drugbank_vocabulary:Drug]*. Вредности предиката *dc:title* и *dc:description* инстанце *drugbank:DB00544* су *Fluorouracil* и *A pyrimidine analog that is an antineoplastic antimetabolite. It interferes with DNA synthesis by blocking the thymidylate synthetase conversion of deoxyuridylic acid to thymidylic acid. [PubChem]*. Поред предиката из RDFS и DC речника, који су генерално заступљени у сваком биоинформатичком репозиторијуму, онтологије често моделују и корисничке предикате за представљање стринг вредности. Табела 7.1 садржи само неке од примера стандардизованих и кориснички дефинисаних предиката актуелних биоинформатичких репозиторијума (укључујући и CPCTAS) који су типа података (*datatype property*). Кодомени ових предиката су стринг вредности (*xsd:string*).

Табела 7.1 Примери стандардизованих и кориснички дефинисаних предиката Bio2RDF, Chem2Bio2RDF, EMBL-EBI и CPCTAS репозиторијума који су типа података (*datatype property*) и чији су кодомени стринг подаци (*xsd:string*)

Репозиторијум	**Предикати типа података (*datatype property*)**	
	Стандардизовани	Кориснички
Bio2RDF	*rdfs:label* *dc:tittle* *dc:identifier* *bio2rdf_vocabulary:identifier* *bio2rdf_vocabulary:namespace*	*drugbank_vocabulary:gene-name* *drugbank_vocabulary:drugbank-id* *drugbank_vocabulary:transmembrane-regions* *drugbank_vocabulary:synonym* *kegg_vocabulary:formula* *kegg_vocabulary:mol_weight*
Chem2Bio2RDF	*rdfs:label*	*pubchem:synonyms* *bindingdb:Name* *bindingdb:Title*
EMBL-EBI	*rdfs:label* *dc:description* *skos:altLabel* *skos:prefLabel*	*chembl:targetType* *chembl:componentType* *chembl:organismName* *chembl:targetConfDesc*
CPCTAS	*dc:title*	*pibas:name* *pibas:hasTargetName* *pibas:targetType* *pibas:result* *pibas:theAimOfExperiment* *pibas:storingConditionOfActiveSubstance*

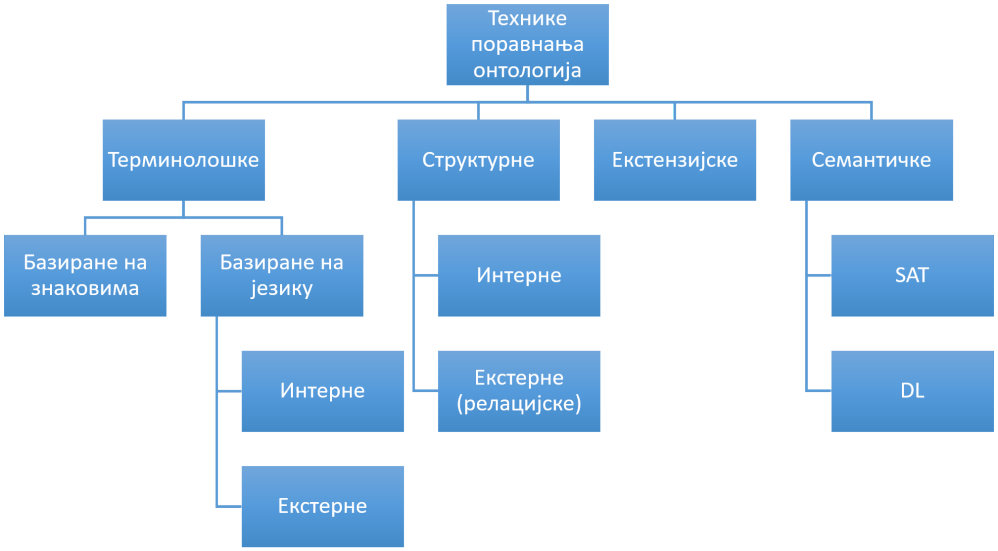
Мерење сличности између текстуалних података даље се може вршити неком од техника онтолошког поравнања представљених у наредном одељку.

7.2.1 Технике онтолошког поравнања

Технике[89] поравнања онтологија дефинишу рачунање сличности између онтолошких елемената. Аутори радова [138,139,140] приказују подробан преглед различитих приступа и класификација техника поравнавања онтологија. Ehrig у [138] тумачи актуелне приступе у теорији поравнања и представља две класификације дефинисане према ортогоналним димензијама: хоризонтална димензија (која се односи на ниво података, ниво онтологије и ниво контекста) и вертикална димензија (представља знање о домену које се може поставити на било коју хоризонталну димензију). Shvaiko и др. [139] предлажу класификацију која се темељи на врсти улазних података и она се цитира у већини извора који се баве овом тематиком [140]. Према датој класификацији, основни скупови техника за поравнање који се базирају на сличности њихових елемената јесу терминолошке технике или технике које се темеље на низовима (стринговима), структурне, екстензијске и семантичке технике (Слика 7.1). У наставку је дат детаљнији опис терминолошких и екстензијских техника[90]. Такође, наглашен је и значај комбиновања техника и људских ресурса у процесу онтолошког поравнања.

[89] У наставку се не прави разлика између термина *техника* и *мера*.

[90] Остале технике нису од превеликог интереса за ово истраживање.



Слика 7.1 Шема класификације техника поравнања на основу врсте улазних података *[140]*

7.2.1.1 Терминолошке технике

Терминолошке технике се користе за упоређивање знаковних низова, а могу се применити на кодомене предиката (објектне вредности типа стринг) како би се детектовали слични елементи. Оне се могу користити за упоређивање како инстанци, тако као и других елемената онтологије [100]. Диференцирају се две врсте техника: технике базиране на знаковним низовима и технике базиране на језику. Друга група техника дели се на интерне и екстерне. Технике које се базирају на знаковним низовима посматрају низ као секвенцу знакова (слова) [100]. Ове технике се могу поделити у неколико група [141]:

1. Технике базиране на карактерима (енгл. *caracter-based*)
   - **Левенштајнова удаљеност (*Levenshtein distance*)** [142] израчунава минимални број знаковних операција (уметања, брисања и замене знакова) за трансформацију једног стринга у други. (**LD**)
   - **Јарова мера (*Jaro measure*)** [143] проналази речи са правописном грешком. (**JaroM**)
   - Техника ***Needleman-Wunsch*** [144] примењује глобално упоређивање стрингова. Ова техника је изузетно погодна када два стринга имају сличну дужину и заједнички степен сличности. Такође, уз помоћ ње могуће је одредити вероватноћу да се две секвенце развијају из истих стрингова [135]. (**NW**)
   - Техника ***Smith-Waterman*** [145] (верзија методе Needleman-Wunsch) одређује сличне регије између две секвенце стринга. Уместо да посматра укупну секвенцу, овај алгоритам упоређује сегменте свих могућих дужина и оптимизује меру сличности. (**SW**)
   - ***UnsmoothedJS*** [146] је тип Jensen-Shannon технике за два различита језичка модела. Ово је популарна техника мерења сличности између две (или више) расподеле вероватноће. (**NW**)
   - **Техника најдужег подстринга (*Longest common substring*)** [147] рачуна сличност између стрингова на основу њиховог заједничког најдужег подстринга. (**LCS**)
2. Q-grams технике [148]
   - ***Bigrams (2-grams)*** техника пореди два стринга и рачуна број заједничких *n-gram*-ова између њих. Један *2-gram* представља секвенцу стринга (подстринг) која има 2 карактера. На пример, *2-grams* секвенце речи *fluorouracil* су: *fl*, *uo*, *ro*, *ur*, *ac* и *il*.
   - ***Trigrams (3-grams)*** техника пореди два стринга и рачуна број заједничких *n-gram*-ова између њих. Један *3-gram* представља секвенцу стринга (подстринг) која има 3 карактера. На пример, *3-grams* секвенце речи *fluorouracil* су: *flu*, *oro*, *ura* и *cil*.
3. Технике базиране на токенима (енгл. *token-based*)
   - **Коефицијент коцке (*Dice coefficient*)** [149] дефинисан је као двоструки број заједничких

израза низова који се пореде у односу на укупан број израза у оба стринга. Резултат 1 ове технике означава идентичне векторе, док вредност 0 одговара ортогоналним векторима. (**DC**)

- **Џакардова мера (*Jaccard measure*)** [150] се примењује над векторима $\vec{X}$ и $\vec{Y}$. У овом случају, сваки вектор одговара неком ентитету. Скаларни производ вектора $\vec{X}$ и $\vec{Y}$, и Еуклидова норма сваког вектора користе се за израчунавање мере сличности. (**JM**)
- **Мера косинусне сличности (*Cosine similarity measure*)** [151] се примењује над векторима $\vec{X}$ и $\vec{Y}$. Она користи Еуклидово косинусно правило за одређивање сличности. (**CSM**)

4. Фонетске технике (енгл. *phonetic-based*)
   - ***Soundex*** [152] техника посматра фонетску сличност између стрингова, односно између изговора тих стрингова.
   - ***Metaphone*** [153] техника представља побољшање претходне технике која се примењује у енглеском језику.

Прва група техника у основи има за циљ да израчуна број заједничких карактера два стринга и генерално се користи за утврђивање типографских грешака (нпр. *Fluorouraci**il*** и *Fluorourac**ul***) [141]. Друга и трећа група техника је у стању да маневрише употребом различитих конвенција за описивање података (нпр. *Diamminodichloride, Platinum* и *Platinum, Diamminodichloride*). У овом случају сличност се израчунава анализом заједничких токена (речи). Технике засноване на кратерима и q-grams технике израчунавају сличност на основу секвенце карактера које су појављују у два стринга, док технике засноване на токенима деле стрингове у речи (симболе или токене) користећи као гранични карактер неки интерпункцијски знак или празан простор (енгл. *whitespace*), а затим израчунавају сличност између два скупа токена [141]. Последња група техника утврђује фонетску сличност стрингова, односно сличност између артикулације тих стрингова. Ове технике се користе за верификацију да ли су два стринга истоветна, чак и ако су другачије написани. Ово је честа околност у језицима где не важи правило да једном гласу одговара једно слово. На пример, у енглеском језику термини *Licence*, *License* и *Licensing* представљају сличне стрингове. Табела 7.2 сумира неке од формула за имплементацију терминолошких техника која се користе за истраживања спроведена у дисертацији. Сва мерења су нормализована на скали [0-1].

Табела 7.2 Преглед формула неких од терминолошких техника онтолошког поравнања (извор *[141]*)

Терминолошке технике	**Формула**	**Ознаке**
Левенштајнова удаљеност	$Lev(s,t) = 1 - \frac{M_{\lvert x\rvert,\lvert y\rvert}}{\max(\lvert x\rvert,\lvert y\rvert)}$, $M_{i,j} = \begin{cases} M_{i-1,j-1} \; if \; x[i]=y[i] \\ 1 + \min(M_{i-1,j}, M_{i,j-1}, M_{i-1,j-1}) \end{cases}$	M матрица
Trigrams	$Trigram(s,t) = \frac{trig(s) \cap trig(t)}{average(\lvert trig(s)\rvert \cup \lvert trig(t)\rvert}$	s,t - стрингови
Коефицијент коцке	$Dice(s,t) = \frac{2 \times \lvert s \cap t\rvert}{\lvert s\rvert + \lvert t\rvert}$	s,t - стрингови
Џакардаова мера	$Jac(s,t) = \frac{\lvert s \cap t\rvert}{s \cup t}$	s,t - стрингови
Мера косинусе сличности	$\cos(s,t) = \frac{\sum_{i=1}^{\lvert\Sigma\rvert} s_i t_i}{\sqrt{\sum_{i=1}^{\Sigma}(s_i)^2 \sum_{i=1}^{\Sigma}(t_i)^2}}$	s,t - стрингови

Чест и неизоставан проблем примене датих техника јесте низак ниво тачности. Свака од техника резултује другачијим вредностима приликом примене, па је неретко неопходно комбиновати више техника из једне категорије како би се добили прецизнији резултати. Пре саме примене ових техника често се спроводе одговарајући поступци нормализације, односно прилагођавање низова неком критеријуму: замена свих великих почетних слова малим словима, замена слова са дијакритичким знаковима (знакови различитог облика, нпр. тачке или цртице који се додају неком слову) итд [100].

Методе које се заснивају на језику корисне су ако се употребљавају врло слични низови како би се означили исти концепти [100]. На пример, може се запазити сличност између речи *diabetes* и *sugar in blood*, али не и између речи *sugar* и *glucose*. „*Осим тога, два низа могу бити врло слична, али притом означавати два различита концепта, нпр. hotel и hostel, а исто важи и за све хомониме. Такође, проблем се може јавити и код коришћења речи из различитих језика како би се означио исти елемент. Интерне језичне технике не упоређују само издвојене знаковне низове, односно речи, већ и делове текста који се могу раставити на речи*“ [100], нпр. *Medical treatments*. „*Овакве технике се темеље на процесирању природног језика како би се из текста издвојили термини са значењем. Упоређивање термина и њихових односа помаже у откривању сличности између елемената онтологија који су дефинисани у облику фраза. И код ове технике потребно је спровести поступке нормализације који се односе на три главне врсте варијација: морфолошке (варијације облика речи која произлази из истог корена), синтаксичке (варијације граматичке структуре) и семантичке (варијације у значењу, обично коришћењем надређеног или подређеног појма)*“ [100]. У овом кораку се поређење елемената онтологије поклапа са процесима рударења текстуалних података (енгл. *text-data mining*) [154], па је ова технологија детаљније представљена у одељку 7.3. Екстерне језичне технике често се служе спољним изворима као што су лексикони, вишејезични речници, семантичко-синтаксички лексикони, терминологије итд. [100]. Примери ових речника су UMLS (*Unified Medical Language System*) [155], који дефинише термине у домену медицинског знања и WordNet [156], који представља најчешће коришћену лексичку базу података за енглески језик. Коришћење речника је поуздан приступ јер се ослања на људске верификоване ресурсе, док је мана недостатак прилагодљивости променама у зависности од структуре онтологије [157], као и слаба покривеност одговарајућих домена. Језички извори „*могу бити дефинисани за један језик или специфично обликовани за одређени домен. Помоћу таквих извора могу се лакше решити проблеми с појављивањем синонима. Узимањем у обзир објашњења значења речи садржаног у речнику повећава се могућност налажења правих кандидата за поравнање. Слично је и с хомонимима, хипонимима и хиперонимима*“ [100].

7.2.1.2 Екстензијске технике

Истраживање у [100] указује да екстензијске технике упоређују екстензије елемената, које могу бити креиране од других елемената, нпр. инстанци. Овакав вид поређења је познат као поравнање на нивоу инстанце. Инстанце представљају одличну прилику за упоређивање онтологија.

Дефиниција 23. Претпоставимо да су инстанце $i_1 \in O_1$ и $i_2 \in O_2$ представљене на следећи начин: $i_1 = [v_1, \dots, v_m]$, $i_2 = [v_1, \dots, v_n]$ где су m и n бројеви предиката датих инстанца, а $v_j,\ 1 \leq j \leq m, n$ је вредност j-тог предиката. Циљ екстензијских техника, за две дате инстанце i_1 и i_2 је одређивање сличности $sim(v_h, v_k)$ где $v_h \in i_1$ и $v_k \in i_2$, за сваки пар одговарајућих предиката у i_1 и i_2.

Дакле, ове технике се заснивају на претпоставци да се сличност између инстанци може одредити поређењем вредности њихових предиката (објектних вредности). За сваку специфичну вредност предиката $v_j,\ 1 \leq j \leq m, n$ могу се користити одговарајуће технике поравнања. Како су текстуални подаци најфреквентнији тип представљања података у онтологији, тако је већина ових техника фокусирана на израчунавање сличности између стринг (текстуалних) вредности и у овом случају екстензијске технике се надограђују терминолошким техникама: техникама заснованим на карактерима, техникама заснованим на токенима или фонетски базираним техникама. Кад су израчунате све вредности $sim(v_h, v_k)$ где $v_h \in i_1$ и $v_k \in i_2$, могуће је одредити (с обзиром на праг сличности), да ли су инстанце i_1 и i_2 сличне (или идентичне).

7.2.1.3 Комбиновање техника и улога експерата

Презентоване технике могу се комбиновати на различите начине са циљем добијања коначног резултата. „*Два основна приступа за комбиновање различитих техника јесу секвенцијална и паралелна композиција. Код секвенцијалне композиције различите технике поравнавања изводе се секвенцијално, једна за другом. У том случају има смисла прво применити технике које се заснивају на знаковним низовима, потом технике темељене на структури (и на крају технике темељене на семантици). Други приступ*

подразумева извођење неколико техника паралелно, а потом се бира један од резултата према неком критеријуму (најмања или највећа сличност) или се обједињују све добијене вредности и израчунава коначна сличност према некој дефинисаној формули“ [100].

Битну улогу у поступцима онтолошког поравнања има и корисник - инжењер знања (експерт). Његова присутност може бити заступљена у различитим фазама. Прва интервенција корисника огледа се у одређивању начина не који ће се комбиновати технике поравнања, што подразумева да корисник прилично добро познаје систем и целокупни ток поравнања. „*Други начин интервенције корисника је постављање почетних вредности поравнања, чиме је дефинисано почетно стање на којем ће се темељити даље израчунавање. И треће, корисникова улога је битна на крају поступка када он даје повратну информацију о резултатима*“ [100].

У одељку 7.4 извршен је преглед актуелне литературе наведених техника у разним доменима.

7.2.2 Семантичка сличност и семантичка повезаност података

Велики број проблема у оквиру области природног процесирања језика - NLP (*Natural Language Processing*) [157] зависи од употребе неке мере сличности текстова. Израчунавање сличности између текстова, односно текстуалних вредности (стрингова) у циљу детекције сличности између онтолошких елемената, осим што је класичан истраживачки проблем у сфери природног процесирања језика, кључни је задатак и многих биомедицинских и биоинформатичких апликација [158,159]. Онтологија повезује одговарајуће текстуалне податке са одређеним концептима користећи како уграђене, тако и кориснички дефинисане предикате, што је представљено у одељку 7.1.1. Текстуални подаци заправо морају бити повезани са онтологијама, као главним ризницама за његово формално представљање, јер су онтологије концептуални модели који имају за циљ да подрже и недвосмислено размене знање и пруже оквир за његову интеграцију [55]. Због тога се може рећи да онтологије одражавају потенцијалне интерпретације термина и као такве се могу се користити за подршку аутоматском семантичком тумачењу текстуалних информација [136]. Рачунари не могу једноставно маневрисати семантичком информацијом објекта, у ствари његовим значењем у унапред дефинисаном контексту. Узрок томе је што се значење најчешће описује у природном језику и због тога процес онтолошког поравнања није нимало једноставан. У домену биоинформатике експерти неретко могу бити у дилеми око тачности резултата, због тога што егзистира суштинска разлика између термина семантичке сличности (енгл. *semantic similarity*) и семантичке повезаности (енгл. s*emantic relatedness*). Pedersen и др. [158] дефинишу ове термине на следећи начин „*Семантичка повезаност је општи појам повезаности концепата, док је семантичка сличност посебан случај повезаности који је везан за сличност концепата*“. Дакле, семантичка сличност говори о томе како су два ентитета слична једна другој, а повезаност је шири израз, повезан са било којом врстом семантичког односа између ентитета [160]. На пример, термини *diabetes* и *sugar in blood* су синоними и зато су семантички слични, а самим тим и повезани. Међутим, термини *clostridium perfringens* и *gangrene* имају низак ниво сличности, јер први термин означава бактерију, а други здравствено стање [160]. Симултано, ова два термина су семантички повезана, јер је бактерија *clostridium perfringens* узрочник гангрене [160]. Дакле, речи често не морају бити синоними да би биле семантички сличне и то је управо оно што дефинише семантичка повезаност.

Семантичка сличност и семантичка повезаност су концепти који заправо означавају метрику за одређивање дистанце значења између два појма [161] (докумената, израза, фраза или речи). Ипак између ових термина постоји суштинска разлика. Семантичка сличност укључује само *јесте-је* (енгл. *is-a*) односе између појмова, док семантичка повезаност укључује сваку везу између два појма [159]. Прецизније речено, термин семантичка сличност је метрика дефинисана над скупом неких појмова, где је идеја о удаљености између појмова базирана на сличности између њиховог значења или семантичког садржаја [162]. За разлику од семантичке сличности, термин семантичке повезаности „*означава облик мерења који квантитативно идентификује однос између појмова заснован на сличности или блискости њиховог значења. Формално говорећи семантичка повезаност је дефинисана као облик семантичких или функционалних асоцијација између речи, а не само лексичких односа као што је синонимија или хипонимија. Циљ семантичке повезаности јесте да пажљиво моделује такве асоцијације*“ [163]. Дакле,

две речи су семантички повезане ако претендују да се користе једна поред друге, односно ако постоји вероватноћа да се оне појаве заједно. У том случају важи да оне припадају истој семантичкој категорији. У складу са тим речи *clostridium perfringens* и *gangrene* су семантички повезане.

Основна разлика између семантичке сличности и семантичке повезаности података јесте у начину њиховог израчунавања [161]. Семантичка сличност се мери утврђивањем дистанци између термина коришћењем онтологија и овај приступ се често назива тополошким приступом [162]. У овом случају мера сличности између појмова подразумева утврђивање најкраће удаљености између два термина унутар онтологије користећи *jecme-je* односе између термина [164]. Постоји велики број мера за одређивање семантичке сличности [164]. На пример, Leacock и Chodorow [165] дефинишу сличност између појмова применом обрасца $sim = -\log length\ 2 * D$, где *length* представља дужину најкраћег пута између два концепта коришћењем бројања чворова, а *D* је максимална дубина таксономије. Метода Wu и Palmer [166] мери дубину два одређена појма у WordNet таксономији и дубину најнижег заједничког претка *LCS* (*Lowest Common Subsumer*), а затим комбинацијом вредности израчунава сличност по обрасцу $sim = 2 * depth(LSC)depth(concpet1) + depth(concpet2)$. Ресник [167] је увео меру коју враћа IC (*Information Content*) *LCS* чвора два задата појма - $sim = IC(LCS)$, где је IC дефинисано као $IC(c) = -log\ P(c)$, а $P(c)$ је вероватноћа да се појам c открије у великом корпусу[91]. Ова мера користи садржај информација заједничких родитеља и заснива се на правилу: два појма су сличнија ако представљају више заједничких информација, а информације које деле два концепта означени су информацијским садржајем појмова. Lin [168] уводи меру која се темељи на Ресниковој мери и додаје фактор нормализације који се састоји од информацијског садржаја два улазна појма $sim = 2 * IC(LCS\ )IC(concpet_1) + IC(concpet_2)$. Наведеним мерама је заједничко то што одређују ниво повезаности концепата на основу информација из семантичких мрежа. Улазни параметри су онтолошки концепти, а излаз је нумеричка вредност у интервалу од 0 до 1. Дате мере прилично квалитетно функционишу на хијерархији WordNet-a, а како су ове структуре имплементиране коришћењем људског знања, дате мере могу успешно да моделују степен семантичке сличности међу појмовима. Међутим, главни недостатак заснива се на томе што су за примену оваквих приступа неизоставни људски и временски капацитети.

За разлику од семантичке сличности, семантичка повезаност се израчунава коришћењем метода векторске репрезентације речи (енгл. *word embedding*) [169] и неке од техника за рачунање сличности између вектора [161]. Векторска репрезентација речи је скуп техника језичког моделовања у области природног процесирања језика, где су речи или фразе из речника мапиране у векторе реалних бројева [170]. Вектори речи, конституисани су у векторском простору на тај начин, да се речи, које деле заједничке контексте у корпусу, налазе у непосредној близини у самом простору. Овај приступ се често назива статистичким приступом и генерално је заснован на тзв. *дистрибуционој хипотези* по којој се речи са сличним значењем појављују у сличним контекстима [171]. Применом ове хипотезе над неким корпусом текстова могуће је створити семантички простор, у коме је за сваку реч која се појављује у корпусу специфицирано колико пута се та реч појавила у контекстима (под термином „контекст“ подразумева се одређени документ из корпуса). Постоје различити модели који се заснивају на примени векторске репрезентације речи. Један од првих модела у овом домену јесте *word2vec* модел [172], који као улаз узима корпус и производи одговарајући векторски простор, при чему свака јединствена реч у корпусу добија одговарајући вектор у простору. Такође, један од начина представљања текстуалних докумената у форми вектора је применом *bag-of-words* [173] принципа, који разматра број појаве сваког израза (речи) у текстуалном документу [174]. Ова репрезентација резултира векторским представљањем текстуалних докумената, којом се свакој речи (изразу) додељује нумерички значај. Најкоришћенији модел заснован на овој идеји је модел векторског простора (енгл. *vector space model*) [175], који је детаљније представљен у одељку 7.5.9, с обзиром да је коришћен као један од приступа за имплементацију алгоритма сличности представљеном у овом истраживању. Предност векторског представљања речи заснива се на томе што је за креирање семантичког простора нужан једино корпус текстова и скуп програмерских метода за креирање вектора. У овом случају се од корисника не захтевају

[91] Збирка текстова природног језика синтетизована према одређеном критеријуму.

додатни речници нити посебно знање. Генерално, метода векторске репрезентације речи постиже позитивне и афирмативне резултате за многе задатке у NLP домену, о чему је дискутовано у прегледу литературе (одељак 7.4). Међутим, према сазнањима, ова метода је слабије експлоатисана у области онтолошког поравнања. Онтологије одражавају потенцијалне интерпретације термина, односно представљају основу за процесирање информација, и као такве се могу користити за подршку аутоматском семантичком тумачењу текстуалних информација [136]. У случају поравнања инстанци, метода векторског представљања речи била би од великог значаја за представљање објектних вредности (стрингова) у форми вектора. На тај начин би се одређеним терминима из онтологије доделили контекстни вектори, што би обезбедило да се поређење значења речи (онтолошких концепата) сведе на поређење њихових контекстних вектора. Над креираним контекстним векторима се може применити и нека од техника онтолошког поравнања, како би се одредила коначна сличност између инстанци. Међутим, главни проблеми са којима се методе векторског представљања речи суочавају заснивају се на томе да су могућа значења речи комбинована у једну репрезентацију [170] или да корпус садржи речи које су фреквентне, а нису од превеликог значаја за сам корпус. Због тога је пре представљања текстуалних података у форми вектора неопходно извршити њихово претпроцесирање. Тиме се пружа основа за софистицираније рударење текстуалних података, које је самим тим у синергији са терминима онтолошког поравнања и семантичке повезаности.

7.3 Рударење текстуалних података

Због обимности и сложености обраде неструктуираних података настала је специфична област истраживања која се бави обрадом текстуалних података - рударење текстуалних података (енгл. *text-data mining*) или текст аналитика (енгл. *text analytics*) [154]. „*Откривање података из текста означава обједињен скуп лингвистичких, статистичких метода и метода намењених машинском учењу које имају за циљ екстракцију информација из текстуалних извора. Информације, које се добију употребом текст аналитике, намењене су коришћењу у оквиру Big Data система, пословној интелигенцији, разним статистичким анализама и обрадама, предикцији понашања и будућих дешавања*“ [176]. Рударење текстуалних података значајну примену остварује и у домену биоинформатике, јер је ова област богата текстуалним подацима које су смештени у одређеним онтологијама. У случају овог истраживања методе рударења текстуалних података од релевантности су за процес онтолошког поравнања над биоинформатичким подацима. У наставку су објашњени основни процеси рударења текстуалних података кроз домен биоинформатике.

7.3.1 Процеси рударења текстуалних података

Рударење текстуалних података се углавном састоји од следећих процеса [136]:

- **Претраживање информација** - IR (*Information Retrieval*)
- **Екстракција информација** - IE (*Information Extraction*)
- **Рударење података** - DM (*Data Mining*)

Процес IR има за циљ прикупљање и филтрирање релевантних документа [136]. Стручњаци у домену биоинформатике и сродних области у великој мери користе IR процесе за лоцирање релевантних информација (најчешће у облику публикација) на интернету [136]. Многи IR алати су дизајнирани за претрагу посебних база података. На пример, за онтолошку базу података PubMed развијено је софтверско решење [177] које нуди софистицираниј претрагу литературе са компаративним листама кључних речи. Овај систем обавља једноставне аутоматске упите и редукује време анализе. У домену биоинформатике од значаја је не ограничавати IR процесе на тачно мечирање са појмовима упита, јер двосмислени термини (енгл. *term ambiguity*) и онтолошке варијације могу проузроковати претраживање ирелевантних информација (енгл. *low precision*) или превидети релевантне информације (енгл. *low recall*) [136]. Двосмисленост је инхерентна карактеристика природног језика и јавља се када се исти термин користи за више концепата (термин *покретач* (енгл. *promoter*) у домену биологије представља везујуће место у ДНК ланцу на коме се РНК полимераза везује како би иницирала транскрипцију гена, док се у домену хемије овај термин дефинише као супстанца која у веома малим количинама може да повећа

активност катализатора) [136]. Дакле, речи у једном речнику могу имати више значења и њихов смисао се може променити у зависности од контекста. Поред тога, хијерархијска организација онтологија и односи између концепата могу се користити за ограничавање упита за претрагу, као и за навигацију корисника кроз огромне количине јавних информација [136]. Mima и др. [178] користе онтологију за извођење софистициране претраге, омогућавајући корисницима да приступају имплицитно наведеним релевантним информацијама кроз хијерархијску експанзију упита. Müller и др. [179] су развили IR систем који ради на нивоу реченице и који користи специфично дизајнирану онтологију за упит корпуса за претраживање информација о одређеним класама биолошких концепата (нпр. генима, ћелијама итд.) и њиховим односима.

Процес IE има за циљ одабир специфичних чињеница о унапред одређеним врстама ентитета и релацијама од значаја [136]. Рани напори IE процеса били су посвећени препознавању назива ентитета (енгл. *Named Entity Recognition - NER*), односно препознавању појмова који означавају специфичне класе биоинформатичких ентитета (нпр. имена гена и протеина), након чега је уследила екстракција специфичних односа између таквих ентитета (нпр. протеин-протеин интеракције), а затим и екстракција сложенијих типова информација (нпр. метаболичких путева) [136]. Екстракција информација зависи од NER-а, али процес мапирања између термина и онтолошких концепата није тако једноставан. Један од главних разлога је тај што многи термини показују висок степен варијације, што заправо означава да се многи нови термини могу препознати као варијанте других. На пример, следећи термини су варијације [136]: *inner mitochondrial membrane vs. mitochondrial inner membrane* или *focal adhesion associated kinase vs. focal adhesion kinase*. Биоинформатичке онтолошке базе података су богате оваквим типовима информација, што са једне стране може бити привилегија у процесу онтолошког поравнања, јер би у том случају неке инстанце биле лакше препознате као идентичне (сличне).

Процес DM се може дефинисати као процес проналажења латентних законитости и веза међу подацима [136]. То је техника која подразумева претрагу података у циљу идентификације тражених узорака и њихових међусобних релација [180]. Уопштено, DM је облик екстракције свежих, интересантних и потенцијално утилитарних информација, садржаних у базама података. Због тога је DM познат и као процес откривања знања у базама података (енгл. *knowledge discovery in databases*) [181]. То је модус којим се сирови подаци обрађују (рударе) и претварају у исплативе информације, а све у циљу доношења функционалних одлука. „*Анализом огромних база података дефинишу се релације, обрасци или форме понашања, неопходне за одлучивање и предвиђање*“ [180]. Из тих разлога се DM неретко поистовећује са процесима откривања и предвиђања знања. Процес откривања знања имплицира корисниково разумевање изречених информација, које имају читљиву форму, док се предвиђање односи на будуће догађаје [182]. Тиме је корисницима омогућено да схвате релације између података и да препознају информације на начин који може допринети бољитку квалитета истраживања. Генерално, основни DM идеал јесте откривање непознатих релација између података: „*из масе података је потребно издвојити мале делове који представљају знање, а онда додатном обрадом створити ново знање*“ [182] и доћи до нових открића. Технике које се примењују у DM процесу су у великој мери реномиране математичке технике, процедуре и алгоритми који су коришћени годинама уназад. „*Иако је DM млада технологија, значајно се користе раније сазнања. Технике које се најчешће примењују углавном су изведене из три главне области*“ [180]: статистике (алгоритми регресије и стабла одлуке), машинског учења (стабла одлуке) и база података (класификација и кластеровање).

С обзиром да онтологије садрже термине, односно текстуалне репрезентације онтолошких концепата, процес онтолошког поравнања представља идеално подручје за примену метода рударења текстуалних података. У онтологији је текстуална репрезентација онтолошког концепта често еквивалентна текстуалном документу, који се може преузети применом IR процеса. На пример, једна од текстуалних репрезентација онтолошког концепта *drugbank:BE0000324* је текстуална (објектна) вредност облика *catalytic activity transferase activity methyltransferase activity transferase activity, transferring one-carbon groups 5,10-methylenetetrahydrofolate-dependent methyltransferase activity thymidylate synthase activity*. Ова вредност може представљати један текстуални документ. Текстуални документ може представљати и комбинацију више текстуалних репрезентација различитих онтолошких концепта. Као резултат можемо

имати велики број текстуалних докумената који представљају корпус, а који служи за даљу анализу. Генерално, IR процеси се другачије третирају у различитим истраживањима. Мао у истраживању [183] дефинише проблем онтолошког поравнања као IR проблем. Он сматра да ако се концепти у онтологији разматрају као документи у корпусу, онда је проблем утврђивања сличних концепата заправо еквивалентан проблему откривања сличних докумената. IR процес у овом истраживању мери сличност између упита (који се генеришу на основу профила сваког концепта у једној онтологији) и докумената, а затим се документа рангирају у складу са резултатом. Borbinha и др. [184] користе *tf-idf* меру [185] као један од примера IR процеса у циљу утврђивања сличних докумената у корпусу. Ова мера се користи да одреди тежину термина у корпусу и детаљније је размотрена у одељку 7.5.9. Процес IE се у случају онтолошког поравнања може третирати као могућност екстракције одређених релација или специфичних концепата, која се даље могу користити за процесирање. На пример, текстуалне (објектне) вредности предиката који се односе на специфичне идентификаторе као што су *dcterms:identifier*, *kegg_vocabulary:internal-id* или *bio2rdf_vocabulary:identifier* (за дефинисање јединствених идентификационих ознака инстанци), *chembl:cellosaurusId* (за означавање ID ћелијске линије у специфичној *Cellosaurus* онтологији) или *chembl:highestDevelopmentPhase* (за дефинисање нивоа развоја супстанце у процесу одобрења лекова), можда неће бити од значаја за даље процесирање. Такође, одређене нумеричке вредности које су представљене предикатима *drugbank_vocabulary:molecular-weight, kegg_vocabulary:exact_mass* или *chembl_molecule:alogp* можда неће бити од превеликог интереса за процес онтолошког поравнања и биће неопходно спровести специфично филтрирање у циљу њихове екстракције из текстуалних докумената. Са друге стране, можда ће од важности бити само специфичне информације, односно речи које имају већу фреквенцију појављивања у корпусу. Ови подаци се могу одредити применом *tf-idf* мере. Такође, могу се и користити NER процеси [186], који могу да елиминишу дупликате у текстуалним документима, што је нарочито важно због онтолошких варијација, а који су све више присутни у домену биоинформатике. DM процеси такође имају обећавајућу улогу у поступку онтолошког поравнања јер се над њима могу применити различите методе кластеровања или класификације текста. Такође, могуће је спровести селекцију и агрегацију одређених онтолошких техника на основу техника рударења података. На пример, Hariri и др. [187] користе стабла одлуке и неуронске мреже да би извршили селекцију и агрегацију онтолошких техника које се темеље на знаковима и на знању.

Кораци и методе рударења текстуалних података опциони су у процесу онтолошког поравнања, зависе од много фактора и подређени су коначном циљу. Такође, не постоји стриктна граница између IR и IE процеса, па често постоји преклапање одговарајућих метода. Методе у домену рударења текстуалних података константно се развијају и тешко је испратити њихово присуство на актуелној информатичкој сцени, због чега су оне ван оквира ове дисертације. Важно је схватити да се свака од наведених метода мора правилно исканалисати и применити на најпогоднији начин за постизање коначног циља. Без обзира на домен примене и на тип методе која се може приметни над текстуалним подацима, круцијално је спровести одређено претпроцесирање текстуалних података, како би се добили што прецизнији резултати приликом поређења докумената. Ово нарочито важи за домен онтолошког поравнања.

7.3.2 Претпроцесирање текста

Претпроцесирање текста је један од незаобилазних приступа у многим алгоритмима рударења текстуалних података и најчешће се заснива на следећим корацима [136]:

- Токенизација (енгл. *tokenization*);
- Утврђивање категорије речи (енгл. *part-of-speech tagging*);
- Филтрирање (енгл. *filtering*);
- Лематизација (енгл. *lemmatization*);
- Стемовање (енгл. *stemming*).

Токенизација [188] је први корак у аутоматском претпроцесирању текста. Она означава процес који идентификује основне текстуалне јединице (токене) који даље не подлежу процесу декомпозиције [136]. Грубо речено, токенизација подразумева декомпозицију текста у низ токена, односно појединачних речи.

Поступак токенизације се често ослања на елиминисање размака (енгл. *whitespace*) и знакова интерпункције. „*Чишћење речи од интерпункције се најчешће обавља регуларним изразом који је креиран да избрише све знакове који нису од интереса и не помажу током касније обраде. Друга могућност је да се користи replace метода доступна у већини програмских језика, при чему се интерпункцијски знакови замењују празним стрингом*“ [189]. Елиминисање знакова интерпункције није условљено само категоријом текста, већ и идејом даље обраде. У случају да је потребно упоређивати целе реченице, не саветује се уклањање знакова интерпункције, док се дати приступ саветује за процес поређења речи [189]. Приликом одређивања токена морају се узети у обзир и неки специјални случајеви: реч која има различите облике у једнини и множини требало би да означава исти токен у сваком од својих облика; речи могу бити скраћене на различите начине, а заправо требало би да представљају исти токен; треба водити рачуна о апострофима, који се често додају неким речима [189]. Механизам за проналажење оваквих речи и њихова замена истим токеном назива се нормализација [190]. Такође, поступак токенизације обухвата и трансформисање великих слова у мала. Посматрајмо објектну вредност инстанце *drugbank:BE0000324*: *catalytic activity transferase activity methyltransferase activity transferase activity, transferring one-carbon groups 5,10-methylenetetrahydrofolate-dependent methyltransferase activity thymidylate synthase activity*. Применом регуларног израза „*[a-z0-9]+*“ и елиминисањем размака резултат токенизације је: [*'catalytic', 'activity', 'transferase', 'activity', 'methyltransferase', 'activity', 'transferase', 'activity', 'transferring', 'one', 'carbon', 'groups', '5', '10', 'methylenetetrahydrofolate', 'dependent', 'methyltransferase', 'activity', 'thymidylate', 'synthase', 'activity'*].

Токенизацију обично прати форма лексичког процесирања која укључује процесе утврђивања категорије речи на основу њихове синтаксичке функције [136]. Под категоријом речи се подразумева утврђивање лексичких класа - именица, глагола, придева итд. Ове информације се добијају коришћењем *Part-Of-Speech* (POS) означивача [191] који су специфични за сваки језик. У датом примеру, класификација би била следећа: именице су *activity*, *transferase*, *methyltransferase*, *carbon*, *groups*, *methylenetetrahydrofolate*, *thymidylate* и *synthase*; придеви су *catalytic* и *antineoplastic*; *transferring* је глагол, док остали токени (речи) представљају бројеве. POS означивачи се обично реализују применом надгледаног машинског учења над корпусом текста који је ручно трениран [192]. За процес утврђивања категорија речи користе се и многи *online* алати који омогућавају једноставну класификацију[92].

Следећа фаза у процесу претпроцесирања текстуалних података подразумева филтрирање. Овај корак се спроводи у циљу елиминације речи које поседују релативно ниско информацијско значење. Ове речи се углавном фреквентно користе и познате су као стоп-речи (енгл. *stop-words*) [193]. Њихова заступљеност у корпусима је често редундантна и успорава спровођење анализе текста, па се зато оне искључују из даље анализе. Ове речи се обично деле у три групе: одредбене речи (*the, a, an, another*), везници (*for, but, or, yet, so*) и предлози (*in, under, towards*) [193]. „*Избор речи које ће се уклонити у овом кораку зависи од касније примене и од језика. Најчешће се речи бирају тако да је њихова учесталост у језику велика, а њихова улога у разумевању и решавању проблема небитна. Ове речи се уклањају ради смањења количине текста који се складишти и обрађује, а самим тим се утиче на убрзање. Додатно, смањује се разуђеност значајног текста*“ [189]. Многи програмски језици имају уграђене листе стоп-речи. Пакет *stop_words* у *Python* програмском језику дефинише листу стоп-речи: [*a, about, above, after, again, against, all, am, an, and, any, are, aren't, as, at, be, because, been, before, being, below, between, both, but, by, can't, cannot, could, couldn't, did, didn't, do, does, doesn't, doing, don't, down, during, each, few, for, from, further, had, hadn't, has, hasn't, have, haven't, having, he, he'd, he'll, he's, her, here, here's, hers, herself, him, himself, his, how, how's, i, i'd, i'll, i'm, i've, if, in, into, is, isn't, it, it's, its, itself, let's, me, more, most, mustn't, my, myself, no, nor, not, of, off, on, once, only, or, other, ought, our, ours, ourselves, out, over, own, same, shan't, she, she'd, she'll, she's, should, shouldn't, so, some, such, than, that, that's, the, their, theirs, them, themselves, then, there, there's, these, they, they'd, they'll, they're, they've, this, those, through, to, too, under, until, up, very, was, wasn't, we, we'd, we'll, we're, we've, were, weren't, what, what's, when, when's, where, where's, which, while, who, who's, whom, why, why's, with, won't, would, wouldn't, you, you'd, you'll, you're, you've, your, yours, yourself, yourselves*].

92 *https://parts-of-speech.info/*

Лематизација речи је још једна форма лексичког процесирања, који разматра морфолошку анализу речи, како би на основу ње био установљен корен и тачно значење речи. Ова метода отклања суфиксе речи, али тако што тежи да разуме значење речи, а затим на основу тога одлучује да ли је уопште уклањање суфикса неопходно и како га спровести [180]. У суштини, методе лематизације мапирају глаголске форме у инфинитиве (*transferring* у *transfer*) и именице на један облик (*transferase* и *methyltransferase* у *transferase*). У овој фази се врши и груписање речи различитих облика тако да се оне могу анализирати као појединачна реч. „*Ова метода се користи, зато што је парсирање целог текста скупа операција, а постоје случајеви у којима је одређивање фраза довољно за даљу обраду. Такође, парсирање није робустан метод*“ [189]. Парсирање реченице се врши у циљу одређивања синтаксне структуре целе реченице (енгл. *full or deep parsing*) или неких њених делова (енгл. *partial or shallow parsing*) [136]. „*Улога синтаксног парсера у претпроцесирању текста јесте да одреди структуру реченица у тексту. Парсирањем се одређује улога речи у реченици, нпр. субјекат, објекат, место радње, време радње и други. За парсирање се користе тагови речи које означавају њихово значење*“ [189]. Слично POS означавању, већина савремених парсера су статистичког типа, односно заснивају се на употреби машинског учења над тренираним корпусом података [192]. Због тога се парсери за сваки речник морају својствено конструисати. Ово је још напорније деловање од креирања POS означивача, јер је мануелно парсирање корпуса реченица оптерећујући процес. Како је POS врло захтеван и често склон дефектима, у пракси се ескивирају методе лексичког процесирања.

Методе стемовања (стемери) имају за циљ издвајање основе (корена) речи [136]. Стемери механички, без комплетног језичког знања, формирају лексички корен речи за већи број улазних речи. На пример, *transfer* је корен речи *transferring* и *transferase*. Након примене методе стемовања може се догодити да речи које су биле дефицитарне у тексту имају основу оних речи које се фреквентније појављују у тексту. Разлог томе је постојање других речи са истом основом. У случају онтолошког поравнања инстанци ово може бити жељени излаз, јер се на тај начин може лако одредити сличност одређених објектних вредности. Речи са заједничким кореном обично имају слично значење и зато је процес стемовања јако цењен, јер се на овај начин свакако редукује количина података коју је потребно даље анализирати. Ипак, мора се водити рачуна и о прецизности резултата, јер неретко се дешава да речи које имају заједничку основу нису семантички сличне [189]. Иначе, методе стемовања су језички зависне. Најпознатији примери стемера за енглески језик су Портеров [194] и Ловинсов [195] алгоритам. Портеров алгоритам (енгл. *Porter's stemming algorithm*) као најчешће коришћен енглески стемер постао је *de facto* стандард. Развио га је Мартин Портер 1980. године. Портеров алгоритам је детаљније представљен у одељку 7.5.9.

7.4 Преглед литературе

Проблем онтолошког поравнања је укорењен још у процесу интеграције података, где егзистира постојање идентичних и сличних података који могу припадати различитим хетерогеним изворима података. Како би се олакшао поступак међусобног повезивања инстанци (енгл. *RDF interlinking*) било је неопходно спровести процес њиховог поравнања. Због тога су задаци онтолошког поравнања, а самим тим и поравнања инстанци, постали један од круцијалних проблема који су разматрани како у домену база података, тако и у NLP сфери [196,197]. Међутим, ови приступи су слабије применљиви у контексту онтологија из тог разлога што не разматрају формалну семантику која је присутна у онтологијама [198]. У литератури се проблему поравнања инстанци приступа са два различита аспекта. Један аспект користи резултате упоређивања инстанци као средство за добијање тачнијих резултата подударања шеме (онтологије) [199], док други аспект користи резултате подударања шеме како би се побољшао процес упоређивања инстанци [200]. Идеја која стоји иза друге перспективе је та да ако су два онтолошка концепта семантички повезана онда је вероватно да су и њихове инстанце врло сличне (идентичне), али ако концепти нису повезани онда њихове инстанце неће бити повезане једна са другом [201]. Дакле, циљ је решити проблем поравнања инстанци и на тај начин допринети детерминацији сличних (идентичних) инстанци. Значај решења овог проблема је нарочито од велике важности у домену биоинформатике, што је представљено у одељку 7.1.

Иницијатива за евалуацију онтолошких поравнања OAEI (*Ontology Alignment Evaluation Initiative*)

датира од 2005. године и до сада је креирано десетине онтолошких система за поравнање [202]. Већина ових приступа је фокусирана на поравнање класа и релација користећи различите технике за текстуално рударење података, најчешће методе кластеровања [203] или лексичке и структурне карактеристике концепата [204]. На пример, AgreementMaker AML [205] врши поравнање класа и релација користећи лексичке и структурне карактеристике концепата. Овај приступ нуди неколико различитих техника поравнања. Најинтерсантнији је *word-matcher* који мери сличност између речи (стрингова), тј. објектних вредности предиката *rdfs:label* или *rdfs:name*. Сличност се мери помоћу тежинске JM (енгл. *weighted Jaccard Measure*) технике чија је вредност једнака 1- JM, а поравнање се примењује над класама. Интересантно је и лексичко процесирање (*lexical-matcher*) које се примењује искључиво над мањим онтологијама, јер је процес подложан грешкама. У овом случају се користи сличност заснована на речима (стринговима), у смислу провере да ли постоје сличне речи, групе сличних речи или акроними који могу да се мечирају са одговарајућим пуним именима. Структурне методе (*structural-matcher*) AML приступа се примењују над класама, како би се одредила њихова сличност на основу сличности њихових надкласа и потклаcа, користећи тежински фактор за прорачун удаљености.

Технике поређења инстанци користе различите приступе као што је терминолошка структура [206], логичко закључивање [207], релационо кластеровање [208] или комбинацију логичких и нумеричких метода [209]. Постоји велики број студија који проучавају примену ових метода, али је јако тешко открити истраживања која се заснивају искључиво на примени једне методе. Велика већина истраживања се заснива на комбинованом приступу, а неретко се користи и знање експерата, што указује на комплексност читаве процедуре. Аутори у [210] генерално разликују две категорије приступа: приступи који користе лингвистичке технике, односно називе и текстуалне описе инстанци (енгл. *linguistic-based approaches*) и приступи који користе технике које су засноване на ограничењима, односно типове података, домене и кључне карактеристике (енгл. *constraint-based approaches*). Прва група техника је заступљена у системима MEDLEY [211] и Hertuda [212], а друга група у истраживањима ASMOV [213] и PARIS [198]. MEDLEY [211] је систем који користи лексичке и структурне методе за израчунавање поравнања инстанци, а такође класа и релација. Лексичке технике подразумевају примену q-grams и LD техника. Над стринговима су у овом случају примењене методе претпроцесирања текста - токенизација и стемовање. Структурне технике одређују поравнање на следећи начин: ако ентитет који припада онтологији има суседа који је у скупу поравнања, онда ентитет са којим је сусед поравнат мора бити сусед било ком потенцијалном поклапању за тај ентитет. Hertuda [212] је врло једноставан систем који користи технике засноване на стринговима. За сваки концепт овај систем преузима ознаке (*rdfs:label*), коментаре (*rdfs:comment*) и специфичне URI фрагменте. На тај начин се за сваки концепт креира скуп термина. Затим се врши токенизација термина, а сличност између скупова термина одређује се применом *Damerau-Levenshtein* технике [214,142]. Овај приступ се делимично може применити над биомедицинским онтологијама FMA-NCI[93] (*Foundational Model of Anatomy*) и FMA-SNOMED[94]. ASMOV [213] је приступ који изводи поравнање из лексичких и структурних информација улазних онтологија, израчунавајући меру сличности између њих. Као метрику за утврђивање сличности између концепата, овај приступ користи дијапазон дефинисаних метода које представља Lin у својој студији [215]. Овај алгоритам такође укључује корак семантичке верификације где се поравнања проверавају, тако да коначни резултат не садржи семантичке недоследности. PARIS [198] је приступ који врши аутоматско поравнање инстанци, при чему се поступак поравнања у овом случају своди на основу процена вероватноће. Претпоставимо да две онтологије деле релацију *r*. Да би се одредила вероватноћа $Pr(x \equiv x')$ да су две инстанце x и x' сличне (еквивалентне) примењује се формула облика:

$$\exists r, y, y' : r(x, y) \wedge r(x', y') \wedge y \equiv y' \wedge f^{-1}(r).$$

Претходни приступи се углавном заснивају на откривању семантичке сличности између онтологија, односно инстанци. Међутим, проблем утврђивања семантичке повезаности постао је јако распрострањен имајући у виду да су онтологије постале ризнице знања и да су све богатије текстуалним информацијама,

93 *https://bioportal.bioontology.org/ontologies/FMA*

94 *http://www.snomed.org/*

које садрже семантички повезане податке. У циљу одређивања семантичке повезаности неки системи за онтолошко поравнање користе WordNet [156] лексичку базу, која дефинитивно решава проблеме значења речи и проблем синонима [216]. Међутим, кључни недостатак WordNet-а јесте његова ниска покривеност. Тачније, многи онтолошки елементи су ван домета ове лексичке базе, па се сличност између појединих елемената не може одредити. Последњих година се за одређивање семантичке повезаности често користи метода векторске репрезентације речи у комбинацији са неком од мера за одређивање сличности (CSM, Еуклидова удаљеност итд.), која се може применити над векторским величинама. Примена векторске репрезентације речи наишла је и на позитиван одзив у многим NLP задацима, укључују парсирање [217], језичко моделовање [218] итд. Међутим, на основу истраживања литературе у домену онтолошког поравнања, откривен је знатно мањи број публикација, које користе методу векторске репрезентације речи у циљу онтолошког поравнања. Репрезентативна истраживања су [219,132,131].

RiMOM [219] користи три различите стратегије поравнања: *name-based*, *metadata-based* и *instance-based*, чији резултати се затим филтрирају и комбинују. Прва стратегија мери број операција које је потребно применити да се један стринг (нпр. ознака ентитета) конвертује у други. Друга стратегија креира документе који садрже све речи ознака (*rdfs:label*) и коментара (*rdfs:comment*). На основу ових докумената, а применом *tf-idf* мере креирају се вектори над којима се примењује CSM. Трећа техника подразумева да се за одговарајући ентитет (класу) преузму све речи које су везане за све инстанце те класе. Над овако дефинисаним документима примењује се претходна техника. Ове стратегије се итеративно понављају све док постоје кандидати за поравнање.

Истраживање у [132] дефинише хибридни приступ који користи методу векторске репрезентације речи за процес онтолошког поравнања. Процес онтолошког поравнања се у овом приступу врши на нивоу ентитета и међусобно се пореде ознаке (*rdfs:label*), имена (*rdfs:name*) и коментари (*rdfs:comment*). Над сваким паром елемената за поређење $(name_1, name_2)$, $(label_1, label_2)$ и $(comment_1, comment_2)$ се примењује метода векторске репрезентације речи одбацивањем мање фреквентнијих речи (оних који се појављују мање од 5 пута у корпусу). За сваки пар ентитета $e_1 \in Entitiy_1$ и $e_2 \in Entitiy_2$, који припадају одговарајућим онтологијама O_1, O_2 ($Entitiy_i = C_i \cup P_i \cup I_i, i = 1,2$, унија класа (*C*), предиката (*P*) и инстанци (*I*)), сличност се одређује као максимална вредност претходно примењених метода: $sim = max\{sim(name_1, name_2), sim(label_1, label_2), sim(comment_1, comment_2)\}$. У овом случају се као мера сличности користи метода удаљености. Аутори овог истраживања су на основу експерименталних приступа дошли до закључка да комбинација онтолошког поравнања на нивоу класе уз примену методе векторске репрезентације постиже боље перформансе од других метода.

Аутори истраживања [131] предлажу апликацију која користи *word2vec* модел како би спровели процес онтолошког поравнања (класа и релација) и извршили одређену предикцију. За две улазне онтологије (изворну O^s и циљну O^t) и сваки пар ентитета $(e^s, e^t) \in O^s \times O^t$ потребно је утврдити њихову семантичку еквивалентност. Ентитети који се пореде морају притом имати исти скуп предиката, који се користе да означе имена (*rdfs:label*), алијасе и дефиниције. Кључан проблем, који овај приступ настоји да реши јесте селекција ентитета који ће се поредити. У циљу селекције, овај приступ користи меру инверзне учесталости документа (*idf*) токена над одређеним објектним вредностима. Овај приступ најпре проналази све заједничке токене (w_{s+t}), а затим рачуна *idf* сваког токена над датим сетом w_{s+t}, и на крају врши сумирање *idf*-ова ($idf_{total} = \sum_{i \in w_{s+t}} idf(i)$). Токени са вишим *idf* вредностима појављују се ређе у онтологији и више доприносе значењу одређеног ентитета. Сумирање *idf*-ова се врши за сваки циљни ентитет и као излаз утврђује се вредност од *K* циљних ентитета са највишим вредностима за сваки изворни ентитет, што директно редукује број парова за поређење $|O_S| \times K$. Објектне вредности датих ентитета се затим конветују у векторе, користећи *word2vec* модел, над којима се врши одређено тренирање и предвиђање вероватноће да су два ентитета еквивалентна. Евалуација ове методе је извршена над одређеним биоинформатичким базама података као што су HGNC, MeSH и OMIM.

Методе онтолошког поравнања користе се у домену биоинформатике и сродних наука за различите задатке. На пример, занимљив задатак је *largebio* (енгл. *Large BioMed Track*) који се састоји у откривању

поравнања између онтологија FMA-NCI, FMA-SNOMED CT и NCI Thesaurus[95] (*National Cancer Institute Thesaurus*). Ове онтологије су семантички богате и садрже десетине хиљада класа [220]. Такође, многи задаци онтолошког поравнања имају за циљ откривање сличних података: откривање сличних болести [221] или детекцију редундантних клиничких података [222]. Иако не постоји велики број система који се експлицитно бави поравнањем онтологија у домену биоинформатике, може се закључити да се сва претходно анализирана решења могу применити над онтологијама у овој области. Решења представљена у истраживањима [131,212] експлицитно користе биоинформатичке ресурсе за проблем онтолошког поравнања. Приликом примене онтолошког поравнања свакако је неопходно разграничити термине семантичке сличности и семантичке повезаности података, пре него што се спроведу одговарајући приступи.

На основу прегледа литературних података, може се закључити да је тема онтолошког поравнања изузетно популарна последњих година. Много апликација је развијено комбинацијом различитих приступа, што је допринело решавању једног од горућих проблема у домену семантичког веба, али и биоинформатике. Ипак, одсуство метода које користе векторске репрезентације речи отвара пут ка даљем усавршавању. У наставку је предложен алгоритам за проблем онтолошког поравнања, односно детекције сличних података, који је иницијално развијен за потребе Платформе, али који се може применити и у другим областима које подржавају онтолошко представљање података. Алгоритам се заснива на утврђивању семантичке повезаности између онтолошких концепата (инстанци) применом екстензијских техника онтолошког поравнања, методе векторске репрезентације речи и мере косинусне сличности.

7.5 Алгоритам детекције сличних података на Платформи

Као што је већ наглашено у одељку 7.1 откривање потенцијално сличних података у домену биоинформатике, има велики утицај на планирање будућих експеримента. Један од кључних принципа приликом извођења експеримената заснива се на хипотези да слични ентитети често деле сличне физичко-хемијске особине и биолошке активности [120]. На тај начин се могу уштедети ресурси и избећи експериментисање са компонентама које могу довести до потенцијално сличних резултата, али и указати кориснику да његово истраживање иде у добром смеру. Полазећи од дате хипотезе, тежило се ка развоју алгоритма који ће допринети решавању овог битног проблема у домену биоинформатике.

Алгоритам, који је развијен за потребе Платформе [7], има за циљ да детектује сличност између инстанци одговарајућих онтолошких база података, које су добијене као резултат извршавања предефинисаних упита. Предложени алгоритам се заснива на примени екстензијских техника онтолошког поравнања (у смислу поравнања инстанци), примени мере тежине термина (*tf-idf* мере), претпроцесирању текстуалних података, представљању текстуалних вредности у форми вектора (применом модела векторског простора) и одређивању угла косинуса између њих (применом мере косинусне сличности). Генерално, алгоритам се базира на примени лингвистичких техника и методе векторског представљања речи, док се као мера сличности примењује CSM. Свеобухватно, може се уочити да је примена екстензијских техника надограђена техником која се базира на токенима (CSM). На крају, од пресудног значаја је и улога експерата, која се огледа у пружању повратних информација о резултатима рада алгоритма.

У наставку је дат сажетији опис алгоритма: за улазне URI спецификације (инстанце онтолошких база података) утврђују се њихови предикати и објектне вредности; затим се обавља селекција предиката на основу мере тежине термина (*tf-idf* мере) над корпусом података (текстуалним документима који садрже објектне вредности инстанци); селектовани предикати се разматрају за сваку инстанцу и над њиховим објектним вредностима примењују се методе претпроцесирања текста (токенизација, филтрирање и стемовање); применом модела векторског простора текстуалне вредности се конвертују у векторске величине; над паровима вектора се затим примењује мера косинусе сличности како би се одредила сличност, при чему се одбацују они парови вектора са сличношћу испод 0.52 (праг сличности је одређен као приближна средња вредност резултата многобројиних тестирања CSM мере, укључујући и резултате

[95] *https://ncithesaurus-stage.nci.nih.gov/ncitbrowser/*

које представља Табела 7.6); на крају се врши агрегација резултата и њихово сортирање, од највећих (најсличнијих) до најмањих вредности. Листинг 7.1 представља псеудо-код алгоритма.

Улаз: $\langle (URI_1, endpoint_1), (URI_2, endpoint_2), \dots, (URI_n, endpoint_n) \rangle$

Излаз: $\langle (URI_1, URI_2, \dots, URI_m) \rangle, 0 \leq m \leq n$

1: излаз = [ ]
2: за сваки URI_i, $1 \leq i \leq n$ ради
3: креирати SELECT SPARQL упит $SQ_i, 1 \leq i \leq n$
4: извршити $SQ_i, 1 \leq i \leq n$
5: сачувати RDF триплете $T_j = (s, p, o), 1 \leq j \leq m, m \leq n$, и издвоји $T_j(\text{o}), 1 \leq j \leq m, m \leq n$ вредности у форми стрингова и груписати их у текстуалне документе $d_m, 1 \leq m \leq n$
6: над скупом текстуалних датотека $D = \{d_1, d_2, \dots, d_m\}$, $1 \leq m \leq n$ применити одређивање *tf-idf* мере и дефинисати скуп речи *words* којом се представљају термини од значаја
7: на основу скупа речи *words* врши се селекција предиката $p_k, k \gg 1$ из $T_j(p), 1 \leq j \leq m, m \leq n$ где, свака реч из $T_j(\text{o})$ припада $words$
8: на основу селектованих предиката извршити прикупљање свих објектних вредности $T_j(\text{o}), 1 \leq j \leq m, m \leq n$, за сваки URI_i, $1 \leq i \leq n$
9: за сваки $T_j(\text{o}), 0 \leq j \leq m, , m \leq n$ извршити методе претпроцесирања текста
10: конвертуј $T_j\langle \text{o} \rangle$ у $\overrightarrow{v_j}$ $1 \leq j \leq m, m \leq n$ користећи *VSM*
11: креирај $LCS = [\,]$
12: за сваки пар $\langle URI_l, URI_p \rangle, 1 \leq l, p \leq n, l \neq p$ ради
13: израчунај *CS* за $cs_{lp} = \langle \overrightarrow{v_l}, \overrightarrow{v_p} \rangle, 0 \leq l, p \leq m$
14: креирај $\text{cs}_{\text{lp}} = [\,]$
15: ако је $cs_{lp} \geq 0.52$
16: додај cs_{lp} у cs_lp
17: за сваки cs_lp израчунај $CS = \sum_1^{i \cdot j} cs_t$,
18: додај CS у LCS
19: сортирај LSC
20: сваки елемент у LSC одговара једном од улазних параметра $\langle URI_l, endpoint_l \rangle, 1 \leq l \leq n$
21: покупи $\langle URI_l \rangle$ из LSC и додај у излаз
22: уклони дупликате из листе излаз

Листинг 7.1 Псеудо-код алгоритма за детекцију сличних података на Платформи

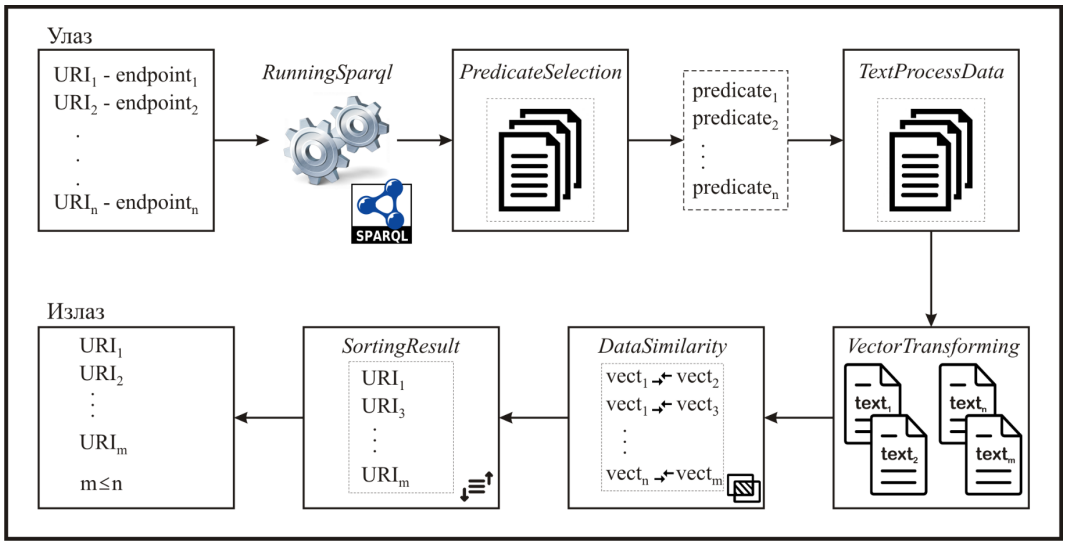
Алгоритам је имплементиран у *Python* програмском језику користећи неке од метода из пакета (модула):

- *sys* - системске функције
- *math* - основни математички оператори
- *validators* - валидација вредности
- *SPARQLWrapper* - извршавање SPARQL упита
- *JSON* - обрада JSON података
- *numpy* - управљање *n*-димензионалним низовима и извршавање софистицираних функција
- *collections* - управљање специјализованим типовима података (*dict*, *list*, *set* и *tuple*)
- *operator* - управљање скупом ефикасних функција које одговарају основним операторима *Python* програмског језика
- *re* - управљање операцијама који одговарају регуларним изразима
- *TextBlob* - процесирање текстуалних података
- *stop_words* - дефинисање листе стоп-речи
- *porter* - управљање операцијама који изводе процесирање текстуалних података у виду уклањања суфикса

7.5.1.1 Модел детекције сличних података

Компонента *DetectingSimilarDataItems*, која припада управљачком слоју архитектуре Платформе (Слика 5.1), обавља процес детекције сличних података. За спровођење овог задатка имплементиран је посебни модел (Слика 7.2), који се састоји од одређених оператора (подоператора), а који спроводи одговарајуће

кораке предложеног алгоритма.



Слика 7.2 Модел детекције сличних података предложен на Платформи

У наставку је детаљније представљен рад модела, анализа одређених делова кода у циљу потпунијег описа појединих корака алгоритма, као и опис математичких појмова, који се користе за имплементацију алгоритма.

7.5.2 Улазни параметри

Улазни параметри модела су вредности које су добијене извршавањем предефинисаних упита. Ове вредности су заправо инстанце, односно URI спецификације, које припадају одговарајућим базама података. Означимо улазне параметре листом парова:

$$\langle (URI_1, endpoint_1), (URI_2, endpoint_2), \dots, (URI_n, endpoint_n) \rangle.$$

Други елемент сваког пара одговара *endpoint*-у инстанце. У циљу ефикасније демонстрације рада алгоритма, као улазне параметре посматраћемо само одређене инстанце резултата предефинисаног упита за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* (Слика 6.9). Овакав вид ограничења био је неопходан имајући у виду да је број резултујућих инстанци знатно већи и да би било јако несистематично представити све кораке алгоритма над великим бројем улазних параметара. Нека је листа улазних параметара дата са:

$$\text{улаз} = [(pibas: TestTarget1, http://cpctas-lcmb.pmf.kg.ac.rs:3030/PIBAS/sparql), (drugbank: BE0000324, http://drugbank.bio2rdf.org/sparql), (kegg: 5f47d0b54b4d81097410bcc4cf01cf71, http://kegg.bio2rdf.org/sparql), (chembl\_target: 3160, https://www.ebi.ac.uk/rdf/services/sparql), (chembl\_target: CHEMBL1075416, https://www.ebi.ac.uk/rdf/services/sparql)]$$

Улазни параметри се користе за даље спровођење корака (3-5) алгоритма, а које обавља *RunningSparql* оператор. Овај оператор извршава одговарајуће SELECT SPARQL упите, чији је циљ преузимање свих предиката и објектних вредности улазних инстанци, односно креирање корпуса. Ови кораци одговарају IR процесима. Слика 7.3 представља део кода алгоритма који се извршава за поменуте кораке. Промењиве *?instance* и *?instance_endpoint* биће замењене улазним параметрима ($URI_i, endpoint_i$), $1 \leq i \leq n$. Извршавање упита се у овом случају изводи применом одговарајућих метода *SPARQLWrapper* пакета.

```
from stop_words import get_stop_words
from textblob import TextBlob as tb

stop_words = get_stop_words('english')
non_selected_predicates = ["http://rdf.ebi.ac.uk/terms/chembl#organismName",
"http://rdf.ebi.ac.uk/terms/chembl#targetType",
"http://bio2rdf.org/drugbank_vocabulary:organism"]
document=""

sparql = SPARQLWrapper(?instance_endpoint)
sparql.setQuery(
   """
   SELECT DISTINCT ?predicate ?object
   WHERE
   {

     <%s> ?predicate ?object

   }
   """ % (?instance)
   sparql.setReturnFormat(JSON)
   final_results = sparql.query().convert()
   for result in final_results["results"]["bindings"]:
     if result["predicate"]["value"] not in non_selected_predicates:
       if validators.url(result["object"]["value"])!= True:
         if result["object"]["value"] not in document:
           if result["object"]["value"] not in stop_words:
             document+=" "+str((result["object"]["value"]))
   doc=tb(document)
```

Слика 7.3 Део кода алгоритма за детекцију сличних података који извршава *RunningSparql* оператор у циљу креирања корпуса (текстуалних датотека) за улазне параметре *?instance* и *?instance_endpoint*

Резултат упита се филтрира, чиме се преузимају само стринг објектне вредности, односно одбацују се оне које су URL типа. Ово филтрирање се постиже применом *url* функције *validators* модула. Разлог због кога се директно у оквиру SPARQL упита не врши филтрирање по стринг вредностима је тај што неке објектне вредности нису директно означене као стринг вредности придруживањем *xsd:string* кодомена. На пример, објектна вредност предиката *dc:tittle* инстанце *drugbank:BE0000324* је означена са *Thymidylate synthase@en* и није директно окарактерисана као стринг, док је вредност предиката *dc:identifier* обележена као стринг *drugbank:BE0000324^^xsd:string*. Ово је један од честих проблема са којима се суочавају многе онтологије, укључујући и PIBAS онтологију. Тако је за инстанцу *pibas:TestTarget1* вредност предиката *pibas:hasTargetName*, означена са *thymidylate synthase* и није директно окарактерисана као стринг.

Селектоване објектне вредности једне инстанце i се затим групишу у текстуални документ d_i применом *TextBlob* пакета. Треба напоменути да се приликом креирања докумената одбацују сви стрингови који припадају групи стоп-речи. За одређивање ове листе речи користи се *get_stop_words* функција *stop_words* пакета. С обзиром да базе података предефинисаних упита користе енглески језик за текстуалну репрезентацију података, *stop_words* листа садржи речи енглеског језика које су представљене у одељку 7.3.1.

Табела 7.3 представља резултат рада *RunningSparql* оператора. Прва колона је резервисана за улазне параметре, док друга означава садржај текстуалне датотеке која одговара улазном параметру. Трећа колона означава назив датотеке. Скуп свих датотека означен је са $D = \{d_1, d_2, d_3, d_4, d_5\}$ и он представља корпус.

Табела 7.3 Резултат рада *RunningSparql* оператора за улазне параметре (биолошке мете) у циљу формирања корпуса

Улазни параметри	**Садржај текстуалне датотеке**	**Датотека**
pibas:TestTarget1	thymidylate synthase chemical agens	d_1
drugbank:BE0000324	Thymidylate synthase [drugbank:BE0000324] >Thymidylate synthase PVAGSELPRRPLPPAAQERDAEPRPPHGELQYLGQIQHILRCGVRKDD RTGTGTLSVFGM QARYSLRDEFPLLTTKRVFWKGVLEELLWFIKGSTNAKELSSKGVKI WDANGSRDFLDSL	d_2

	GFSTREEGDLGPVYGFQWRHFGAEYRDMESDYSGQGVDQLQRVIDT IKTNPDDRRIIMCA WNPRDLPLMALPPCHALCQFYVVNSELSCQLYQRSGDMGLGVPFNI ASYALLTYMIAHIT GLKPGDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDF KAEDFQIEGY NPHPTIKMEMAV TYMS >942 bp ATGCCTGTGGCCGGCTCGGAGCTGCCGCGCCGGCCCTTGCCCCCC GCCGCACAGGAGCGG GACGCCGAGCCGCGTCCGCCGCACGGGGAGCTGCAGTACCTGGG GCAGATCCAACACATC CTCCGCTGCGGCGTCAGGAAGGACGACCGCACGGGCACCGGCAC CCTGTCGGTATTCGGC ATGCAGGCGCGCTACAGCCTGAGAGATGAATTCCCTCTGCTGACA ACCAAACGTGTGTTC TGGAAGGGTGTTTTGGAGGAGTTGCTGTGGTTTATCAAGGGATCC ACAAATGCTAAAGAG CTGTCTTCCAAGGGAGTGAAAATCTGGGATGCCAATGGATCCCGA GACTTTTTGGACAGC CTGGGATTCTCCACCAGAGAAGAAGGGGACTTGGGCCCAGTTTAT GGCTTCCAGTGGAGG CATTTTGGGGCAGAATACAGAGATATGGAATCAGATTATTCAGGA CAGGGAGTTGACCAA CTGCAAAGAGTGATTGACACCATCAAAACCAACCCTGACGACAG AAGAATCATCATGTGC GCTTGGAATCCAAGAGATCTTCCTCTGATGGCGCTGCCTCCATGCC ATGCCCTCTGCCAG TTCTATGTGGTGAACAGTGAGCTGTCCTGCCAGCTGTACCAGAGA TCGGGAGACATGGGC CTCGGTGTGCCTTTCAACATCGCCAGCTACGCCCTGCTCACGTACA TGATTGCGCACATC ACGGGCCTGAAGCCAGGTGACTTTATACACACTTTGGGAGATGCA CATATTTACCTGAAT CACATCGAGCCACTGAAAATTCAGCTTCAGCGAGAACCCAGACCT TTCCCAAAGCTCAGG ATTCTTCGAAAAGTTGAGAAAATTGATGACTTCAAAGCTGAAGAC TTTCAGATTGAAGGG TACAATCCGCATCCAACTATTAAAATGGAAATGGCTGTTTAG Nucleotide transport and metabolism catalytic activity transferase activity methyltransferase activity transferase activity, transferring one-carbon groups 5,10-methylenetetrahydrofolate-dependent methyltransferase activity thymidylate synthase activity physiological process cellular metabolism nucleotide metabolism pyrimidine nucleotide biosynthesis pyrimidine nucleotide metabolism nucleobase, nucleoside, nucleotide and nucleic acid metabolism pyrimidine deoxyribonucleoside monophosphate biosynthesis dTMP biosynthesis pyrimidine nucleoside monophosphate biosynthesis 18p11.32 35585.0 Human TS TSase EC 2.1.1.45 7.0 None	
kegg:5f47d0b54b4d81 097410bcc4cf01cf71	thymidylate synthase inhibitor [HSA:7298] [KO:K00560] [kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71]	d_3
chembl_target:3160	CHEMBL3160 Thymidylate synthase	d_4
chembl_target:CHEM BL1075416	CAMA-1 CHEMBL1075416	d_5

На основу добијених резултата приступа се следећем кораку алгоритма - утврђивању тежине термина, како би се извршила адекватна селекција предиката.

7.5.3 Утврђивање тежине термина

Процес утврђивања тежине термина спроводи *TF-IDFMeasure* оператор. Овај оператор користи математичке принципе представљене у одељку 7.5.9.1. Улазни параметри овог оператора су текстуалне датотеке - документи (корпус D). Над документима се најпре примењује процес израчунавања *tf* мере према формули (**1**). У овом случају оператор *TF-IDFMeasure*, одређује број појављивања датог термина у разматраном документу. Затим се спроводи операција уз помоћ које се одређује *idf* мера. Ова мера се одређује према формули (**2**). На овај начин су веће тежине додељене терминима који нису толико фреквентни у корпусу. Према формули (**3**) оператор *TF-IDFMeasure* одређује *tf-idf* меру. На овај начин одређују се термини који су фреквентнији у корпусу и који могу бити од значаја за примену метода онтолошког поравнања. Слика 7.4 представља програмски код за израчунавање датих тежинских мера. Променљива *doc* означава један текстуални документ, док променљива *doc_list* означава корпус.

```
def tf(word, doc):
    return doc.words.count(word) / len(doc.words)

def n_containing(word, doc_list):
    return sum(1 for doc in doc_list if word in doc)

def idf(word, doc_list):
    return math.log(len(doc_list) / (1 + n_containing(word, doc_list)))

def tfidf(word, doc, doc_list):
    return tf(word, doc) * idf(word, doc_list)
```

Слика 7.4 Део кода алгоритма за детекцију сличних података за рачунање *tf*, *idf* и *tf-idf* тежинских мера

Селекција термина подразумева преузимање првих 5 речи са највећом *tf-idf* мером, који се складиште у *words* листу. Табела 7.4 представља резултат спровођења корака (6) за корпус D.

Табела 7.4 Резултат рада *TF-IDFMeasure* оператора у циљу формирања *words* листе

Датотека	**Речи**	***tf***	***idf***	***tf-idf***
d_1	thymidylate	0.25	0.22314	0.05579
	synthase	0.25	0.0	0.0
	chemical	0.25	0.91629	0.22907
	agens	0.25	0.91629	0.22907
d_2	activity	0.06742	0.91629	0.06177
	Nucleotide	0.05618	0.91629	0.05148
	metabolism	0.05618	0.91629	0.05148
	nucleotide	0.05618	0.91629	0.05148
	pyrimidine	0.04494	0.91629	0.04118
	biosynthesis	0.04494	0.91629	0.04118
	Thymidylate	0.03371	0.51083	0.01722
	thymidylate	0.03371	0.22314	0.00752
	synthase	0.03371	0.0	0.0
	methyltransferase	0.02247	0.91629	0.02059
	monophosphate	0.02247	0.91629	0.02059
	transferase	0.02247	0.91629	0.02059
	nucleoside	0.02247	0.91629	0.02059
	CTCCGCTGCGGCGTCAGGAAGGACGACCGCACGGGCACCGGCACCCTGTCGGTATTCGGC	0.01124	0.91629	0.0103
	ATTCTTCGAAAAGTTGAGAAAATTGATGACTTCAAAGCTGAAGACTTTCAGATTGAAGGG	0.01124	0.91629	0.0103
	cellular	0.01124	0.91629	0.0103
	TGGAAGGGTGTTTTGGAGGAGTTGCTGTGGTTTATCAAGGGATCCACAAATGCTAAAGAG	0.01124	0.91629	0.0103
	process	0.01124	0.91629	0.0103
	TACAATCCGCATCCAACTATTAAAATGGAAATGGCTGTTTAG	0.01124	0.91629	0.0103
	GCTTGGAATCCAAGAGATCTTCCTCTGATGGCGCTGCCTCCATGCCATGCCCTCTGCCAG	0.01124	0.91629	0.0103
	7.0	0.01124	0.91629	0.0103
	GACGCCGAGCCGCGTCCGCCGCACGGGGAGCTGCAGTACCTGGGGGCAGATCCAACACATC	0.01124	0.91629	0.0103
	nucleobase	0.01124	0.91629	0.0103
	ATGCCTGTGGCCGGCTCGGAGCTGCCGCGCCGGCCCTTGCCCCCCGCCGCACAGGAGCGG	0.01124	0.91629	0.0103
	CTGCAAAGAGTGATTGACACCATCAAAACCAACCCTGACGACAGAAGAATCATCATGTGC	0.01124	0.91629	0.0103
	TSase	0.01124	0.91629	0.0103
	GFSTREEGDLGPVYGFQWRHFGAEYRDMESDYSGQGVDQLQRVIDTIKTNPDDRRIIMCA	0.01124	0.91629	0.0103
	acid	0.01124	0.91629	0.0103
	catalytic	0.01124	0.91629	0.0103
	ACGGGCCTGAAGCCAGGTGACTTTATACACACTTTGGGAGATGCACATATTTACCTGAAT	0.01124	0.91629	0.0103
	physiological	0.01124	0.91629	0.0103
	CTCGGTGTGCCTTTCAACATCGCCAGCTACGCCCTGCTCACGTACATGATTGCGCACATC	0.01124	0.91629	0.0103
	18p11.32	0.01124	0.91629	0.0103
	TS	0.01124	0.91629	0.0103
	ATGCAGGCGCGCTACAGCCTGAGAGATGAATTCCCTCTGCTGACAACCAAACGTGTGTTC	0.01124	0.91629	0.0103
	2.1.1.45	0.01124	0.91629	0.0103
	CTGTCTTCCAAGGGAGTGAAAATCTGGGATGCCAATGGATCCCGAGACTTTTTGGACAGC	0.01124	0.91629	0.0103
	QARYSLRDEFPLLTTKRVFWKGVLEELLWFIKGSTNAKELSSKGVKIWDANGSRDFLDSL	0.01124	0.91629	0.0103
	GLKPGDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKAEDFQIEGY	0.01124	0.91629	0.0103
	NPHPTIKMEMAV	0.01124	0.91629	0.0103

	None	0.01124	0.91629	0.0103
	PVAGSELPRRPLPPAAQERDAEPRPPHGELQYLGQIQHILRCGV RKDDRTGTGTLSVFGM	0.01124	0.91629	0.0103
	35585.0	0.01124	0.91629	0.0103
	one-carbon	0.01124	0.91629	0.0103
	transferring	0.01124	0.91629	0.0103
	EC	0.01124	0.91629	0.0103
	CATTTTGGGGCAGAATACAGAGATATGGAATCAGATTATTC AGGACAGGGAGTTGACCAA	0.01124	0.91629	0.0103
	CTGGGATTCTCCACCAGAGAAGAAGGGGACTTGGGCCCAGT TTATGGCTTCCAGTGGAGG	0.01124	0.91629	0.0103
	bp	0.01124	0.91629	0.0103
	groups	0.01124	0.91629	0.0103
	nucleic	0.01124	0.91629	0.0103
	deoxyribonucleoside	0.01124	0.91629	0.0103
	942	0.01124	0.91629	0.0103
	drugbank	0.01124	0.91629	0.0103
	TYMS	0.01124	0.91629	0.0103
	5,10-methylenetetrahydrofolate-dependent	0.01124	0.91629	0.0103
	TTCTATGTGGTGAACAGTGAGCTGTCCTGCCAGCTGTACCAG AGATCGGGAGACATGGGC	0.01124	0.91629	0.0103
	WNPRDLPLMALPPCHALCQFYVVNSELSCQLYQRSGDMGLGV PFNIASYALLTYMIAHIT	0.01124	0.91629	0.0103
	BE0000324	0.01124	0.91629	0.0103
	dTMP	0.01124	0.91629	0.0103
	CACATCGAGCCACTGAAAATTCAGCTTCAGCGAGAACCCAG ACCTTTCCCAAAGCTCAGG	0.01124	0.91629	0.0103
	transport	0.01124	0.91629	0.0103
d_3	HSA:7298	0.14286	0.91629	0.1309
	kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71	0.14286	0.91629	0.1309
	KO	0.14286	0.91629	0.1309
	synthase	0.14286	0.0	0.0
	K00560	0.14286	0.91629	0.1309
	thymidylate	0.14286	0.22314	0.03188
	inhibitor	0.14286	0.91629	0.1309
d_4	CHEMBL3160	0.33333	0.91629	0.30543
	Thymidylate	0.33333	0.51083	0.17028
	synthase	0.33333	0.0	0.0
d_5	CAMA-1	0.5	0.91629	0.45815
	CHEMBL1075416	0.5	0.91629	0.45815
Резулат:	***words***=[*K00560, CHEMBL3160, Thymidylate, HSA:7298, agens, kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71, nucleotide, synthase, chemical, CAMA-1, pyrimidine, thymidylate, CHEMBL1075416, activity, inhibitor, Nucleotide, metabolism, KO*]			

На основу листе тежински значајних речи (*words*) приступа се следећем кораку алгоритма - селекцији предиката. Овај приступ утиче на избор објектних вредности над којима се примењују методе претпроцесирања текстуалних података.

7.5.3.1 Селекција предиката

Многе инстанце у онтолошким базама података неретко имају велики број предиката, а самим тим и велики број објектних вредности које се могу користити за онтолошко поравнање. На пример, инстанца *chembl_target:CHEMBL3160* Chembl/EMBL-EBI базе података је представљена са 1048, док је инстанца *drugbank:BE0000324* Drugbank/Bio2RDF базе података представљена са 48 различитих триплета. Број парова ентитета за поређење само ове две инстанце износио би 1048×48. За онтолошко поравнање пет инстанци, коришћених за демонстрацију рада алгоритма, тај број био би знатно већи што подразумева да је неопходно извршити оптимизацију алгоритма, како би се уштедели рачунарски ресурси. Оператор *RunningSparql* је у том погледу први извршио оптимизацију алгоритма, на тај начин што је обзир узео само триплете са објектним вредностима типа стринг (*xsd:string*), а одбацио све URL објектне вредности. Међутим, приликом поређења инстанци, многе објектне вредности не морају бити од интереса. Из тог разлога је алгоритам радом *TF-IDFMeasure* оператора додатно оптимизован, тако што је формирана листа тежински значајних речи - *words.*

Корак селекције предиката (7) поверен је оператору *PredicateSelection*, који на основу *words* листе врши селекцију предиката чије објектне вредности садрже неке од елемента ове листе. Издвојени предикати се преузимају и складиште у листу *selected_predicates*. Напоменућемо да су предикати *dc:tittle* и *rdfs:label* подразумевано укључени у листу, јер су од јако велике важности за поравнање инстанци и такође се користе као обавезни предикати за поређење и у другим приступима [205,212,219,132,131]. Приступ који се користи за селекцију предиката у [131] делимично се подудара са приступом представљеним у овом

истраживању. Оба приступа користе тежинске мере речи како би одредили значајне ентитете за поређење. Међутим основна разлика је у томе што аутори истраживања [131] користе унапред дефинисане атрибуте над којима врше процес спровођења тежинских мера. Иако не специфицирају атрибуте којима располажу у овом процесу, може се наслутити да се ради о атрибутима који припадају *rdfs* (*rdfs:name*, *rdfs:label*) и *owl* (*owl:sameAs*) речницима. Приступ представљен у дисертацији настоји да селектује и додатне предикате, који су од користи за поређење ентитета, што се може сматрати побољшањем. Ипак, алгоритам представљен у дисертацији има и један вид ограничења: за ChEMBL/EMBL-EBI и Drugbank/Bio2RDF базе података предикати *chembl:organismName, chembl:targetType* и *drugbank_vocabulary:organism* се не узимају у обзир приликом креирања корпуса. Ова листа предиката је коду дефинисана променљивом *non_selected_predicates*. Постојећи услов је био неопходан с обзиром да су објектне вредности ових предиката - речи *human*, *homo*, *sapiens, single* и *protein* - често припадале листи *words*. Поређењем само ових објектних вредности, инстанце које реално нису сличне имале би особину сличности. Због тога су ови предикати искључени из анализе још приликом рада *RunningSparql* оператора. Процеси које спроводи *PredicateSelection* могу се генерално окарактерисати као IE процеси. Табела 7.5 презентује резултат рада *PredicateSelection* оператора.

Табела 7.5 Резултат рада *PredicateSelection* оператора у циљу селекције предиката (*selected_predicates*)

Улазни параметри	**Предикати и објектне вредности**	**Датотека**
pibas:TestTarget1	*pibas:hasTargetName* thymidylate synthase *pibas:targetType* chemical agens	d_1
drugbank:BE0000324	*rdfs:label* Thymidylate synthase [drugbank:BE0000324] *dc:title* Thymidylate synthase *dc:identifier* BE0000324 *bio2rdf_vocabulary:namespace* Drugbank *drugbank_vocabulary:amino-acid-sequence* >Thymidylate synthase PVAGSELPRRPLPPAAQERDAEPRPPHGELQYLGQIQHILRCGVRKDDRT GTGTLSVFGM QARYSLRDEFPLLTTKRVFWKGVLEELLWFIKGSTNAKELSSKGVKIWD ANGSRDFLDSL GFSTREEGDLGPVYGFQWRHFGAEYRDMESDYSGQGVDQLQRVIDTIKT NPDDRRIIMCA WNPRDLPLMALPPCHALCQFYVVNSELSCQLYQRSGDMGLGVPFNIASY ALLTYMIAHIT GLKPGDFIHTLGDAHIYLNHIEPLKIQLQREPRPFPKLRILRKVEKIDDFKA EDFQIEGY NPHPTIKMEMAV *drugbank_vocabulary:gene-name* TYMS *drugbank_vocabulary:gene-sequence* >942 bp ATGCCTGTGGCCGGCTCGGAGCTGCCGCGCCGGCCCTTGCCCCCCGCC GCACAGGAGCGG GACGCCGAGCCGCGTCCGCCGCACGGGGAGCTGCAGTACCTGGGGCA GATCCAACACATC CTCCGCTGCGGCGTCAGGAAGGACGACCGCACGGGCACCGGCACCCT GTCGGTATTCGGC ATGCAGGCGCGCTACAGCCTGAGAGATGAATTCCCTCTGCTGACAACC AAACGTGTGTTC TGGAAGGGTGTTTTGGAGGAGTTGCTGTGGTTTATCAAGGGATCCACA AATGCTAAAGAG CTGTCTTCCAAGGGAGTGAAAATCTGGGATGCCAATGGATCCCGAGA CTTTTTGGACAGC CTGGGATTCTCCACCAGAGAAGAAGGGGACTTGGGCCCAGTTTATGG CTTCCAGTGGAGG CATTTTGGGGCAGAATACAGAGATATGGAATCAGATTATTCAGGACA GGGAGTTGACCAA CTGCAAAGAGTGATTGACACCATCAAAACCAACCCTGACGACAGAAG AATCATCATGTGC GCTTGGAATCCAAGAGATCTTCCTCTGATGGCGCTGCCTCCATGCCAT GCCCTCTGCCAG TTCTATGTGGTGAACAGTGAGCTGTCCTGCCAGCTGTACCAGAGATCG GGAGACATGGGC CTCGGTGTGCCTTTCAACATCGCCAGCTACGCCCTGCTCACGTACATG ATTGCGCACATC ACGGGCCTGAAGCCAGGTGACTTTATACACACTTTGGGAGATGCACAT ATTTACCTGAAT CACATCGAGCCACTGAAAATTCAGCTTCAGCGAGAACCCAGACCTTTC CCAAAGCTCAGG	d_2

	ATTCTTCGAAAAGTTGAGAAAATTGATGACTTCAAAGCTGAAGACTTT CAGATTGAAGGG TACAATCCGCATCCAACTATTAAAATGGAAATGGCTGTTTAG *drugbank_vocabulary:general-function* Nucleotide transport and metabolism *drugbank_vocabulary:go-function* catalytic activity *drugbank_vocabulary:go-process* physiological process; pyrimidine nucleoside monophosphate biosynthesis; nucleobase, nucleoside, nucleotide and nucleic acid metabolism *drugbank_vocabulary:name* Thymidylate synthase *drugbank_vocabulary:synonym* TS *drugbank_vocabulary:transmembrane-regions* None *drugbank_vocabulary:locus* 18p11.32 *drugbank_vocabulary:theoretical-pi* 7.0 *drugbank_vocabulary:molecular-weight* 35585.0	
kegg:5f47d0b54b4d81097410bcc4cf01cf71	*rdfs:label* thymidylate synthase inhibitor [HSA:7298] [KO:K00560] [kegg_resource:5f47d0b54b4d81097410bcc4cf01cf71] *dc:title* thymidylate synthase inhibitor [HSA:7298] [KO:K00560]	d_3
chembl_target:CHEMBL3160	*rdfs:label* Thymidylate synthase *dc:title* Thymidylate synthase *chembl:chemblId* CHEMBL3160	d_4
chembl_target:CHEMBL1075416	*rdfs:label* CAMA-1 *dc:title* CAMA-1 *chembl:chemblId* CHEMBL1075416	d_5
selected_predicates=[*rdfs:label,dc:title, pibas:hasTargetName, pibas:targetType, dc:identifier, drugbank_vocabulary:amino-acid-sequence, drugbank_vocabulary:general-function, drugbank_vocabulary:go-function, drugbank_vocabulary:go-process, drugbank_vocabulary:name, kegg_vocabulary:modifier, chembl:chemblId*]		

Када је листа предиката одређена, приступа се (8) кораку алгоритма - прикупљању објектних вредности улазних параметара URI_i, $1 \leq i \leq n$. Наиме, за сваки улазни параметар проверава се да ли његови предикати припадају листи *selected_predicates*. Уколико постоји поклапање, објектна вредност датог предиката се преузима и користи за наредни корак алгоритма - претпроцесирање текстуалних података, како би се створиле добре основе за методу векторске репрезентације речи. Пре него што се приступи овом кораку, анализиране су терминолошке технике и објашњено је зашто је одабран CSM (а самим тим и векторска репрезентација речи), као приступ који има потенцијално добар ефекат на рачунање сличности.

7.5.3.2 Избор терминолошке технике за рачунање сличности

У циљу избора терминолошке технике (која се примењује за рачунање сличности између текстуалних докумената) најпре је вршена детаљна анализа литературе. На основу истраживања [223] иницијално се може закључити да нити једна техника нема епитет „најбоље“ и да се не може са сигурношћу утврдити која је најадекватнија, јер њена примена зависи од потребе и ситуације. Као што је већ речено у одељку 7.2.1 технике засноване на кратерима и q-grams технике израчунавају сличност на основу секвенце карактера који су појављују у два стринга, док технике засноване на токенима деле стрингове у речи (симболе или токене) користећи као гранични карактер неки интерпункцијски знак или празан простор, а затим рачунају сличност између два скупа токена. У циљу поређења терминолошких техника извршено је њихово тестирање над паровима стрингова. Табела 7.6 представља резултат тестирања. Примери поређења су делимично преузети из литературе [141] и проширени биоинформатичким терминима. Сва мерења, изузев код техника базираних на карактерима, су нормализована на скали од [0,1]. Што је вредност ближа нули код техника базираних на карактерима, односно јединици код осталих техника, сличност између стрингова је већа. За тестирање техника коришћене су *Python* функције[96] које имплементирају формуле, које садржи Табела 7.2. Треба напоменути и то да су све улазне вредности, пре него што је мера сличности израчуната, конвертоване у стрингове са малим словима применом *Python* методе *lower*.

[96] Имплементација функција је доступна на *https://figshare.com/articles/Implementacija_terminoloskih_metoda/7577789*.

Табела 7.6 Резултати примене терминолошких техника над тестним паровима у циљу селекције најадекватније технике

Парови стрингова за поређење (стринг$_1$vs. стринг$_2$)	**Технике базиране на карактерима**		**q-gramms технике**	**Технике базиране на токенима**		
	LD	LCS	Trigrams	DC	JM	CSM
Bar vs. Car	1	2	1.0	**0.67**	**0.5**	0.0
Test vs. Rest	1	3	0.67	**0.86**	**0.75**	0.0
Cow vs. Caw	1	1	1.0	**0.67**	**0.5**	0.0
Pizza Buffa vs. Buffa Pizza	8	5	0.0	1.0	1.0	1.0
Microsoft Corporation vs. Microsft Corporation	1	14	0.33	0.96	0.92	0.5
Ruby L'otel vs. Ruby Lotel	1	6	0.43	0.95	0.9	0.41
Lenovo Inc vs. Lenovo	4	6	0.2	0.77	0.62	0.71
Thymidylate synthase vs. Synthase Thymidylate	16	11	0.0	1.0	1.0	1.0
Paul JONES vs. Paul JOHNSON	4	7	0.8	**0.9**	**0.82**	0.5
Out of the Blue vs. Of the Blue	4	11	0.25	1.0	1.0	0.87
Camphorato platinum vs. Platinum	11	8	0.57	0.76	0.61	0.71
Infuzions vs. W Infuzions	2	9	0.0	0.87	0.78	0.71
Activated protein vs. Protein actiavted	15	7	0.0	1.0	1.0	1.0
Thymidylate synthase vs. Thymidlyate synthase	2	12	0.42	1.0	1.0	0.5
Cisplatin vs. Cislpatin	2	4	0.73	1.0	1.0	**0.0**
Fluorouracilo vs. Fluorouracil	1	12	0.091	0.86	0.75	**0.0**
Dakota Brand of Fluorouracil vs. Gry Brand of Fluorouracil	6	22	0.28	0.84	0.72	0.75
Fluorouracil-GRY vs. Fluorouracil	4	12	0.26	0.8	0.67	0.71
Uracil, 5-fluoro- vs. Fluorouracil	16	6	0.43	0.97	0.94	**0.0**
Glutathione Homocysteine Transhydrogenase vs. Glutathione CoA glutathione Transhydrogenase	10	19	0.32	0.67	0.5	0.72
Проценат успешности				80%	80%	85%

Резултати (Табела 7.6) показују да прве две групе техника (технике базиране на карактерима и q-grams технике) раде генерално добро за типографске грешке. На пример, за парове стрингова *Micros**o**ft Corporation* и *Microsft Corporation*, односно *Cis**pl**atin* и *Cis**lp**atin*, обе технике дају прилично висок ниво сличности. Међутим, ове технике не успевају да утврде сличност кад се ред речи промени у стринговима. На пример, за пар стрингова *Thymidylate Synthase* и *Synthase Thymidylate* оне не дају добре резултате. Овај недостатак надомешћују технике базиране на токенима. Како су за потребе тестирања алгоритма, документи за поређење често поседовали стрингове са истим речима али различитог распореда, фокусирали смо се на технике базиране на токенима - DC, JM и CSM. Ове технике у већини случајева показују добре резултате, али се као боља за нијансу издвојила CSM. Лоши резултати примене ових техника су означени масним (енгл. *bold*) словима. У неким истраживањима [224,225] CSM је процењена као једна од најпопуларнијих. Популарност ове технике произилази из чињенице да уколико се над стринговима примене још додатни кораци текстуалног претпроцесирања, резултати ће бити бољи. На пример, ако се реч *Fluorouracilo* сведе на коренски облик *Fluorouracil* онда је резултат примене CSM над паровима стрингова *Fluorouracilo* и *Fluorouracil* једнак 1. Недостатак ове мере је низак ниво сличности у случају типографских грешака. На пример, за стрингове *Cis**pl**atin* и *Cis**lp**atin* вредност CSM је једнака нули. Међутим, проценат оваквих случајева међу онтолошким подацима је знатно мањи или занемарљив. CSM је зато усвојена као најбољи кандидат за утврђивање сличности. Такође, на основу резултата спроведних тестирања, релативно виоска успешност CSM технике (85%) додатно оправдава њен избор. Формулација и начин рачунања ове технике представљене су у одељку 7.5.9.2.

Примена CSM-а у случају алгоритма представљеног у дисертацији подразумева представљање речи коришћењем векторске репрезентације. Пре него што речи добију форму вектора извршава се њихово претпроцесирање. У наставку је реализација овог корака алгоритма детаљније представљена.

7.5.4 Претпроцесирање текстуалних података

Претпроцесирање текстуалних података, односно селектованих објектних вредности, обавља *TextProcessData* оператор. Рад овог оператора подељен је на неколико других подоператора, који имају одређену улогу у циљу коначне реализације обраде података:

- *LowerCaseTransformation* (*LCT*) оператор - изводи процес трансформације свих великих слова у мала слова;
- *RegularExpressTransformation* (*RET*) оператор - врши филтрирање стрингова применом регуларних израза;
- *StopWordFiltering* (*SWF*) оператор - отклања речи са релативно ниским информацијским значењем (стоп-речи);
- *StemmingRemoval* (*SR*) оператор - врши уклањање морфолошких суфикса речи.

Процеси које се спроводе радом ових подоператора одговарају корацима претпроцесирања текста представљеним у одељку 7.3.2. За сваки од њих користе су уграђене *Python* функције (методе): за процес трансформације великих слова у мала користи се уграђена *Python* метода *lower*; за процес филтрирања применом регуларних израза користи се одговарајућа *compile* функција *re* модула; за процес уклањања стоп-речи користи се листа речи која је одређена *get_stop_words* функцијом *stop_words* пакета; уклањање суфикса врши се применом Портеровог алгоритма [194] - *porter* модула.

Посматрајмо стринг *Nucleotide transport and metabolism*. Применом *LCT* подоператора вредност стринга постаје *nucleotide transport and metabolism*. Применом *SWF* подоператора уклања се стоп реч *and* која припада листи стоп речи (*stop_words*). Подоператор *RET* има задатак да уклони све карактере који не припадају скупу бројева или слова: *re.compile("[a-z0-9]+", re.I)*. На крају, *SR* подоператор уклања одговарајуће суфиксе применом Портеровог алгоритма. Тако се стринг *nucleotide transport metabolism* коначно трансформише у стринг *nucleotid transport metabol*. Слика 7.5 представља део програмског кода, који приказује функционисање датих подоператора.

```
splitter=re.compile("[a-z0-9]+", re.I)
stemmer=porter.PorterStemmer()

def text_process_data(word):
    """
    Function that normalizes word (string). First step
is converting to lower case. Second step is filtering
by regular expression. Third step is removing stop
word. Last step is stemming process.
    @word: string
    """
    w_lower=word.lower()
    w=splitter.findall(w_lower)
    if w[0] not in stop_words:
     ws=stemmer.stem(w[0],0,len(w[0])-1)
    return ws
```

Слика 7.5 Део кода алгоритма за детекцију сличних података који извршава *TextProcessData* оператор у циљу претпроцесирања текстуалних података

Табела 7.7 представља резултата рада *TextProcessData* оператора, односно његових подоператора.

Табела 7.7 Резултат рада *TextProcessData* подоператора у циљу претпроцесирања текстуалних података

Стрингови	**LCT**	**RET**	**SWF**	**SR**
Nucleotide transport and metabolism	nucleotide transport and metabolism	nucleotide transport and metabolism	nucleotide transport metabolism	nucleotid transport metabol
Thymidylate synthase	thymidylate synthase	thymidylate synthase	thymidylate synthase	thymidyl synthas
catalytic activity	catalytic activity	catalytic activity	catalytic activity	catalyt activ
physiological process	physiological process	physiological process	physiological process	physiolog process
Thymidylate synthase [drugbank:BE0000324]	thymidylate synthase [drugbank:be0000324]	thymidylate synthase drugbank be0000324	thymidylate synthase drugbank be0000324	thymidyl synthas drugbank be0000324
drugbank:BE0000324	drugbank:be0000324	drugbank be0000324	drugbank be0000324	drugbank be0000324
thymidylate synthase inhibitor [HSA:7298] [KO:K00560]	thymidylate synthase inhibitor [hsa:7298] [ko:k00560]	thymidylate synthase inhibitor hsa 7298 ko k00560	thymidylate synthase inhibitor hsa 7298 ko k00560	thymidyl synthas inhibitor hsa 7298 ko k00560
CAMA-1	cama-1	cama 1	cama 1	cama 1
Nucleobase, nucleoside,	nucleobase, nucleoside,	nucleobase	nucleobase	nucleobase

nucleotide and nucleic acid metabolism	nucleotide and nucleic acid metabolism	nucleoside nucleotide and nucleic acid metabolism	nucleoside nucleotide nucleic acid metabolism	nucleoside nucleotide nucleic acid metabol

Претпроцесирани стрингови се даље користе за наредну фазу алгоритма - конвертовање текстуалних у векторске вредности.

7.5.5 Трансформација текстуалних у векторске вредности

Оператор *VectorTransforming* спроводи корак (10) алгоритма, конвертујући текстуалне вредности у векторе и припремајући их за финалну анализу. У овом кораку примењује се *bag-of-words* принцип [173], који разматра број појаве сваке речи у одређеном документу. Овај приступ се поклапа са одређивањем *tf* мере према формули (**1**). Дата репрезентација доводи до векторског представљања текстуалних вредности, којом се свакој речи додељује нумерички значај. Модел који је у овом случају коришћен и који је заснован на идеји трансформације речи у векторе јесте VSM, који је детаљније представљен у одељку 7.5.9.1. Слика 7.6 представља програмски код за имплементацију ове фазе алгоритма.

```
def doc_vec(doc, key_idx):
    """
    Convert document to vector.
    @doc: string (document)
    @key_idx: dictionary
    """
    v = zeros(len(key_idx))
    for word in splitter.findall(doc):
        keydata = key_idx.get(stemmer.stem(word, 0, len(word)-1).lower(),None)
        if keydata: v[keydata[0]] = 1
    return v
```

Слика 7.6 Део кода алгоритма за детекцију сличних података који извршава *VectorTransforming* оператор који обавља процес трансформације текстуалних података у векторске вредности

Табела 7.8 представља резултат рада *VectorTransforming* оператора над одабраним улазним параметрима. Резултат је делимичан с обзиром да је укупан број комбинација за поређење стрингова знатно већи.

Табела 7.8 Делимични приказ резултата рада *VectorTransforming* оператора

Ознака парова за поређење	**Парови стрингова за поређење**	**Векторске величине**
$p_1(d_2 - d_4$; *drugbank_vocabulary:general-function - dc:title)*	nucleotid transport metabol	[1. 1. 0. 0. 1.]
	thymidyl synthas	[0. 0. 1. 1. 0.]
$p_2(d_2 - d_4$; *drugbank_vocabulary:go-function - dc:title)*	thymidyl synthas activ	[1. 1. 1.]
	thymidyl synthas	[0. 1. 1.]
$p_3(d_2 - d_4$; *drugbank_vocabulary:go-function - rdfs:label)*	thymidyl synthas activ	[1. 1. 1.]
	thymidyl synthas	[0. 1. 1.]
$p_4(d_2 - d_4$; *drugbank_vocabulary:title - dc:title)*	thymidyl synthas	[1. 1.]
	thymidyl synthas	[1. 1.]
$p_5(d_2 - d_4$; *drugbank_vocabulary:title - rdfs:label)*	thymidyl synthas	[1. 1.]
	thymidyl synthas	[1. 1.]
$p_6(d_2 - d_5$; *rdfs:label - rdfs:label)*	thymidyl synthas drugbank be0000324	[0. 1. 0. 1. 1. 1.]
	cama 1	[1. 0. 1. 0. 0. 0.]
$p_7(d_3 - d_4$; *dc:title - dc:title)*	thymidyl synthas inhibitor HSA 7298 KO K00560 kegg resourc 5f47d0b54b4d81097410bcc 4cf01cf71	[1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
	thymidyl synthas	[0. 0. 0. 0. 0. 0. 0. 0. 1. 1.]
p_8 $(d_3 - d_4$; *rdfs:label - dc:title)*	thymidyl synthas inhibitor hsa 7298 ko K00560	[1. 1. 1. 1. 1. 1. 1.]
	thymidyl synthas	[0. 0. 0. 0. 0. 1. 1.]
p_9 $(d_3 - d_4$; *dc:title - rdfs:label)*	thymidyl synthas inhibitor hsa 7298 ko K00560	[1. 1. 1. 1. 1. 1. 1.]
	thymidyl synthas	[0. 0. 0. 0. 0. 1. 1.]
$p_{10}(d_2 - d_5$; *drugbank_vocabulary:name - dc:title)*	thymidyl synthas	[0. 0. 1. 1.]
	cama 1	[1. 1. 0. 0.]
$p_{11}(d_2 - d_5$; *drugbank_vocabulary:name - rdfs:label)*	thymidyl synthas	[0. 0. 1. 1.]

	cama 1	[1. 1. 0. 0.]
$p_{12}(d_3 - d_5$; *drugbank_vocabulary:go-process - dc:title)*	pyrimidin nucleosid monophosph biosynthesi	[0. 1. 0. 1. 1. 1.]
	cama 1	[1. 0. 1. 0. 0. 0.]
$p_{13}(d_3 - d_5$; *drugbank_vocabulary:go-process - rdfs:label)*	nucleobase nucleoside nucleotide nucleic acid metabol	[0. 1. 0. 1. 1. 1. 1. 1.]
	cama 1	[1. 0. 1. 0. 0. 0. 0. 0.]
$p_{14}(d_2 - d_3$; *rdfs:label - drugbank_vocabulary:name)*	thymidyl synthas inhibitor HSA 7298 KO K00560 kegg resourc 5f47d0b54b4d81097410bcc	[1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
	thymidyl synthas	[0. 0. 0. 0. 0. 0. 0. 0. 1. 1.]
$p_{15}(d_2 - d_5$; *drugbank_vocabulary:go-function - rdfs:label)*	catalyt activ	[0. 1. 0. 1.]
	cama 1	[1. 0. 1. 0.]
$p_{16}(d_1 - d_2$; *pibas:hasTargetName - dc:title)*	thymidyl synthas	[1. 1.]
	thymidyl synthas	[1. 1.]
$p_{17}(d_1 - d_4$; *pibas:hasTargetName - rdfs:label)*	thymidyl synthas	[1. 1.]
	thymidyl synthas	[1. 1.]

Креиране векторске величине користе се као улазни параметри следећег корака алгоритма - рачунање сличности применом CSM-а.

7.5.6 Рачунање сличности

Рачунање сличности спроводи *DataSimilarity* оператор. Овај оператор најпре израчунава CSM над свим паровима вектора $(v_k, v_l), (URI_i, p_j, v_j)$, $1 \le i \le n, 1 \le j \le m$. Израчунавање ове мере постиже се функцијом *get_cosine* (Слика 7.7), која имплементира формулу (**4**). Улазни параметри функције су вектори над којима се примењују одговарајуће функције (*dot* и *norm*) *numpy* пакета у циљу одређивања излазне величине. Прва метода израчунава скаларни производ два вектора, а друга метода количник норми вектора. Повратна вредност функције одговара количнику ове две величине.

```
def get_cosine(vec1, vec2):
    """
    Function for calcaulation of cosine similarity measure between vectors.
    @parm:
    vec1: vector
    vec2: vector
    """
    numerator = dot(vec1,vec2)
    denominator = (norm(vec1) * norm(vec2))
    if not denominator or denominator==0.0:
       return 0.0
    else:
       return float(numerator) / denominator
```

Слика 7.7 Део кода алгоритма за детекцију сличних података који извршава *DataSimilarity* оператор који обавља израчунавање CSM-а. Функција *get_cosine* имплементира формулу (**4**)

Оператор *DataSimilarity* у следећем кораку одбацује оне парове вектора, код којих је примена *get_cosine* функције дала резултате ниже од 0.52 (праг сличности). Ове вредности се складиште у *cs_lp* листе. Свака листа одговара једном пару инстанци који се пореде. Дакле, ако су се поредили вектори који одговарају објектним вредностима инстанци *drugbank:BE0000324* и *chembl:CHEMBL3160*, резултати *get_cosine* функције смештају се у листу која одговара том пару инстанци. Табела 7.9 представља резултат рада *DataSimilarity* оператора над одговарајућим (селектованим) паровима вектора.

Табела 7.9 Резултат рада *DataSimilarity* оператора

Ознака парова за поређење	**Парови стрингова за поређење**	**CSM**
$p_1(d_2 - d_4$; *drugbank_vocabulary:general-function - dc:title)*	nucleotid transport metabol	0.0
	thymidyl synthas	
$p_2(d_2 - d_4$; *drugbank_vocabulary:go-function - dc:title)*	thymidyl synthas activ	0.82
	thymidyl synthas	
$p_3(d_2 - d_4$; *drugbank_vocabulary:go-function - rdfs:label)*	thymidyl synthas activ	0.82
	thymidyl synthas	
$p_4(d_2 - d_4$; *drugbank_vocabulary:title - dc:title)*	thymidyl synthas	1.0
	thymidyl synthas	

$p_5(d_2 - d_4$; *drugbank_vocabulary:title - rdfs:label)*	thymidyl synthas	1.0
	thymidyl synthas	
$p_6(d_2 - d_5$; *rdfs:label - rdfs:label)*	thymidyl synthas drugbank be0000324	0.0
	cama 1	
$p_7(d_3 - d_4$; *dc:title - dc:title)*	thymidyl synthas inhibitor HSA 7298 KO K00560 kegg resourc 5f47d0b54b4d81097410bcc 4cf01cf71	0.45
	thymidyl synthas	
p_8 $(d_3 - d_4$; *rdfs:label - dc:title)*	thymidyl synthas inhibitor hsa 7298 ko K00560	0.53
	thymidyl synthas	
p_9 $(d_3 - d_4$; *dc:title - rdfs:label)*	thymidyl synthas inhibitor hsa 7298 ko K00560	0.53
	thymidyl synthas	
$p_{10}(d_2 - d_5$; *drugbank_vocabulary:name - dc:title)*	thymidyl synthas	0
	cama 1	
$p_{11}(d_2 - d_5$; *drugbank_vocabulary:name - rdfs:label)*	thymidyl synthas	0
	cama 1	
$p_{12}(d_3 - d_5$; *drugbank_vocabulary:go-process - dc:title)*	pyrimidin nucleosid monophosph biosynthesi	0.0
	cama 1	
$p_{13}(d_3 - d_5$; *drugbank_vocabulary:go-process - rdfs:label)*	nucleobase nucleoside nucleotide nucleic acid metabol	0.0
	cama 1	
$p_{14}(d_2 - d_3$*; rdfs:label - drugbank_vocabulary:name)*	thymidyl synthas inhibitor HSA 7298 KO K00560 kegg resourc 5f47d0b54b4d81097410bcc	0.45
	thymidyl synthas	
$p_{15}(d_2 - d_5$; *drugbank_vocabulary:go-function - rdfs:label)*	catalyt activ	0.0
	cama 1	
$p_{16}(d_1 - d_2$*; pibas:TestTarget1 - dc:title)*	thymidyl synthas	1.0
	thymidyl synthas	
$p_{17}(d_1 - d_4$*; pibas:TestTarget1 - rdfs:label)*	thymidyl synthas	1.0
	thymidyl synthas	
Резултат: $cs\_lp_{d_2-d_4} = [0.82, 0.82, 1.0, 1.0], cs\_lp_{d_3-d_4} = [0.53, 0.53], cs\_lp_{d_1-d_2} = [1.0], cs\_lp_{d_1-d_4} = [1.0]$		

Када су утврђене сличности између парова вектора, односно одређене cs_lp листе, може се приступити реализацији последњих корака алгоритма - сумирању коначних вредности и одређивању излазних параметара.

7.5.7 Излазни параметри

Реализацију последњих корака алгоритма (17-21) обавља *SortingResult* оператор. Улазни параметри овог оператора су претходно креиране $cs\_lp_{m*m} = [cs_1, cs_2, \dots, cs_k], 1 \le m \le n$ листе, које одговарају паровима датотека $d_i - d_j, 1 \le i, j \le n, i \ne j$, односно инстанцама $URI_i - URI_j, 1 \le i, j \le n, i \ne j$ које се пореде, а које садрже CSM вредности. Оператор *SortingResult* најпре извршава сабирање CSM вредности сваке cs_lp листе. Ове вредности се затим смештају у LSC листу. Листа се потом сортира - од највећих до најмањих вредности. Излазни параметри (URI спецификације) се затим смештају у резултујућу листу из које се отклањају дупликати. Табела 7.10 представља кораке (17-21) алгоритма за улазне параметре:

$$cs\_lp_{d_2-d_4} = [0.82, 0.82, 1.0, 1.0], cs\_lp_{d_3-d_4} = [0.53, 0.53],\ cs\_lp_{d_1-d_2} = [1.0], cs\_lp_{d_1-d_4} = [1.0].$$

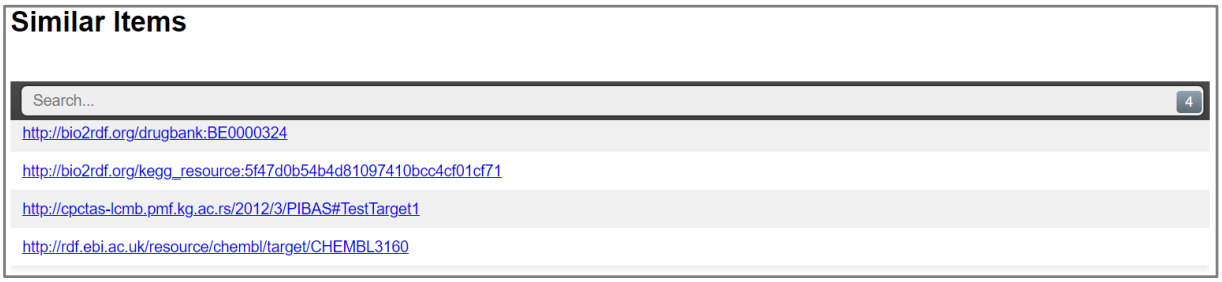
Табела 7.10 Резултат рада *SortingResult* оператора у циљу одређивања излазних параметара

Улазни параметри: $cs\_lp_{d_2-d_4} = [0.82, 0.82, 1.0, 1.0], cs\_lp_{d_3-d_4} = [0.53, 0.53], cs\_lp_{d_1-d_2} = [1.0], cs\_lp_{d_1-d_4} = [1.0]$		
Корак 17	**Корак 18**	**Корак 19**
$cs\_lp_{d_2-d_4} = [3.64], cs\_lp_{d_3-d_4} = [1.06],$ $cs\_lp_{d_1-d_2} = [1.0], cs\_lp_{d_1-d_4} = [1.0]$	LSC $= [cs_{lp_{d_2-d_4}}, cs_{lp_{d_3-d_4}}, cs_{lp_{d_1-d_2}}, cs\_lp_{d_1-d_4}]$	LSC $= [cs_{lp_{d_2-d_4}}, cs_{lp_{d_3-d_4}},$ $cs_{lp_{d_1-d_2}}, cs\_lp_{d_1-d_4}]$
Корак 20: излаз = [(drugbank: BE0000324, http://drugbank. bio2rdf. org/sparql), (chembl_target: 3160, https:// www. ebi. ac. uk/rdf/services/sparql), (kegg: 5f47d0b54b4d81097410bcc4cf01cf71, http://kegg. bio2rdf. org/ sparql), (pibas: TestTarget1, http://cpctas − lcmb. pmf. kg. ac. rs: 3030/PIBAS/sparql), (chembl_target: 3160, https:// www. ebi. ac. uk/rdf/services/sparql)]		

Корак 21: излаз = [(***drugbank***: ***BE0000324***, ***kegg***: ***5f47d0b54b4d81097410bcc4cf01cf71***, ***http***:// ***pibas***: ***TestTarget1***, ***chembl_target***: **3160**)]

7.5.8 Презентација резултата

Резултат примене методе за детекцију сличних података (кликом на дугме *Detect similar data*) јесте табеларни приказ резултата на новој веб страници са URI спецификацијама, односно сличним инстанцама (Слика 7.8).



Слика 7.8 Резултат примене методе детекције сличних података на Платформи за улазне параметре *pibas:TestTarget1*, *drugbank:BE0000324*, *kegg:5f47d0b54b4d81097410bcc4cf01cf71*, *chembl_target:3160* и *chembl_target:CHEMBL1075416*

7.5.9 Математички алати

7.5.9.1 Модел векторског простора

Модел векторског простора (енгл. *Vector Space Model - VSM*) је алгебарски модел који је представио Salton и др. [175] и који је првобитно примењен за индексирање и проналажење информација. У општем смислу, VSM омогућава презентацију и компарацију текстуалних података помоћу n-димензионалних вектора. Векторска репрезентација је најчешћа метода приказа текстуалног садржаја у циљу обраде и анализе података [226]. У том случају, у моделу векторског простора, свака реч представља променљиву која има нумеричку вредност која верификује тежину речи у документу. Процес одређивања тежине речи обично следи након неке форме процесирања текстуалног садржаја - токенизације, филтрирања, лематизације или стемовања. Иначе, процес филтрирања може бити јако важан за дефинисање тежине речи, јер се уклањањем стоп-речи смањује димензионалност векторског простора, што у великој мери утиче на перформансе алгоритма за рачунање тежине речи [226]. Један од најјаснијих начина израчунавања тежине термина у документима јесте *tf* (*Term Frequency*) мера.

Дефиниција 4. Нека је дат корпус $D = \{d_1, d_2, \dots d_D\}$ и нека $V = \{w_1, w_2, \dots w_v\}$ буде скуп различитих речи (термина) у корпусу. Мера *tf* означава учесталост појављивања термина (речи) у документу и израчунава на се следећи начин [227]:

$$tf(w_i) = c(w_i, d_j), 1 \leq i \leq V, 1 \leq j \leq D \quad \text{Ако се реч не појављује у } d_j \text{ онда је } tf(w_i) = 0. \qquad \textbf{(1)}$$

Дакле, на овај начин се могу утврдити термини који се фреквентније појављују у документу, односно они који су релевантнији за сам документ. Антагонистички приступ јесте инверзна фреквенција докумената - *idf* (*Inverse Document Frequency*), која се користи да одреди значај термина који нису толико заступљени у корпусу.

Дефиниција 5. Нека је учесталост речи $w_i \in V$ у документу $d \in D$ дато са $f_d(w)$ и нека је број докумената који поседују реч w означен са $f_D(w)$. Мера *idf* речи $w_i \in V$ се уовом случају рачуна на следећи начин [227]:

$$idf(w_i) = \log\left(\frac{|D|}{f_D(w)}\right), 1 \leq i \leq V, \text{ где је } |D| \text{ број документа у колекцији } D. \qquad \textbf{(2)}$$

Тежина речи је знатно делотворнија мера код дужих текстова, јер се код сажетијих текстова све речи појављују неколико пута и та информација може навести на погрешан закључак [226]. Због тога је концепција валоризовати све термине који нису уобичајени у корпусу (релативно виско *idf*), а при томе имају незанемарљив број појављивања у датом документу (релативно висок *tf*). Метрика за мерење оваквих термина у VSM-у јесте *tf-idf* и рачуна се према формули [227]:

$$tf - idf(w_i) = tf(w_i) \times idf(w_i), 1 \leq i \leq V \quad \textbf{(3)}$$

7.5.9.2 Мера косинусне сличности

Када су документи (односно текстуалне вредности) представљени као вектори, сличност између њих одговара корелацији између самих вектора [228]. Сличност докумената s и t може се израчунати применом једне од најпроминентнијих мера сличности, мере косинусне сличности. Ова вредност се рачуна као косинус угла између два вектора. Због тога се дата мера назива мером косинусне сличности. У наставку следи дефиниција.

Дефиниција 6. Нека су дата два документа *s* и *t*, тада је мера косинусне сличности једнака [228]:

$$\cos(s,t) = \frac{s * t}{\|s\| \, \|t\|} = \frac{\sum_{i=1}^{n} s_i t_i}{\sqrt{\sum_{i=1}^{n} (s_i)^2 \sum_{i=1}^{n} (t_i)^2}} \quad \textbf{(4)}$$

Поступак се извршава тако што се скаларни производ вектора, подели производом Еуклидових норми вектора. Као резултат примене мере косинусне сличности, добија се вредност у интервалу од 0 до 1. Ако се упоређују два идентична документа, онда је резултат примене ове мере једнак 1, односно два документа се сматрају потпуно идентичним [228]. Такође, врло битна карактеристика мере косинусне сличности јесте аутономија у односу на дужину докумената.

7.5.9.3 Портеров алгоритам

За уклањање суфикса речи, у циљу издвајња основе или корена речи, најчешће се користи Портеров алгоритам. У наставку следи дефиниција.

Дефиниција 7. Ако се променљивом P означе суседни сугласници, а променљивом K суседни самогласници у речи, онда се свака реч може представити формулом облика [194]:

$$[P](KP)^n[V], \quad \textbf{(5)}$$

где угласте заграде означавају опциону појаву променљиве. Ознака *n* указује колико пута се структура KP појављује у речи. На основу речи дефинише се и листа правила, која означава када је могуће елиминисати одговарајуће префиксе. Правила су облика [194]:

$$if(con)\ l_1\ then\ l_1 \rightarrow\ l_2, \quad \textbf{(6)}$$

што означава да ако је испуњен услов (con), а реч се завршава са l_1, онда се l_1 замењује са l_2. При томе је l_2 најчешће празна секвенца. Услов највише зависи од броја n, а l_1 су специфични суфикси који су атрактивни за примену овог правила. Услови могу бити и слојевити, због чега се l_1 може заменити са више правила.

8 Резултати и дискусија

Главни исход дисертације јесте софтверско решење (Платформа) које омогућава откривање знања у домену биоинформатике извршавањем предефинисаних Federated SPARQL упита уз методу додавања кориснички селектоване базе података, као и динамичког филтрирања резултата упита у циљу побољшања релевантности резултата. Као напредна метода Платформе издваја се могућност детекције сличних података над резултатима извршавања предефинисаних упита. У првом делу овог поглавља приказани су резултати примене основних метода Платформе кроз различите тестне сценарије, као и компарација резултата са другим значајним софтверским решењима у овом домену. У другом делу поглавља тестиран је алгоритам за детекцију сличних података. Тестирање је спроведено над резултатима предефинисаних упита, али и над улазним параметрима PubChem, UniChem[97], UniProt[98] и canSAR[99] база података, како би се показала независност алгоритма од Платформе. Такође, да би се нагласио значај селекције предиката, који се сматра кључним кораком за успешан исход примене алгоритма, спроведено је тестирање методе без примене овог корака. Такође, извршена је и компарација резултата алгоритма са неким од адекватних приступа у овом домену. У последњем делу поглавља су разјашњена одређена ограничења која утичу на перформансе Платформе.

8.1 Анализа и дискусија резултата основних метода

У овом одељку су кроз различите тестне сценарије представљене релевантности метода: извршавања предефинисаних упита, динамичког филтрирања резултата упита и додавања кориснички селектоване базе података. Значај ових метода је аргументован у смислу задовољавања критеријума за шаблоне (*) (одељак 4.5.3). Прве две методе су демонстриране у свим сценаријима, док је трећа метода демонстрирана у последњем сценарију. Тестни сценарији су спроведени над улазним параметром - активном супстанцом, која поседује званични статус лека (верификована од стране FDA). У овом случају као активна супстанца (*тест супстанца 1*) коришћен је лек *Fluorouracil*[100]. Резултати упита за ову активну супстанцу укључују и податке из CPCTAS базе. За други вид тестирања, односно за тестирање супстанце која званично није одобрена од стране FDA, коришћена је активна супстанца (*тест супстанца 2*) чија је *InChiKey* вредност дата са RGVRUQHYQSORBY-JIGXQNLBSA-N[101]. Овај приступ размотрен је са становишта корисника Лабораторије.

Сценарио 1

Откривање информација о лековима доступно је извршавањем предефинисаног упита за шаблон (*d*) *Find info about drug*. Претпоставимо да је корисник заинтересован да открије информације о *тест супстанци 1*. Резултат извршавања предефинисаног упита овог шаблона јесте листа лекова, односно инстанци (Слика 8.1): по једна инстанца из PIBAS/CPCTAS, Drugbank/Bio2RDF и Chembl/EMBL-EBI базе података, као и две инстанце из Kegg/Bio2RDF базе података. Свака инстанца је линкована и може се приступити њеној дескрипцији[102]. На основу анализе дескрипције може се проверити да ли супстанца крши неко од правила Липинског представљено у одељку 4.5.3, а затим се може спознати и њен механизам акције.

97 *https://www.ebi.ac.uk/unichem*

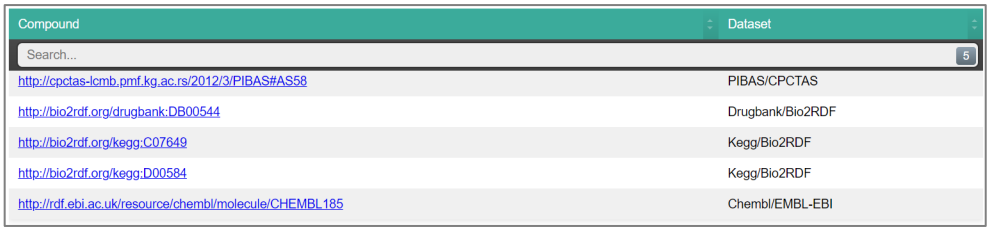
98 *https://www.uniprot.org*

99 *https://cansarblack.icr.ac.uk*

100 *InChiKey*: GHASVSINZRGABV-UHFFFAOYSA-N, *SMILES*: FC1=CNC(=O)NC1=O

101 *Canonical SMILES*:
COc1cccc2C(=O)c3c(O)c4C[C@@](O)(CCO)C[C@H](O[C@H]5C[C@H](N)[C@H](O)[C@H](C)O5)c4c(O)c3C(=O)c12

102 Дескрипције су разматране у одељку 6.1.3. У наставку су изостављни њихови визуелни прикази.



Compound	Dataset
Search...	5
http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#AS58	PIBAS/CPCTAS
http://bio2rdf.org/drugbank:DB00544	Drugbank/Bio2RDF
http://bio2rdf.org/kegg:C07649	Kegg/Bio2RDF
http://bio2rdf.org/kegg:D00584	Kegg/Bio2RDF
http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL185	Chembl/EMBL-EBI

Слика 8.1 Резултат извршавања предефинисаног упита за откривање информација о леку *Fluorouracil* на Платформи

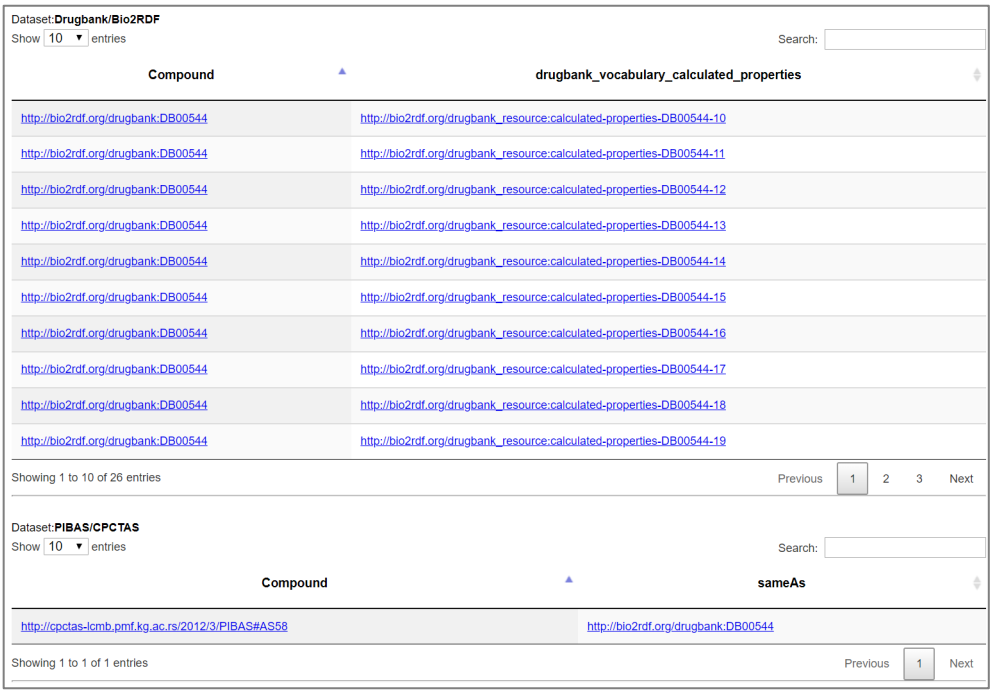
Посматрајмо најпре инстанцу *drugbank:DB00544*, односно њену дескрипцију (кликом на линк *http://bio2rdf.org/drugbank:DB00544*). Дескрипција садржи велики број предиката и објектних вредности које је описују. Информације о правилима биорасположивости су доступне кроз објектну вредност предиката *drugbank_vocabulary:calculated-properties*, која представља молекуларне дескрипторе. У овом случају постоје две *logP* вредности које су израчунате применом софтверских решења ALOGPS[103] и ChemAxon[104]. Прва вредност износи -0.58, а друга -0.66 и обе су у границама погодних. Молекуларна тежина је такође израчуната применом ChemAxon софтверског решења и износи 130.0772. Вредности *HBA* и *HBD* су у границама дозвољених. За разлику од Drugbank/Bio2RDF базе података, инстанце *kegg:C07649* и *kegg:D00584* пружају информације једино о молекуларној тежини (вредност 130.077194) кроз објектну вредност предиката *kegg_vocabulary:mol_weight*. Инстанца *chembl_molecule:CHEMBL185* такође садржи све предикате из којих се може одредити да ли је неко од правила Липинског прекршено. Предикати за одређивање ових вредности су: *chembl_molecule:alogp*(*:hbd*, *:hba*). Све ове вредности поклапају се са вредностима дефинисаним у Drugbank/Bio2RDF бази података. База података PIBAS/CPCTAS не поседују такве типове предиката, али с обзиром да је она интегрисана са Bio2RDF репозиторијумом применом својства *pibas:sameAs*(*pibas:AS58*, *drugbank:DB00544*)[105], може се закључити да важе исте карактеристике. Потврда да правила Липинског нису прекршена, односно да дато јединење заиста представља лек јесте и вредност *rdf:type* предиката (*drugbank_vocabulary:Drug*), што је експлицитно дефинисано у Drugbank/Bio2RDF бази података. И остали предикати су од важности и сугеришу на неке од битних физичко-хемијских особина лека који се користе за разне QSAR методе. Анализом објектних вредности одређених предиката може се утврдити и механизам акције датог лека. На пример, у Drugbank/Bio2RDF бази података за ову улогу је експлицитно задужен предикат *drugbank_vocabulary:mechanism-of-action*: *The precise mechanism of action has not been fully determined, but the main mechanism of fluorouracil is thought to be the binding of the deoxyribonucleotide of the drug (FdUMP) and the folate cofactor, N5–10-methylenetetrahydrofolate, to thymidylate synthase (TS) to form a covalently bound ternary complex. This results in the inhibition of the formation of thymidylate from uracil, which leads to the inhibition of DNA and RNA synthesis and cell death. Fluorouracil can also be incorporated into RNA in place of uridine triphosphate (UTP), producing a fraudulent RNA and interfering with RNA processing and protein synthesis*. Ово директно указује на биолошку мету (*thymidylate synthase*) која се може експлоатисати у експерименталним истраживањима истог или сличног лека. Предикати који дефинишу механизам акције нису директно доступни у другим базама података. На пример, непосредна информација о механизму акције није доступна у Kegg/Bio2RDF и PIBAS/CPCTAS базама података, али уз подршку кориснички детерминисаних предиката *drugbank_vocabulary:x-kegg*(*kegg:C07649*, *drugbank:DB00544)* и *pibas:sameAs*(*pibas:AS58*, *drugbank:DB00544*) могућ је индиректан начин откривања информација. Ово је заправо последица интеграције података и због тога је од велике користи укључити што више база у предефинисани упит како би корисник открио комплементарне информације. Интеграција је присутна и кроз кориснички дефинисане предикате: *drugbank_vocabulary:x-chebi*, *kegg_vocabulary:x-pubchem.compound, drugbank_vocabulary:x-pharmgkb*, *chembl:moleculeXref* итд. Захваљујући *cross-references* подацима изводе се директни закључци многих параметара који нису доступни у једној бази података.

103 *http://www.vcclab.org/lab/alogps/*

104 *https://chemaxon.com/*

105 Нови вид интеграције PIBAS/CPCTAS базе података са Drugbank/Bio2RDF базом података.

Као што је већ наведено у одељку 4.5.3 анализа дескрипције појединачних инстанци може бити од велике важности у откривању информација од интереса, али велики број предиката који описује инстанце може збунити и оптеретити корисника у процесу анализе. Додатно, многе дескрипције не морају приказивати све предикате инстанце и то може бити одређено правилима репозиторијума. Метода динамичког филтрирања резултата предефинисаног упита у овом случају олакшава процес претраге и анализе података. Број предиката у оквиру панела који одговара одређеној бази података зависи од саме инстанце и може бити различит од онога што је представљено на дескрипцијама. Слика 8.2 представља резултат примене методе динамичког филтрирања резултата за селектоване предикате *drugbank_vocabulary:calculated-properties* и *pibas:sameAs*. На овај начин корисник може лако проверити са којом базом је интегрисана PIBAS/CPCTAS база података, али и открити молекуларне дескрипторе *мест супстанце 1*. У овом случају се од корисника, који је потпомогнут описом предиката (линкова у оквиру панела или дескрипције самих предиката) очекује да поседује одређено доменско знање. На пример, селекција предиката *drugbank_vocabulary:calculated-properties* подразумева претходно знање о молекуларним дескрипторима, односно корисник мора знати да су *calculated properties* резултати математичке или логичке процедуре која претвара хемијске информације у нумерички облик. Као што се може приметити, резултат филтрирања за Drugbank/Bio2RDF базу података није идентичан објектној вредности која је представљена у оквиру дескрипције, већ је сваки дескриптор представљен као инстанца класе *drugbank_vocabulary:Resource*.


Dataset:**Drugbank/Bio2RDF**

Show 10 entries Search:

Compound	drugbank_vocabulary_calculated_properties
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-10
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-11
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-12
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-13
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-14
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-15
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-16
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-17
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-18
http://bio2rdf.org/drugbank:DB00544	http://bio2rdf.org/drugbank_resource:calculated-properties-DB00544-19

Showing 1 to 10 of 26 entries Previous 1 2 3 Next

Dataset:**PIBAS/CPCTAS**

Show 10 entries Search:

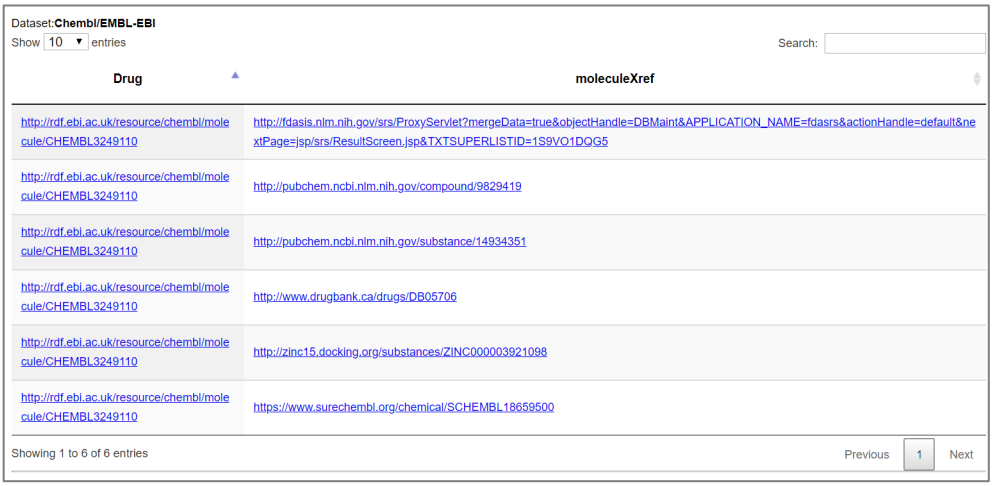
Compound	sameAs
http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#AS58	http://bio2rdf.org/drugbank:DB00544

Showing 1 to 1 of 1 entries Previous 1 Next

Слика 8.2 Резултат извршавања методе динамичког филтрирања резултата упита за откривање информација о леку *Fluorouracil* на Платформи (за селектоване предикате *drugbank_vocabulary:calculated-properties* и *pibas:sameAs*)

Извршавањем шаблона *Find info about drug* може се открити да ли једињење (активна супстанца) има пионирску улогу у експерименталним истраживањима. Иако предефинисани упит (*d*) покрива важне и проминентне биоинформатичке базе података, не може се са кредибилношћу тврдити да неко једињење већ није нашло примену у експерименталним процесима. Разлог томе је што су многа једињења, која су у фази истраживања и која званично немају статус лека, латентна од шире биоинформатичке јавности. Претпоставимо да је корисник Лабораторије синтетисао *мест супстанцу 2*. За новосинтетисано једињење, корисник пре свега жели да провери да ли је оно већ познато биоинформатичкој заједници, а затим и да планира експериментална истраживања на основу доступних података. Извршавањем предефинисаног упита, корисник открива инстанцу *chembl_molecule:CHEMBL3249110*. Анализом дескрипције можемо закључити да су сва правила Липинског прекршена (молекуларна тежина износи 529.54, док су *HBA* и *HBD* вредности, респективно, 11 и 6). Дакле, синтетисано једињење нема статус

лека. Метода динамичког филтрирања резултата у овом случају за селектовани предикат *chembl_molecule:Xref* открива *cross-reference* податке (Слика 8.3). Објектне вредности овог предиката су заправо веб странице одговарајућих база које се односе на дату супстанцу. Кликом на линк *http://www.drugbank.ca/drugs/DB05706* корисник може открити информације као што су механизам акције (*As an anthracycline, GPX-100 has antimitotic and cytotoxic activity through a number of proposed mechanisms of action: GPX-100 forms complexes with DNA by intercalation between base pairs, and it inhibits topoisomerase II activity by stabilizing the DNA-topoisomerase II complex, preventing the religation portion of the ligation-religation reaction that topoisomerase II catalyzes*) или биолошке мете (*DNA topoisomerase 2-alpha* и *DNA*) које су у интеракцији са датим леком, а које могу утицати на планирање експерименталних приступа. У овом случају можемо приметити да је дато једињење доступно у *online* Drugbank бази података, а није доступно у Drugbank/Bio2RDF онтолошкој бази података. Често се дешава обрнуто, или је случај да подаци нису линковани. Ово је такође један од проблема са којим се корисници сусрећу приликом претраге онтолошких репозиторијума. Предикат *foaf:depiction* који је доступан у оквиру панела приликом примене методе динамичког филтрирања такође је битан јер нуди информацију о хемијској структури лека[106]. Коришћењем неког екстерног алата, као што је QSAR Toolbox Helpdesk[107] и традиционалном методом компарације хемијских структура може се утврдити сличност супстанци. Ово може сугерисати на потенцијалну интеграцију Платформе са другим софтверским решењима.


Dataset:**Chembl/EMBL-EBI**
Show 10 entries
Search:

Drug	moleculeXref
http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL3249110	http://fdasis.nlm.nih.gov/srs/ProxyServlet?mergeData=true&objectHandle=DBMaint&APPLICATION_NAME=fdasrs&actionHandle=default&nextPage=jsp/srs/ResultScreen.jsp&TXTSUPERLISTID=1S9VO1DQG5
http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL3249110	http://pubchem.ncbi.nlm.nih.gov/compound/9829419
http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL3249110	http://pubchem.ncbi.nlm.nih.gov/substance/14934351
http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL3249110	http://www.drugbank.ca/drugs/DB05706
http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL3249110	http://zinc15.docking.org/substances/ZINC000003921098
http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEMBL3249110	https://www.surechembl.org/chemical/SCHEMBL18659500

Showing 1 to 6 of 6 entries Previous 1 Next

Слика 8.3 Резултат примене методе динамичког филтрирања резултата упита за откривање информација о *мест супстанци 2* на Платформи (за селектовани предикат *chembl_molecule:Xref*)

Анализом претходних података извршава се релативно једноставна претрага значајних информација (потенцијалног) лека. Откривене информације представљају круцијални фактор за побољшање биолошких активности, што доприноси ефикаснијим експерименталним истраживањима. Применом метода потенцијално се откривају и информације о активним супстанцама које нису познате широј биоинформатичкој заједници. То може утицати на популаризацију нових лекова и успостављању евентуалне сарадње између истраживача.

Сценарио 2

Један од најважнијих корака у процесу откривања лекова јесте идентификација биолошке мете. Метода извршавања предефинисаних упита на Платформи за шаблон (*a*) *Find targets for the drug*, има за циљ да корисницима омогући идентификацију биолошких мета кроз јавно доступне онтолошке базе података. Резултат извршавања предефинисаног упита овог шаблона за *мест супстанцу 1* јесте листа биолошких мета, односно инстанци одговарајућих онтолошких база података која је представљена у одељку 6.1.3.1 (Слика 6.9). На основу статистичког приказа резултата (Слика 5.5) уочава се да највише података о

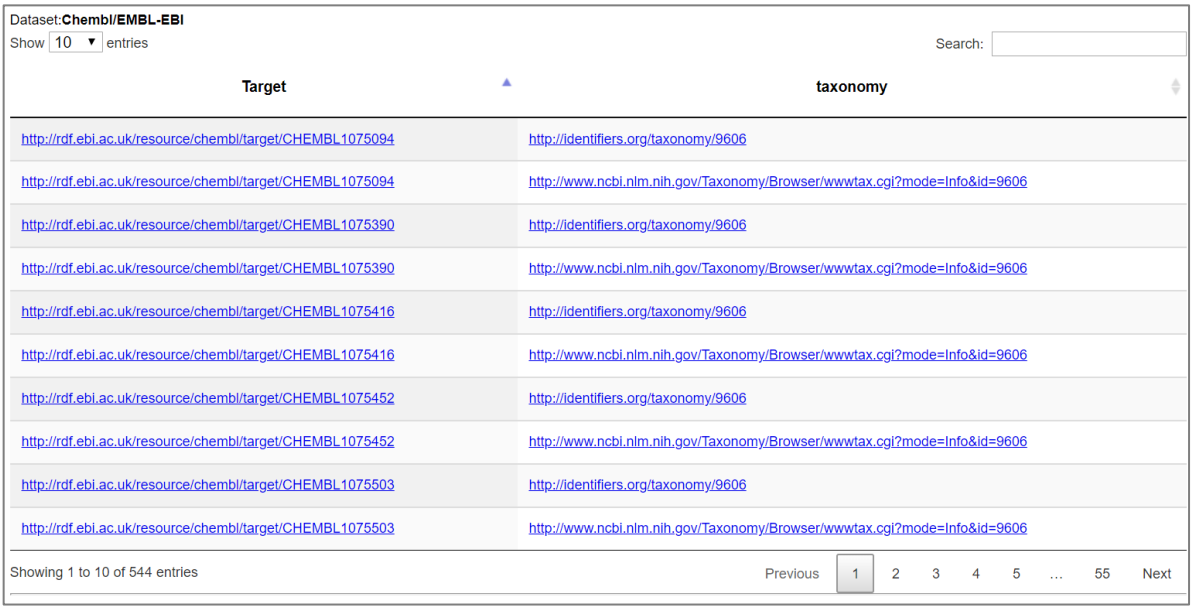
106 *https://www.ebi.ac.uk/chembl/compound/displayimage_large/CHEMBL3249110*

107 *https://qsartoolbox.freshdesk.com/support/home*

биолошким метама потиче из Chembl/EMBL-EBI (284), затим из Drugbank/Bio2RDF (3) и Kegg/Bio2RDF базе података (1). База података PIBAS/CPCTAS у овом случају нуди 1 биолошки систем (*pibas:TargetTest1*) који је тестног типа, док база Bindingdb/Chem2Bio2RDF нема повратних вредности. Постојање релативно великог броја резултата које произилази из EMBL-EBI репозиторијума, последица је тога што се он константно ажурира и што настоји да буде у току са актуелним дешавањима на биоинформатичкој сцени. Ово нарочито важи за Chembl/EMBL-EBI базу података, која се активно користи у многим истраживањима и софтверским решењима у овом домену. Статистички резултат може бити од значаја за корисника јер га потенцијално усмерава на базе података које су „богате“ подацима и које му могу понудити информације од интереса. Federated SPARQL упит за шаблон (*a*) дизајниран је тако да врши селекцију свих мета које су у интеракцији са одговарајућим леком. Додатно, селекција је вршена тако да су биране само мете које су у могућности да успоставе интеракцију са леком - што је постигнуто клаузулом *?ic50value<300000.0* (Слика 6.6 а). Напоменућемо да овај вид филтрирања није био изводљив над Drugbank/Bio2RDF и Kegg/Bio2RDF базама података. Тиме је аутоматски задовољен критеријум да су мете у интеракцији са болешћу и овакав вид селекције сугерише на позитивну идентификацију биолошких мета. Дескрипције биолошких мета обично садрже предикате које указују на дату релацију. На пример, инстанце Kegg/Bio2RDF (Drugbank/Bio2RDF) базе података поседују објектна својства *drugbank_vocabulary:is_target_of* (*kegg_vocabulary:is_target_of*) која упућују на конекцију са *мест супстанцом 1* али и осталим лековима које користе исту биолошку мету за третман лечења канцера. За идентичну сврху инстанца PIBAS/CPCTAS базе података користи предикат *pibas:isTargetOf*. До ових података није лако доћи кроз LODEStar претраживач у Chembl/EMBL-EBI бази података, јер би се иницијално морало приступити неком од есеја кроз предикат *chembl:Activity*, а затим би двоструком применом предиката *chembl:hasActivity* било могуће открити лекове који су у интеракцији са метом. На овај начин корисник може проширити интересовање и открити друге лекове који користе исту биолошку мету. Ово је од важности јер је доказано да се сличне болести лече сличним лековима, а то би потенцијално значило да се могу користити и исте (или сличне) биолошке мете у процесу експериментисања над сличним лековима. Имена биолошких мета такође су доступна кроз одговарајућа уграђена својства *rdfs:label* или *dc:title*: *MOLT-4* (*chembl_target:CHEMBL614177*), *MDA-MB-231* (*chembl_target:CHEMBL400*), *Thymidylate synthase* (*chembl_target:CHEMBL3160, drugbank:BE0000324* и *kegg:5f47d0b54b4d81097410bcc4cf01cf71*), *DNA* (*drugbank:BE0004796*), *RNA* (*drugbank:BE0004810*) итд. Назив биолошке мете *pibas:TargetTest1 je thymidylate synthase* и дефинисан је предикатом *pibas:hasTargetName*. Инстанце Kegg/Bio2RDF и Drugbank/Bio2RDF базе, за разлику од инстанци PIBAS/CPCTAS и Chembl/EMBL-EBI база података, не садрже предикат који наводи тип биолошке мете. На основу доменског знања може се закључити да су инхибитори *thymidylate synthase* заправо хемијски агенси који инхибирају дати ензим и да они имају примену у антиканцерогеној хемотерапији. Слично је и за ДНК и РНК биолошке мете, за које корисник мора знати да означавају нуклеинске киселине. У Chembl/EMBL-EBI бази података својство *chembl:targetType* открива тип биолошке мете, док је у PIBAS/CPCTAS бази података предикат *pibas:tragetType* задужен за ову нотацију. Са друге стране претходно описане конекције биолошких мета са лековима омогућавају проверу да ли лек представља хемијско једињење мале молекуларне тежине или биолошко једињење. На пример, у DrugBank/Bio2RDF бази података *мест супстанца 1* је дефинисана као хемијско једињење (*drugbank_vocabulary:Small-molecule*) молекуларне тежине 130.077194 (овај податак је дефинисан као део објектне вредности *drugbank_vocabulary:calculated-properties* предиката). До сличних информација се може доћи и кроз Chembl/EMBL-EBI базу података. Постојање других лекова (нпр. Gemcitabine [drugbank:DB00441] и Lomustine [drugbank:DB01206] који се користе у терапији лечења канцера) који су везани за предикат *drugbank_vocabulary:is_target_of* и који су типа *small-molecule*, само су додатни показатељи позитивне идентификације таргета. База података Chembl/EMBL-EBI пружа и информације о такозваном месту везивања (енгл. *binding site*), односно локацијама на биолошкој мети на којој се одвија хемијска интеракција са одређеном активном супстанцом (механизам деловања или механизам акције). Ова информација је доступна кроз предикат *chembl:BindingSite*. Место везивања се може дефинисати на различитим нивоима грануларности (на нивоу подјединице, на нивоу протеина и нивоу остатка) и ова карактеристика се користи да покаже хемијску специфичност за типове лиганда који се могу везати, као и афинитет - меру јачине хемијске интеракције. Међутим, овај податак није доступан за све инстанце. На

пример, таргети *chembl_target:CHEMBL61417* и *chembl_target:CHEMBL400* не поседују овакав тип података јер ове инстанце представљају биолошке мете које су типа ћелијских линија. Са друге стране, инстанца *chembl_target:CHEMBL3160* указује на локацију везивања - протеин *thymidylate synthase*. Идентификација биолошке мете се у великој мери темељи и на доменском знању. На основу истраживања [112] може се закључити да *тест супстанца 1* функционише блокирањем активности *thymidylate synthase*, која метилује деоксиуридин монофосфат (dUMP) да би се формирао тимидин монофосфат (dTMP). Ове информације могу утицати на селекцију биолошке мете, односно *thymidylate synthase* може бити потенцијално повољна биолошка мета за даља истраживања. На основу објектних вредности инстанце која представља протеин *thymidylate synthase*, могу се планирати даљи експериментални приступи, који би били квалитетнији и довели до бржих позитивних резултата. Истраживање [229] образлаже како се биолошка мета *thymidylate synthase* експлоатише за даље побољшање. Такође, за једињења која су слична леку, откривање потенцијалних биолошких мета у раној етапи испитивања је врло пожељно, како би се ескивирали нежељени ефекти у клиничким анализама. Дакле, резултат шаблона *Find targtes for the drug* је од круцијалне важности јер указује на то да ли се неко у биоинформатичкој заједници бави сличном тематиком и да ли будућа спроведена експериментална истраживања могу бити од користи за заједницу.

Корисник се у овом сценарију суочава са релативно великим бројем биолошких мета. Треба напоменути да би резултат био већи да није извршено филтрирање за IC_{50} вредност. Применом методе динамичког филтрирања резултата, корисник убрзава процес претраге. На пример, једноставном селекцијом предиката *chembl:targetType* и *pibas:targetType* идентификују се све протеинске биолошке мете (Слика 6.14). Такође, једноставно се може открити таксономија биолошких мета селекцијом *chembl:taxonomy* предиката Chembl/EMBL-EBI базе (Слика 8.4). Како су резултати организовани и по извору података и по типу одабраног предиката, корисник има могућност лакшег сналажења и брже претраге већег броја података. Овакав вид репрезентације резултата омогућава кориснику да самостално пореди резултате између инстанци у оквиру једне или више база података.



Dataset:**Chembl/EMBL-EBI**

Show 10 entries Search:

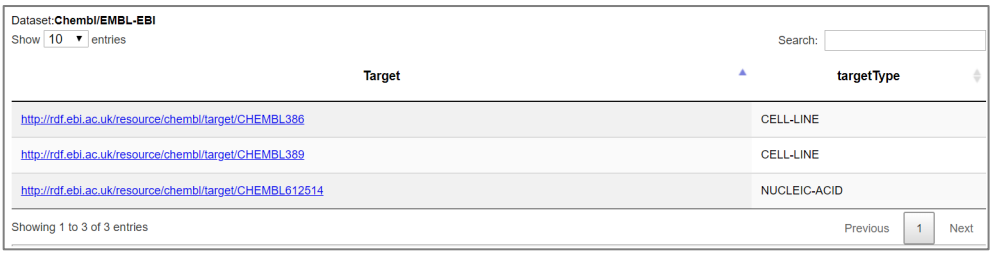
Target	taxonomy
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075094	http://identifiers.org/taxonomy/9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075094	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075390	http://identifiers.org/taxonomy/9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075390	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075416	http://identifiers.org/taxonomy/9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075416	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075452	http://identifiers.org/taxonomy/9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075452	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075503	http://identifiers.org/taxonomy/9606
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL1075503	http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=9606

Showing 1 to 10 of 544 entries Previous 1 2 3 4 5 … 55 Next

Слика 8.4 Резултат примене методе динамичког филтрирања резултата упита за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* на Платформи (за селектовани предикат *chembl:taxonomy*)

Извршавањем предефинисаног упита за *тест супстанцу 2* откривају се следеће биолошке мете: *chembl_target:CHEMBL612514*, *chembl_target:CHEMBL386* и *chembl_target:CHEMBL389*. Оне су потенцијално одговарајуће и корисник се на основу механизма акције (који је утврђен у претходном тестном сценарију) може определити за једну од њих. На основу механизма акције, корисника може занимати и тип биолошких мета. Применом методе динамичког филтрирања резултата (Слика 8.5) за предикат *chembl:targetType* корисник се може одлучити за биолошку мету *chembl_target:CHEMBL612514*

која је типа нуклеинска киселина.


Dataset:Chembl/EMBL-EBI
Show 10 entries
Search:

Target	targetType
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL386	CELL-LINE
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL389	CELL-LINE
http://rdf.ebi.ac.uk/resource/chembl/target/CHEMBL612514	NUCLEIC-ACID

Showing 1 to 3 of 3 entries Previous 1 Next

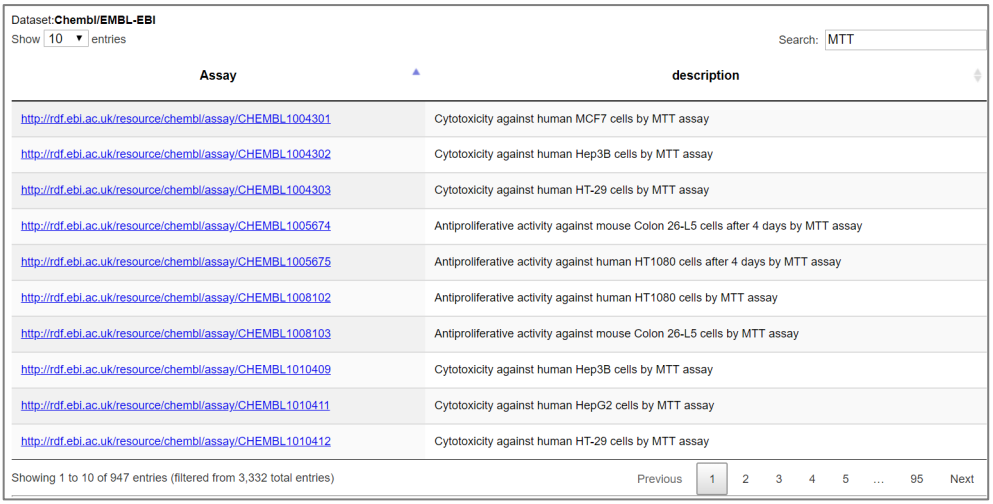
Слика 8.5 Резултат примене методе динамичког филтрирања резултата упита за откривање биолошких мета које су у интеракцији са *тест супстанцом 2* на Платформи (за селектовани предикат *chembl_target:targetType*)

На основу анализе резултата спроведене у овом сценарију може се закључити да методе извршавања предефинисаних упита и динамичког филтрирања резултата за шаблон (*a*) имају могућност да корисницима сервирају информације од интереса у циљу идентификације биолошких мета, које потенцијално могу утицати на процес дизајна лекова.

Сценарио 3

Након селекције биолошке мете и процене исплативости читавог процеса, врши се клонирање мете и развија есеј који омогућава мерење њене активности. Метода извршавања предефинисаних упита на Платформи за шаблон (*b*) шаблон *Find assays for the drug*, има за циљ да омогући идентификацију есеја који су у интеракцији са одговарајућом активном супстанцом. За *тест супстанцу 1* статистика упита је следећа: 1 есеј из CPCTAS/PIBAS и 3332 есеја из Chembl/EMBL-EBI базе. База података PubChem/Chem2Bio2RDF у овом случају нема повратних резултата. Релативно велики број инстанци из Chembl/EMBL-EBI базе децидиран је показатељ колико се различитих есеја може применити за један лек. Као консеквенца следи чињеница да је идентификација есеја прави изазов. Инстанца *pibas:UM05* CPCTAS/PIBAS базе података је у овом случају само тест инстанца и она не подржава дескрипцију. Касније се кроз методу динамичког филтрирања резултата упита (односно кроз селекцију предиката *pibas:name*) може уочити да ова инстанца означава MTT есеј (тестови редукције тетразолијумове боје). Инстанце Chembl/EMBL-EBI базе података садрже детаље кроз *dc:description* предикат. Овај податак је од важности за корисника јер указује на интеракцију између биолошке мете, ћелијске линије и типа есеја. Информација о биолошкој мети која се везује за есеј доступна је и кроз предикат *chembl:hasTarget*. Такође, корисника може занимати и податак који се односи на класификацију есеја (везујући, ADME, функционални или физичко-хемијски) што је доступно кроз предикат *chembl:assayType*. Класификација есеја може подразумевати и верификацију коришћене ћелијске линије, што је одређено предикатом *chembl:assayCellType*.

Метода динамичког филтрирања резултата упита и у овом сценарију олакшава претрагу с обзиром да се корисник суочава са анализом великог броја података. У случају да је корисник заинтересован за есеје које су у интеракцији са биолошком метом *thymidylate synthase*, селекцијом предиката *dcterms:description* и додатним филтрирањем резултата по кључној речи *thymidylate synthase* могу се спознати овакви типови података. До сличних информација се може доћи селекцијом предиката *chembl:hasTarget* и филтрирањем резултата по кључној речи *CHEMBL3160,* што подразумева да корисник има одређено искуство са претходним сценаријем. Такође, због једноставне примене на више узорака и ниских трошкова, MTT есеји су често у фокусу корисника. До информација о овим есејима се може доћи филтрирањем резултата по кључној речи *MTT* (Слика 8.6). Такође, селекција предиката *cco:hasCellLine* и додатно филтрирање резултата по кључној речи *human colon* индицирају на тип ћелијске линије која се може користити за експериментални приступ у лечењу канцера дебелог црева.

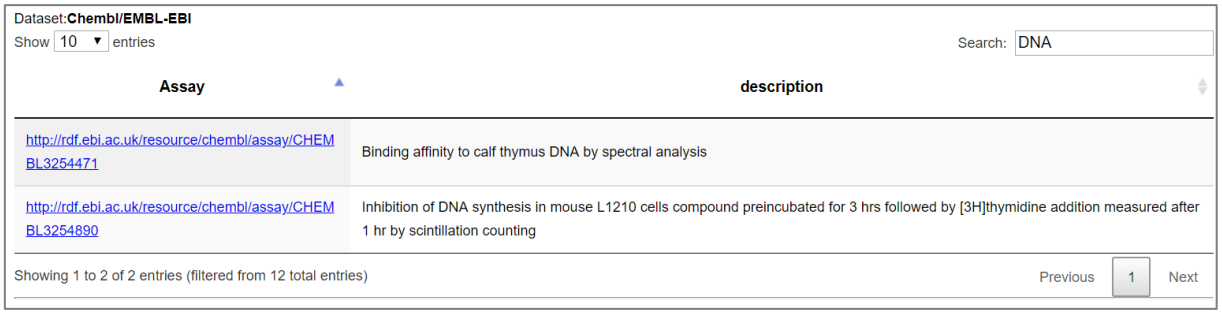

Dataset:**Chembl/EMBL-EBI**
Show 10 entries Search: MTT

Assay	description
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1004301	Cytotoxicity against human MCF7 cells by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1004302	Cytotoxicity against human Hep3B cells by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1004303	Cytotoxicity against human HT-29 cells by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1005674	Antiproliferative activity against mouse Colon 26-L5 cells after 4 days by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1005675	Antiproliferative activity against human HT1080 cells after 4 days by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1008102	Antiproliferative activity against human HT1080 cells by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1008103	Antiproliferative activity against mouse Colon 26-L5 cells by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1010409	Cytotoxicity against human Hep3B cells by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1010411	Cytotoxicity against human HepG2 cells by MTT assay
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL1010412	Cytotoxicity against human HT-29 cells by MTT assay

Showing 1 to 10 of 947 entries (filtered from 3,332 total entries) Previous 1 2 3 4 5 … 95 Next

Слика 8.6 Резултат примене методе динамичког филтрирања резултата упита за откривање есеја који су у интеракцији са леком *Fluorouracil* на Платформи (за селектовани предикат *dcterms:description*), филтриран по кључној речи *MTT*

Резултат извршавања шаблона *Find assays for the drug* за *тест супстанцу 2* обухвата 12 есеја из Chembl/EMBL-EBI базе података. База података PubChem/Chem2Bio2RDF нема повратних резултата. Ако је корисник у претходном сценарију изабрао биолошку мету *chembl_target:CHEMBL612514*, која је по типу нуклеинска киселина, уз помоћ методе филтрирања резултата за предикат *dcterms:description* (и додатним филтрирањем резултата по кључној речи *DNA*) може открити комплементарне информације о типовима есеја који се користе у експерименталним приступима (Слика 8.7). Ово може бити од важности и сугерисати кориснику како да планира сопствена истраживања.


Dataset:**Chembl/EMBL-EBI**
Show 10 entries Search: DNA

Assay	description
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL3254471	Binding affinity to calf thymus DNA by spectral analysis
http://rdf.ebi.ac.uk/resource/chembl/assay/CHEMBL3254890	Inhibition of DNA synthesis in mouse L1210 cells compound preincubated for 3 hrs followed by [3H]thymidine addition measured after 1 hr by scintillation counting

Showing 1 to 2 of 2 entries (filtered from 12 total entries) Previous 1 Next

Слика 8.7 Резултат примене методе динамичког филтрирања резултата упита за откривање есеја који су у интеракцији са *тест супстанцом 2* на Платформи (за селектовани предикат *dcterms:description*), филтриран по кључној речи *DNA*

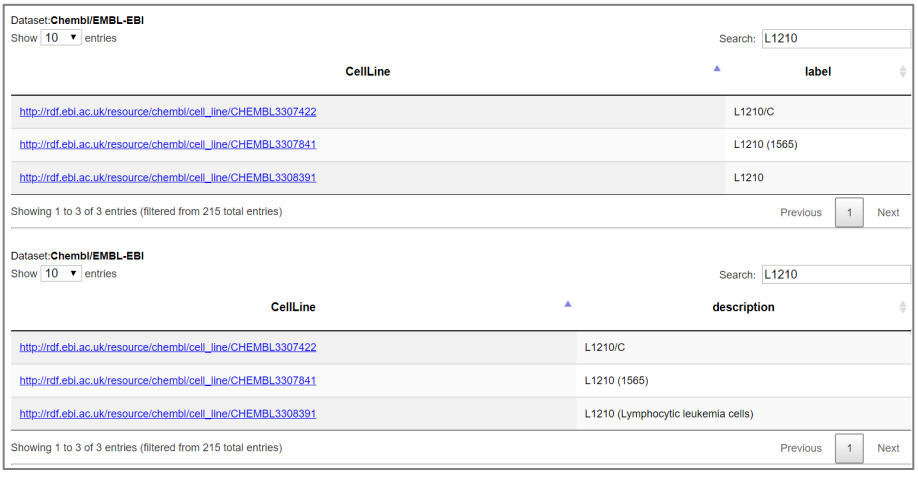
Напоменућемо да је извршавање шаблона (*b*) спроведено за кључну реч која означава *SMILES* параметар активне супстанце. Проблем варирања резултата се у овом случају може појавити због потенцијалне егзистенције више различитих валидних *SMILES* параметара за један лек. На пример, у Chembl/EMBL-EBI и PIBAS/CPCTAS бази података овај параметар за *тест супстанцу 1* има вредност FC1=CNC(=O)NC1=O, док је у PubChem/Chem2Bio2RDF бази података дефинисан са C1=C(C(=O)NC(=O)N1)F. Различите вредности *SMILES* параметара се аутоматски одражавају на резултат упита. На пример, за другу вредност параметра, резултат упита би обухватао само инстанцу *pubchem_bioassay:931*. Такође, могу постојати разлике између канонических и изомерних *SMILES* параметра, што би се аутоматски одразило и на типове предиката који се користе у предефинисаним упитима. На пример, у PubChem/Chem2Bio2RDF бази података се за ову намену користе предикати *pubchem:openeye_iso_smiles* и *pubchem:openeye_can_smiles*. Управо из тих разлога је исправније користити *InChiKey* параметар који јединствено идентификује неку супстанцу. На Платформи је хотимично примењено коришћење *SMILES* параметра како би се указало на различите начине представљања података у домену биоинформатике, као и на проблеме са којима се истраживачи у овом

случају сусрећу, у смислу да резултати претраге могу варирати.

Сценарио 4

Избор ћелијске линије захтева успостављање равнотеже између избора одговарајућег модела и одабира ћелијске линије са којом се могу вршити испитивања. Метода извршавања предефинисаних упита на Платформи за шаблон (*c*) *Find cell lines for drug*, указује на ћелијске линије које се потенцијално могу применити у експерименталном раду. За *тест супстанцу 1* статистика упита је следећа: 1 ћелијска линија из CPCTAS/PIBAS базе података и 215 ћелијских линија из Chembl/EMBL-EBI базе података. Ћелијска линија *pibas:MS05* је тест ћелијска линија. Кроз дескрипције појединачних инстанци Chembl/EMBL-EBI базе података могу се открити ознаке ћелијских линија представљене предикатом *rdfs:label*. Треба напоменути да су ћелијске линије које се користе у Chembl/EMBL-EBI бази података преузете из EFO[108] (*Experimental Factor Ontology*) и CLO[109] (*Cell Line Ontology*) онтологија и као такве можда нису увек адекватно представљене. На пример, термин *H4* се користи да означи ћелијску линију ATCC CRL-1548 (ћелијска линија хепатома пацова) или ћелијску линију ATCC HTB-148 (ћелијска линија хуманих неуроглиома) што може додатно збунити корисника. Ово је још само један од проблема са којима се корисници могу ухватити у коштац приликом процеса претраге и откривања релевантних информација. Кроз предикат *chembl:isCellLineForTarget* може се открити повезаност између ћелијске линије и биолошке мете, а кроз предикат *chembl:isCellLineForAssay* повезаност између есеја и ћелијске линије. Од велике важности је и предикат *chembl:cellosaurusId* који означава интеграцију Chembl/EMBL-EBI и *Cellosaurus*[110] онтологије, која представља велику базу ћелијских линија. Уз помоћ *Cellosaurus* онтологије могу се открити додатне информације, међу којима је и детаљан преглед литературних података.

Метода динамичког филтрирања резултата упита и у овом сценарију олакшава претрагу података. Селекцијом предиката *dcterms:description* и *rdfs:label*, и додатним филтрирањем резултата по кључној речи *L1210* потврђује се постојање ћелијских линија одговарајућег типа. На тај начин се могу открити инстанце *chembl:3307442*, *chembl:3307841* и *chembl:3308931* (Слика 8.8). Слично, селекцијом предиката *chembl:isCellLineForAssay* и филтрирањем по кључној речи *CHEMBL1010412* (на основу претходног сценарија) може се утврдити веза ка MTT есеју. Селекција предиката *chembl:chemblId* омогућава и приступ литературним подацима кроз *Cellosaurus* базу података, па би се корисник могао лакше одлучити за неку од ћелијских линија.



Слика 8.8 Резултат примене методе динамичког филтрирања резултата упита за откривање ћелијских линија које су у интеракцији са леком *Fluorouracil* на Платформи (за селектоване предикате *rdfs:label* и *dcterms:description*), филтриран по кључној речи *L1210*

108 *https://www.ebi.ac.uk/efo/*

109 *https://www.ebi.ac.uk/ols/ontologies/clo*

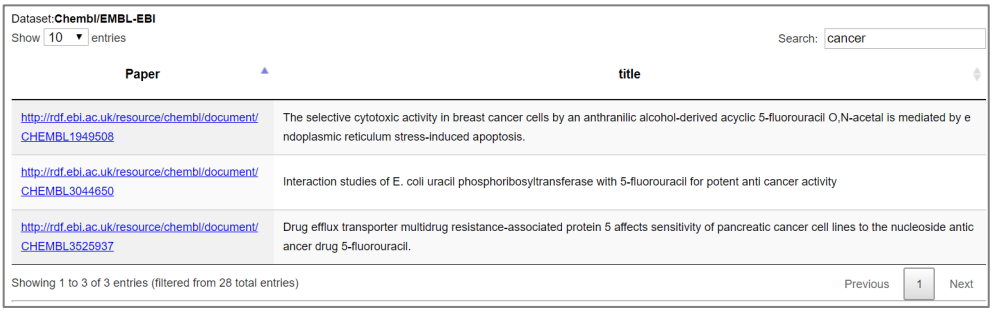
110 *https://web.expasy.org/cellosaurus/*

За *тест супстанцу 2* резултат извршавања предефинисаних упита подразумева две ћелијске линије из Chembl/EMBL-EBI базе података. Ако се корисник у претходном сценарију одлучио за инстанцу *chembl_assay:CHEMBL3254890*, методом динамичког филтрирања резултата упита се за предикат *chembl:isCellLineForAssay* (и додатним филтрирањем по кључној речи *CHEMBL3254890*) добијају информације о ћелијској линији *chembl_cell_line*:*CHEMBL3308391* која се користи за овај есеј (Слика 8.9). Ово је од важности и сугерише кориснику како да планира сопствена истраживања.

Слика 8.9 Резултат примене методе динамичког филтрирања резултата упита за откривање ћелијских линија које су у интеракцији са *тест супстанцом 2* на Платформи (за селектовани предикат *chembl:isCellLineForAssay*), филтриран по кључној речи *CHEMBL3254890*

Сценарио 5

Сваки озбиљан научно-истраживачки рад подразумева и адекватан преглед литературе. Претпоставимо да је корисник заинтересован да открије публикације које су у вези са *тест супстанцом 1*. На Платформи је у оквиру *Research* теме доступан шаблон (*e*) *Find papers with a tittle for the keyword*, који обезбеђује претрагу публикација на основу задате кључне речи. Резултат извршавања овог упита обухвата преко 12000 публикација у PubMed/Bio2RDF и 28 публикација у Chembl/EMBL-EBI бази података. База података Reference/CPCTAS у овом случају нема повратних резултата, што се може оправдати чињеницом да је њено последње ажурирање обављено октобра 2016. године. Приступ дескрипцијама PubMed/Bio2RDF базе података овога пута нема превише ефекта, јер су подаци прилично оскудни. У овом случају се до релевантнијих података може доћи методом динамичког филтрирања. Подаци о наслову публикација доступни су захваљујући предикатима *dc:tittle* и *rdfs:label*. Подаци о аутору се могу открити на основу предиката *pubmed_vocabulary:author*, а подаци о датуму објављивања публикације на основу предиката *pubmed_vocabulary:date-completed* и *dc:date*. Предикат *dc:abstract* омогућава преузимање информација о апстракту публикација. Претпоставимо да је корисник заинтересован само за наслове публикација Chembl/EMBL-EBI базе података. Слика 8.10 садржи резултат примене методе динамичког филтрирања за селектовани предикат *dc:tittle*. Додатно, корисник може бити заинтересован само за публикације које су свом наслову садрже реч *cancer* с обзиром да се лек *Fluorouracil* користи у терапији лечења канцера. У том случају релевантнији подаци се добијају филтрирањем наслова публикација по кључној речи *cancer*.



Слика 8.10 Резултат примене методе динамичког филтрирања резултата упита за откривање публикација које су у интеракцији са леком *Fluorouracil* на Платформи (за селектовани предикат *dc:tittle*), филтриран по кључној речи *cancer*

За кључну реч *HCT-116* шаблон (*e*) обухвата и инстанце Reference/CPCTAS базе података. У случају да је корисник заинтригиран датом базом, он се кроз опцију динамичког филтрирања резултата може одлучити за један од предиката који заправо припадају *dc* и *bibo* речницима (с обзиром да је

Reference/CPCTAS база података настала њиховим наслеђивањем). Предикати ових речника омогућавају претрагу наслова публикације (*dc:tittle*), аутора (*bibo*:*authorList*), страница публикације (*bibo:pages*) и тип саме инстанце (*rdf:type*). Проблем селекције предиката *bibo*:*authorList* и *bibo*:*isPartOf* је тај што би извршавањем звездастих SPARQL упита резултат одговарао такозваном празном чвору (енгл. *blank node - b0*), па овакви подаци не би имали превеликог значаја за кориснике. Сличан проблем може бити присутан и код других база података.

Напоменућемо да клаузула *FILTER regex(?Title, "%s", "i")* у предефинисаном упиту овог шаблона (Слика 6.6 е) узима у обзир сва појављивања кључне речи без обзира да ли се користе велика или мала слова, што је одређено условом "*i*". Међутим, резултат може варирати у зависности од интерпункцијских знакова који се користе у кључној речи. На пример, ако је задата кључна реч *5-Fluorouracil* (која означава синоним речи *Fluorouracil*) онда је статистика резултата следећа: 10193 публикација из Pubmed/Bio2RDF и 26 публикација из Chembl/EMBL-EBI базе података. То би значило да се клаузула FILTER може проширити додатним условом. Ово сугерише да предефинисани упити увек имају простора за модификацију и побољшање.

Сценарио 6

У овом сценарију је представљен значај примене методе додавања кориснички селектоване базе података. За потребе тестирања коришћена је тест база података, а поступак спровођења методе врши се на начин који је дефинисан у одељку 6.3. Додавањем нове базе повећава се потенцијал откривања нових знања. Многе мање познате базе података могу се на овај начин популаризовати и постати доступне на биоинформатичкој сцени. Једна од предности која ова метода нуди јесте и што корисник може извршити мануелно поређење података између већ укључених база података у предефинисаном упиту са својом базом. Ова сазнања могу указивати кориснику да ли се његова истраживања развијају у добром смеру, или га могу упутити на евентуалне експерименталне промене које могу побољшати даља истраживања.

У овом сценарију посматран је значај додавања тест базе података за откривање биолошких мета које су у интеракцији са леком *Fluoruoracil*, односно надовезаћемо се на *Сценарио 2*. Образац који корисник уноси приликом попуњавања *pop-up* форме (Слика 6.15) је у овом случају облика: *?compound testOntology:hasInChiKey "GHASVSINZRGABV-UHFFFAOYSA-N". ?Experiment testOntology:hasCompound ?compound; testOntology:hasTarget ?Target.* Иницијални резултати *Сценарија 2* након додавања тест базе података су обогаћени за биолошку мету *testOntology:TestTarget1*. Претпоставимо да је корисник селектовао предикате *pibas:hasSynonym* (PIBAS/CPCTAS панел)*, drugbank_vocabulary:x-uniprot* (Drugbank/Bio2RDF панел) и *testOntology:hasSynonym* (TestDataset/TestInitiative панел). Мануелним упоређивањем вредности добијених применом методе динамичког филтрирања резултата (Слика 8.11), корисник може закључити да користи сличну биолошку мету (*Thymidylate synthetase*). Како је тест база података интегрисана са Uniprot/EMBL-EBI базом података, предикати *drugbank_vocabulary:x-uniprot* и *testOntology:hasSynonym,* могу се директно искористи за утврђивање сличности између биолошких мета. Такође, метода динамичког филтрирања резултата упита може се експлоатисати у циљу поређења специфичних података између база. Конкретно, у циљу поређења типа биолошке мете могу се поредити вредности предиката *testOntology:hasTargetType* са вредностима предиката *chembl:TargetType*. На овај начин корисник може потврдити да ли се његова истраживања одвијају у добром смеру. Значај методе додавања кориснички селектоване базе података може се тестирати и за претходно спроведене сценарије на сличне начине.


Dataset:**Drugbank/Bio2RDF**
Show 10 entries Search:

Target	drugbank_vocabulary_x_uniprot
http://bio2rdf.org/drugbank:BE0000324	http://bio2rdf.org/uniprot:P04818
http://bio2rdf.org/drugbank:BE0000324	http://bio2rdf.org/uniprot:TYSY_HUMAN
http://bio2rdf.org/drugbank:BE0004796	N/A
http://bio2rdf.org/drugbank:BE0004810	N/A

Showing 1 to 4 of 4 entries Previous 1 Next

Dataset:**TestDataset/TestInitiative**
Show 10 entries Search:

Target	hasSynonym
http://147.91.205.66:2020/Tests/TestOntology#TestTarget1	http://bio2rdf.org/uniprot:Q53Y97

Showing 1 to 1 of 1 entries Previous 1 Next

Слика 8.11 Резултат примене методе динамичког филтрирања резултата упита (за селектоване предикате *drugbank_vocabulary:x-uniprrot* и *testOntology:hasSynonim*) након додавања кориснички селектоване базе података за откривање биолошких мета које су у интеракцији са леком *Fluorouracil* на Платформи

На основу анализе резултата следи да су методе извршавања предефинисаних упита, динамичког филтрирања резултата упита и додавања кориснички селектоване базе података, од изузетне важности за разна биоинформатичка истраживања. Важно је напоменути да се *DataSources* онтологија увек може проширити новим шаблонима, који би допринели квалитетнијим истраживањима, и да се предефинисани упити могу модификовати или проширити новим обрасцима према захтевима истраживача.

8.1.1 Компарација резултата основних метода Платформе са резултатима актуелних софтверских решења

У овом одељку се врши компарација резултата Платформе са резултатима одговарајућих софтверских решења представљених у одељку 6.4. Циљ компарације се своди се на проверу да ли упити на Платформи могу остварити неки комплементарни резултат у поређењу са резултатима добијеним коришћењем других софтверских решења. С обзиром да се на Платформи упити користе за утврђивање интеракције лекова са биолошким метама, есејима и ћелијским линијама или за откривање информација о публикацијама или лековима, било је пре свега неопходно одабрати софтверска решења коју су способна за такав вид поређења. Најпогоднија решења у том смислу била су Open PHACTS [118], BioSearch [111] и BioCarian [116]. Ова решења подржавају базе података које се поклапају са базама података на Платформи. Платформа SPARQLGraph [114] подразумева визуелно креирање упита и зато није била од интереса за овај вид компарације. Софтверско решење QueryMed [110], иако најсличније Платформи, није узето у обзир с обзиром да се заснива на демо приступу. Платформа GFBio [115] није погодна за поређење јер подржава специфичан скуп података, док је BioQueries [117] систем већ искоришћен за проверу предефинисаних упита.

Задатак тестирања у односу на Open PHACTS, BioSearch и BioCarian платформе заснивао се на поређењу биолошких система, ћелијских линија, есеја и литературних података. Евалуација садржи укупно 19 упита (7 упита за биолошке системе, 7 упита за ћелијске линије, 3 упита за есеје и 2 упита за публикације). Свака група упита тестирана је за одређене *InChiKey*/*SMILES* параметре или текстуалне вредности. Улазни параметри су бирани тако да у се обзир узимају супстанце (лекови) који се користе у процесу лечења канцера. Напоменућемо да је за потребе тестирања, за групу упита која се односи на биолошке мете, ћелијске линије и есеје, изостављено филтрирање по IC_{50} вредностима. За Open PHACTS платформу коришћена је REST API верзија v2.28, док су BioSearch и BioCarian тестирани *online*[111].

[111] *http://ws.nju.edu.cn/biosearch/ u http://www.biocarian.com/search*

Табела 8.1 садржи резултате компарација. Ћелије са садржајем *Нема података* означавају да дати шаблони нису могли да се тестирају и пореде са Платформом. У наставку је извршена компарација резултата Платформе са резултатима одговарајућих софтверских решења[112].

Табела 8.1 Компарација резултата рада методе извршавања предефинисаних упита на Платформи са резултатима рада софтверских решења Open PHACTS *[118]*, BioSearch *[111]* и BioCarian *[116]*

Шаблон	Кључна реч	PIBAS FedSPARQL	Open PHACTS	Поклапање	BioSearch	Поклапање	BioCarin	Поклапање
a	**Fluorouracil GHASVSINZRGABV-UHFFFAOYSA-N**	Chembl: 287 Drugbank: 3 Kegg:1 BindingDB:0 PIBAS:1	Chembl: 262	Chembl: 188	Drugbank: 3 Kegg:1	Drugbank:3 Kegg:1	Нема података	Нема података
a	**Cisplatin DQLATGHUWYMOKM-UHFFFAOYSA-L**	Chembl: 0 Drugbank: 1 Kegg:0 PIBAS:1	Chembl: 0	Chembl: 0	Drugbank:1 Kegg:0	Drugbank:1 Kegg:0	Нема података	Нема података
a	**Paclitaxel RCINICONZNJXQF-MZXODVADSA-N**	Chembl: 683 Drugbank: 6 Kegg:1 PIBAS:0	Chembl: 229	Chembl: 201	Drugbank: 6 Kegg:1	Drugbank: 6 Kegg:1	Нема података	Нема података
a	**Cladribine PTOAARAWEBMLNO-KVQBGUIXSA-N**	Chembl:64 Drugbank: 10 Kegg:0 PIBAS: 0	Chembl:170	Chembl:62	Drugbank: 1 Kegg:0	Drugbank: 10 Kegg:0	Нема података	Нема података
a	**Imatinib KTUFNOKKBVMGRW-UHFFFAOYSA-N**	Chembl:1090 Drugbank: 9 Kegg:4 PIBAS: 0	Chembl:525	Chembl:508	Drugbank: 9 Kegg:4	Drugbank: 9 Kegg:4	Нема података	Нема података
a	**Alvocidib BIIVYFLTOXDAOV-YVEFUNNKSA-N**	Chembl:446 Drugbank:10 Kegg:0 PIBAS: 0	Chembl:437	Chembl:431	Drugbank: 0 Kegg:1	Drugbank: 0 Kegg:0	Нема података	Нема података
a	**Thiotepa FOCVUCIESVLUNU-UHFFFAOYSA-N**	Chembl:113 Drugbank: 1 Kegg:1 PIBAS: 0	Chembl:118	Chembl:110	Drugbank: 1 Kegg:1	Drugbank: 1 Kegg:1	Нема података	Нема података
c	**Fluorouracil GHASVSINZRGABV-UHFFFAOYSA-N**	Chembl: 215 PIBAS:1	Chembl: 26	Chembl: 19	Нема података	Нема података	Нема података	Нема података
c	**Cisplatin DQLATGHUWYMOKM-UHFFFAOYSA-L**	Chembl: 0 PIBAS:1	Chembl: 0	Chembl: 0	Нема података	Нема података	Нема података	Нема података
c	**Paclitaxel RCINICONZNJXQF-MZXODVADSA-N**	Chembl: 576 PIBAS:0	Chembl:0	Chembl:0	Нема података	Нема података	Нема података	Нема података
c	**Cladribine PTOAARAWEBMLNO-KVQBGUIXSA-N**	Chembl: 35 PIBAS: 0	Chembl:26	Chembl:25	Нема података	Нема података	Нема података	Нема података
c	**Imatinib KTUFNOKKBVMGRW-UHFFFAOYSA-N**	Chembl: 625 PIBAS: 0	Chembl:386	Chembl:375	Нема података	Нема података	Нема података	Нема података
c	**Alvocidib BIIVYFLTOXDAOV-YVEFUNNKSA-N**	Chembl: 34 PIBAS: 0	Chembl:22	Chembl:21	Нема података	Нема података	Нема података	Нема података
c	**Thiotepa FOCVUCIESVLUNU-**	Chembl: 85 PIBAS: 0	Chembl:83	Chembl:81	Нема података	Нема података	Нема података	Нема података

[112] Резултат извршавања упита на Платформи доступан је на веб адреси *https://figshare.com/articles/Rezultati_izvrsavanja_predefinisanih_upita_za_proces_validacije/7667387.*

	UHFFFAOYSA-N							
b	**Fluorouracil FC1=CNC(=O)NC1=O**	Chembl: 3374 Pubchem: 0 PIBAS: 1	Chembl: 178	Chembl: 3	Kegg: 0 Drugbank: 0	Нема података	Нема података	Нема података
b	**Cisplatin [NH2-].[NH2-].Cl[Pt+2]Cl**	Chembl:0 Pubchem:3 PIBAS: 5	Chembl: 0	Chembl: 0	Kegg: 0 Drugbank: 0	Нема података	Нема података	Нема података
b	**Paclitaxel CC(=O)O[C@H]1C(=O)[C@]2(C)[C@@H](O)C[C@H]3OC[C@@]3(OC(=O)C)[C@H]2[C@H](OC(=O)c4ccccc4)[C@]5(O)C[C@H](OC(=O)[C@H](O)[C@@H](NC(=O)c6ccccc6)c7ccccc7)C(=C1C5(C)C)C**	Chembl:4054 Pubchem:0 PIBAS: 0	Chembl:1179	Chembl: 1040	Kegg: 0 Drugbank: 0	Нема података	Нема података	Нема података
e	**Fluorouracil**	Chembl: 28 Pubmed: 12201 PIBAS: 0	Нема података	Нема података	Kegg: 0	Нема података	Pubmed:3	Pubmed: 1
e	**Cisplatin**	Chembl: 25 Pubmed:18929 PIBAS: 8	Нема података	Нема података	Kegg:4	Нема података	Pubmed:5	Pubmed: 2

У поређењу са Open PHACTS платформом утврђено је да поклапајући резултати долазе из Chembl/EMBL-EBI базе података па су због тога само они (нумерички) представљени у колони која означава поклапање резултата. Може се уочити да постоји велико подударање за групу упита која се односи на биолошке мете, ћелијске линије и есеје. За већину тестних случајева оба решења нуде комплементарне податке. На пример, за лек *Thiotepa* Платформа нуди 113, а Open PHACTS платформа 118 биолошких мета. Поклапање резултата у овом случају је изузетно велико (110 заједничких биолошких мета). Обе платформе нуде комплементарне информације. Постојање комплементарних података на Open PHACTS платформи може се оправдати чињеницом да су ове инстанце присутне у ранијим верзијама Chembl/EMBL-EBI базе, док Платформа подржава актуелну верзију Chembl/EMBL-EBI базе података кроз *remote endpoint*. Из добијених резултата можемо закључити да оба приступа пружају одличну полазну тачку за откривање података, али да је Платформа за нијансу испред Open PHACTS платформе јер приступа ажурним подацима без обзира на ризик од недоступности *remote endpoint*-а. Додатно, Платформа може пружити и комплементарне податке кроз опцију додавања кориснички селектованих база података.

Ако поредимо резултате са BioSearch платформом може се приметити да поклапајући резултати долазе из Drugbank/Bio2RDF и Kegg/Bio2RDF базе података. У случају биолошких мета постоји потпуно поклапање са резултатима Платформе. Са друге стране, BioSearch платформа не нуди информације о ћелијским линијама и есејима. Током процеса тестирања BioSearch платформе уочено је да се подаци о литератури могу открити и у Kegg/Bio2RDF бази података, док Платформа подржава литературне податке кроз PubMed/Bio2RDF базу података. Из добијених резултата може се закључити да оба решења подржавају актуелне податке Bio2RDF репозиторијума. Платформа додатно нуди опцију додавања кориснички селектоване базе података, па самим тим и већи потенцијал комплементарних резултата.

Компарација са BioCarian платформом вршена је у погледу литературних података, јер остали типови шаблона нису подржани. Иако оба решења подржавају PubMed/Bio2RDF базу података не постоји тотално поклапање резултата. То се може објаснити чињеницом да се на Платформи откривају публикације које садрже кључну реч искључиво у наслову, док BioCarian платформа открива публикације које дату кључну реч садрже и у другим параметрима, нпр. апстракту.

Резултати компарације показују да Платформа представљена у дисертацији представља озбиљну полазну

тачку за откривање података значајних за истраживања у домену биоинформатике (рационалног дизајна лекова). Поклапања која постоје међу резултатима указују да Платформа подржава значајне базе података за биоинформатичка истраживања. Такође, актуелни шаблони на Платформи у великој мери се поклапају са истраживањима других приступа, па се може закључити да Платформа у потпуности иде у корак са популарним софтверским решењима у овој области. Додатно, могућност додавања нових база података на Платформи нуди опцију откривања комплементарног знања које доприноси потенцијално ефикаснијим истраживањима.

8.2 Анализа резултата методе за детекцију сличних података

Први вид тестирања алгоритма за детекцију сличних података заснива се на вредностима које су добијене као резултат извршавања предефинисаних упита. Тестирање је спроведено над 17 тестних случајева: 7 за шаблон (*a*), 3 за шаблон (*b*) и 7 за шаблон (*c*). У неким случајевима су за различите шаблоне коришћени исти улазни параметри. Табела 8.2 представља резултат тестирања[113]. Резултати тестирања су подељени по шаблонима. Друга колона је резервисана за кључну реч, док трећа колона представља број података добијених извршавањем предефинисаног упита. Око 80% кључних речи у складу је са облашћу рада Лабораторије и њихова селекција је вршена на основу одлуке експерата. Остале кључне речи су биране насумично. Четврта колона представља број сличних података добијених применом алгоритма. Последња колона је резервисана за валидацију алгоритма, коју извршава особље Лабораторије. Повратне информације експерата су у овом случају јако важне јер се на тај начин одражава њихова улога у процесу онтолошког поравнања. Процес валидације је обављен мануелно и подразумевао је проверу сличности међу подацима који су добијени применом алгоритма. Приликом валидације није се правила разлика између сличних и идентичних података. Валидација је спроведена провером URI спецификација и поређењем њихових објектних вредности. У случају есеја, потврда сличности подразумевала је и проверу да ли се они као група могу применити над датом активном супстанцом. Као подршка за проверу сличности коришћена су софтверска решења canSAR и TDRTargets[114]. Вредност 0 означава слабу сличност или тоталну несличност међу подацима, вредност 1 означава да су сви подаци слични, док вредност 2 означава да су подаци у великом броју слични.. На основу људске процене утврђено је да је алгоритам у већини тестних случајева произвео позитиван резултат. Установљено је да резултат у великој мери зависи од селектованог прага сличности. Између неких објектних вредности сличност је у интервалу од 0.5 до 0.6, што је показано као релативно ниска гранична вредност (у наставку је презентовано и тестирање алгоритма за различите вредности прага сличности). Такође, установљено је да алгоритам није показао позитивне резултате у случајевима када су егзистирали предикати који нису од значаја за поређење (нпр. *dcterms:identifier*, *bio2rdf_vocabulary:namespace* итд.), или кад су инстанце слабије онтолошки представљене па су се поредили искључиво стрингови чија је сличност била мања од 0.52. Ово сугерише да алгоритам може да се побољша неким типом кластеризације текстуалних података за проширење *non_selected_predicates* листе, или је потребно комбиновати CSM са другим техникама. На основу датих чињеница може се закључити да алгоритам нуди обећавајуће решење за проблем детекције сличних података.

Табела 8.2 Резултати валидације методе за детекцију сличних података над улазним параметрима који су добијени као резултат извршавања предефинисаних упита

Шаблон	Кључна реч	Број улазних параметара	Резултат алгоритма	Валидација
a	**Fluorouracil GHASVSINZRGABV-UHFFFAOYSA-N**	289	68	2
a	**Cisplatin DQLATGHUWYMOKM-UHFFFAOYSA-L**	2	2	1
a	**Paclitaxel RCINICONZNJXQF-MZXODVADSA-N**	689	180	2
a	**Cladribine PTOAARAWEBMLNO-KVQBGUIXSA-N**	74	16	2
a	**Phentermine**	9	5	1

[113] Резултат тестирања је доступан на *https://figshare.com/articles/Rezultat_testiranja_algortima_za_detekciju_slicnih_podataka/8080970.*

[114] *http://tdrtargets.org/*

	DHHVAGZRUROJKS-UHFFFAOYSA-N			
a	**Alvocidib** **BIIVYFLTOXDAOV-YVEFUNNKSA-N**	456	397	2
a	**Thiotepa** **FOCVUCIESVLUNU-UHFFFAOYSA-N**	112	16	1
b	**Fluorouracil** **GHASVSINZRGABV-UHFFFAOYSA-N**	216	158	2
b	**Cisplatin** **DQLATGHUWYMOKM-UHFFFAOYSA-L**	1	0	1
b	**Paclitaxel** **RCINICONZNJXQF-MZXODVADSA-N**	575	282	2
b	**Cladribine** **PTOAARAWEBMLNO-KVQBGUIXSA-N**	35	12	2
b	**Imatinib** **KTUFNOKKBVMGRW-UHFFFAOYSA-N**	552	199	2
b	**Alvocidib** **BIIVYFLTOXDAOV-YVEFUNNKSA-N**	34	15	2
b	**Thiotepa** **FOCVUCIESVLUNU-UHFFFAOYSA-N**	82	66	2
c	**N'-[(2,2-dichloro-1-methylcyclopropyl)carbonyl]cyclohexanecarbohydrazide** **CC1(CC1(Cl)Cl)C(=O)NNC(=O)C2CCCCC2**	7	5	1
c	**Cisplatin** **[NH2-].[NH2-].Cl[Pt+2]Cl**	8	2	1
c	**2-[(3-Fluorobenzyl)thio]-3H-imidazo[4,5-c]pyridine** **C1=CC(=CC(=C1)F)CSC2=NC3=C(N2)C=NC=C3**	3	2	1

Други вид валидације подразумевао је тестирање алгоритма за улазне параметре који су бирани на основу одговарајућих релација у PubChem, UniChem, UniProt и canSAR базама података. База података PubChem садржи информације о хемијским структурама, идентификаторима, физичко-хемијским својствима и биолошким активностима молекула, док UniProt база података садржи информације о биолошким метама. База података UniChem представља односе између хемијских структура у оквиру EMBL-EBI репозиторијума, док је база података canSAR интегрисана база података која садржи широк спектар биоинформатичких података укључујући и податке о ћелијским линијама. База података PubChem искоришћена је као основа за тестирање сличности између супстанци (лекова). Сличне супстанце су утврђене на основу својства *Related Records* и *Synonyms*. На пример, уочена је сличност између супстанци *Diamminedichloroplatinum* и *cis-Dichlorodiammineplatinum(II)* чији су идентификатори 2767 и 441203. Ови идентификатори су искоришћени као основа за утврђивање сличности између инстанци *pubchem_compound:2767* и *pubchem_compound:441203* који припадају онтолошкој бази података Pubchem/Chem2Bio2RDF. База података UniChem искоришћена је као основа за тестирање идентичности између инстанци (лекова). На пример, једноставном претрагом по *InChiKey* параметру GHASVSINZRGABV-UHFFFAOYSA-N (лек *Fluorouracil*) утврђена је идентичност између инстанци *drugbank:DB00544* и *kegg:C07649*, које респективно припадају Drugbank/Bio2RDF и Kegg/Bio2RDF базама података. У UniProt бази података сличне биолошке мете су откривене на основу релације *SimilarProteins*. На пример, утврђено да је постоји сличност између биолошких мета *MAP kinase-activated protein kinase 2* и *Mapkapk2 protein* чији су идентификатори P49137 и P49138. Ови идентификатори су преузети као основа за утврђивање сличности између инстанци *uniprot:P49137* и *uniprot:P49138* EMBL-EBI репозиторијума. Такође, односи унутар UniProt базе података су искоришћени и за утврђивање идентичности између биолошких мета. На пример, уочена је идентичност између инстанци *uniprot:P49137* и *chembl_target:CHEMBL2208* EMBL-EBI репозиторијума. У canSAR бази података сличне ћелијске линије су откривене на основу својства *similar_cell_lines*. Сличност је у овом случају класификована по различитим категоријама укључујући сличност према мутацији гена, према профилу експресије гене и према профилу осетљивости на лекове. На пример, ћелијска линија HCT-116[115] је према првој категорији сличности, слична ћелијској линији *MDA-MB-231*. Ове ознаке су искоришћене за утврђивање сличности између инстанци *chembl_cell_line:CHEMBL3308372* и *chembl_cell_line:CHEMBL3307960* Chembl/EMBL-EBI базе података. Табела 8.3 садржи резултате тестирања. Прва колона је резервисана за улазне параметре који су груписани по типовима података - лековима, биолошким метама и ћелијским линијама. Ознака **И** означава идентичне, ознака **С** сличне, док

[115] *https://cansar.icr.ac.uk/cansar/cell-lines/HCT-116/similar_cell_lines/*

ознака **Н** означава инстанце које немају особину сличности (идентичности). Ове ознаке су утврђене на основу релација у поменутим базама. Друга колона представља резултат примене алгоритма, док трећа колона означава валидацију рада алгоритма на основу мишљења експерата Лабораторије - вредност 1 представља позитиван исход алгоритма, а вредност 0 неуспех. У складу са резултатима може се утврдити да је тачност рада алгоритма релативно висока. Конкретно, алгоритам није потврдио сличност инстанци *kegg:C07666* и *drugbank:DB00250*, као и несличност инстанци *pubchem:2767* и *pubchem:441203*. Код првог пара инстанци утврђено да је сличност између парова стрингова *Dapsone 4,4 Sulfonylbisbenzenamine Diaphenylsulfone kegg C07666* и *Dapsone* испод прага сличности - 0.41, док је код другог пара инстанци сличност између стрингова *MolPort-003-983-464* и *MolPort-003-983-423* изнад прага сличности - 0.75. У последњем случају алгоритам је негативно валидиран јер су инстанце представљене са релативно мање предиката, односно информација од значаја. За консеквенцу, то је подразумевало поређење само објектних вредности које се односе на идентификаторе (нпр. *NCI-H322M* и *CW-2*) код којих није уочена сличност. Ови резултати тестирања указују на несавршеност алгоритма, чиме се оставља довољно простора за његово побољшање: алгоритам би могао да се комбинује са неким другим мерама сличности; могу се применити још радикалније методе претпроцесирања текста; потребно је размотрити различите вредности за праг сличности, итд.

Табела 8.3 Резултати валидације алгоритма за детекцију сличних података за улазне параметре селектоване на основу релација у PubChem, UniProt, UniChem и canSAR базама података

Улазни параметри		**Резултат примене алгоритма**	**Валидација рада алгоритма**
Супстанце (лекови)			
kegg:C07649 drugbank:DB00544	**И**	kegg:C07649 drugbank:DB00544	1
drugbank:DB00515 pubchem:2767	**И**	drugbank:DB00515 pubchem:2767	1
drugbank:DB01229 pubchem:36314	**И**	drugbank:DB01229 pubchem:36314	1
drugbank:DB01229 chembl_molecule:CHEMBL428647	**И**	drugbank:DB01229 chembl_molecule:CHEMBL428647	1
kegg:C07666 drugbank:DB00250	**И**	Нема сличних података!	0
pubchem:16213520 pubchem:3385	**С**	pubchem:16213520 pubchem:3385	1
pubchem:2767 pubchem:24191118	**С**	pubchem:2767 pubchem:24191118	1
pubchem:2767 pubchem:6712951	**С**	pubchem:2767 pubchem:6712951	1
pubchem:2767 pubchem:5702198	**С**	pubchem:2767 pubchem:5702198	1
pubchem:2767 pubchem:5460033	**С**	pubchem:2767 pubchem:5460033	1
pubchem:2767 pubchem:84691	**С**	pubchem:2767 pubchem:84691	1
pubchem:2767 pubchem:441203	**С**	pubchem:2767 pubchem:441203	1
pubchem:2767 pubchem:441204	**Н**	pubchem:2767 pubchem:441203	0
chembl_molecule:CHEMBL428111 pubchem:441212	**Н**	Нема сличних података!	1
pubchem:2768 pubchem:421213	**Н**	Нема сличних података!	1
kegg:C07122 drugbank:DB00466	**Н**	Нема сличних података!	1
Биолошке мете			
uniprot:P04818 chembl_target:CHEMBL3160	**И**	uniprot:P04818 chembl_target:CHEMBL3160	1
uniprot:P04818 drugbank:BE0000324	**И**	uniprot:P04818 drugbank:BE0000324	1
uniprot:P04818 chembl_target:CHEMBL1952	**И**	uniprot:P04818 chembl_target:CHEMBL1952	1
uniprot:P49137 uniprot:P49138	**С**	uniprot:P49137 uniprot:P49138	1
uniprot:P04818 uniprot:Q53Y97	**С**	uniprot:P49137 uniprot:Q53Y97	1
kegg:K00560 uniprot:P04818	**С**	kegg:K00560 uniprot:P04818	1
kegg:K00560 chembl_target:CHEMBL3587	**Н**	Нема сличних података!	1
chembl_target:CHEMBL2208	**Н**	Нема сличних података!	1

chembl_target:CHEMBL3587 chembl_target:CHEMBL4040 chembl_target:CHEMBL614245			
uniprot:P04818 chembl_target:CHEMBL2208	**H**	Нема сличних података!	1
Ћелијске линије			
chembl_cell_line:CHEMBL3307945 chembl_cell_line:CHEMBL3308372	**C**	chembl_cell_line:CHEMBL3307945 chembl_cell_line:CHEMBL3308372	1
chembl_cell_line:CHEMBL3307710 chembl_cell_line:CHEMBL3308161	**C**	Нема сличних података!	0
chembl_cell_line:CHEMBL3307710 chembl_cell_line:CHEMBL3307960	**H**	Нема сличних података!	1
chembl_cell_line:CHEMBL3308723 chembl_cell_line:CHEMBL3308729	**H**	Нема сличних података!	1

Трећи вид тестирања алгоритма подразумевао је тестирање алгоритма без корака селекције предиката. Тестирање без селекције предиката подразумева да су у обзир узети сви предикати улазних параметара који имају стринг објектне вредности, односно да није извршен процес одређивања тежине термина, који обавља *TF-IDFMeasure* оператор. Табела 8.4 представља резултате тестирања алгоритма за оба приступа. Прва колона представља улазне параметре, док друга колона означава резултат рада алгоритма за оба приступа (вредност 1 означава да постоји сличност између података, док вредност 0 означава супротно). Улазни параметри су подељени у три категорије (лекови, биолошке мете и ћелијске линије) и селектовани тако да између њих не постоји реална сличност (идентичност).

Табела 8.4 Резултати примене алгоритма за детекцију сличних података за два основна приступа: са селекцијом и без селекције предиката

Улазни параметри		**Резултат примене алгоритма**	
		Приступ без селекције предиката	**Приступ са селекцијом предиката**
Супстанце (лекови)			
kegg:C07641 drugbank:DB00466	**H**	0	0
drugbank:DB00515 pubchem:2768	**H**	0	0
drugbank:DB01229 pubchem:3385	**H**	0	0
pubchem:2767 pubchem:2768	**H**	0	0
pubchem:2767 pubchem:2769	**H**	0	0
pubchem:2768 pubchem:84691	**H**	0	0
pubchem:2767 pubchem:441205	**H**	0	0
pubchem:2768 pubchem:441203	**H**	0	0
chembl_molecule:CHEMBL2068237 pubchem:36314	**H**	0	0
pubchem:2768 pubchem:421213	**H**	0	0
kegg:C07122 drugbank:DB00466	**H**	0	0
Биолошке мете			
chembl_target:CHEMBL2208 chembl_target:CHEMBL3587	**H**	1	0
uniprot:P49137 chembl_target:CHEMBL3587	**H**	0	0
Ћелијске линије			
chembl_cell_line:CHEMBL3307710 chembl_cell_line:CHEMBL3307960	**H**	1	0
chembl_cell_line:CHEMBL3308723 chembl_cell_line:CHEMBL3308729	**H**	1	0
chembl_cell_line:CHEMBL3308723 chembl_cell_line:CHEMBL3308727	**H**	1	0

На основу резултата може се закључити да приступ без селекције предиката не даје добре резултате уколико улазни параметри припадају ChEMBL/EMBL-EBI бази података. Ово се може објаснити чињеницом да одређени предикати, који се користе за представљање многих инстанци у ChEMBL/EMBL-EBI бази података, имају идентичне објектне вредности. На пример, велики број биолошких мета располаже својством *chembl:organismName*, које представља изворни организам

биолошког система. Објектна вредност овог својства често је представљена са *Homo sapiens* и уколико би се она користила у процесу поређења инстанци, то би аутоматски имало позитиван ефекат на њихову сличност. Због тога је уведена листа *non_selected_predicates*. Спроведено тестирање указује на то да је корак селекције предиката неизбежан за успешну детекцију сличних података.

Као што је већ поменуто, успешност алгоритма зависи и од прага сличности. Алгоритам тренутно користи вредност 0.52, која у великом броју случајева утиче на његов позитиван исход. Међутим, варијације у прагу сличности могу имати различите ефекте на коначни резултат. Табела 8.5 представља резултат тестирања алгоритма за различите вредности прага сличности над улазним параметрима ChEMBL/EMBL-EBI базе података.

Табела 8.5 Резултати примене алгоритма за детекцију сличних података за улазне параметре селектоване из ChEMBL/EMBL-EBI базе података и различите вредности прага сличности (*th*)

Улазни параметри		**Резултат примене алгоритма**			
		Праг сличности $th \leq 0.6$	**Праг сличности** $th \leq 0.7$	**Праг сличности** $th \leq 0.8$	**Праг сличности** $th \leq 0.9$
Биолошке мете					
chembl_target:CHEMBL2208 chembl_target:CHEMBL3587 chembl_target:CHEMBL3221	**H**	0	0	0	0
chembl_target:CHEMBL4526 chembl_target:CHEMBL4696 chembl_target:CHEMBL5836	**C**	CHEMBL4526 CHEMBL5836	CHEMBL4526 CHEMBL5836	CHEMBL4526 CHEMBL5836	0
chembl_target:CHEMBL614917 chembl_target:CHEMBL614919 chembl_target:CHEMBL614922 chembl_target:CHEMBL614923	**C**	CHEMBL614917 CHEMBL614919 CHEMBL614922	0	0	0
chembl_target:CHEMBL2094134 chembl_target:CHEMBL2111354 chembl_target:CHEMBL2111464	**C**	CHEMBL2094134 CHEMBL2111354 CHEMBL2111464	CHEMBL2094134 CHEMBL2111354 CHEMBL2111464	CHEMBL2094134 CHEMBL2111354 CHEMBL2111464	CHEMBL2094134 CHEMBL211135 CHEMBL2111464
Ћелијске линије					
chembl_cell_line:CHEMBL3307710 chembl_cell_line:CHEMBL3307960 chembl_cell_line:CHEMBL3307761	**H**	0	0	0	0
chembl_cell_line:CHEMBL3307743 chembl_cell_line:CHEMBL3307744 chembl_cell_line:CHEMBL3308494	**C**	CHEMBL3307743 CHEMBL3307744 CHEMBL3308494	CHEMBL3307743 CHEMBL3307744 CHEMBL3308494	0	0
chembl_cell_line:CHEMBL3307447 chembl_cell_line:CHEMBL3307450 chembl_cell_line:CHEMBL3307742 chembl_cell_line:CHEMBL3307743	**C**	CHEMBL3307447 CHEMBL3307450 CHEMBL3307742 CHEMBL3307743	CHEMBL3307447 CHEMBL3307450 CHEMBL3307742 CHEMBL3307743	CHEMBL3307447 CHEMBL3307450 CHEMBL3307742 CHEMBL3307743	0
chembl_cell_line:CHEMBL3308372 chembl_cell_line:CHEMBL3308494 chembl_cell_line:CHEMBL3308552 chembl_cell_line:CHEMBL3308836 chembl_cell_line:CHEMBL3308848	**C**	CHEMBL3308552 CHEMBL3308836 CHEMBL3308848	CHEMBL3308552 CHEMBL3308836 CHEMBL3308848	CHEMBL3308552 CHEMBL3308848	0

8.2.1 Компарација алгоритма за детекцију сличних података на Платформи са алгоритмима актуелних решења

Поравнање инстанци је прилично комплексан процес и истраживачи покушавају да се на различите начине ухвате у коштац са овим проблемом. Без обзира на приступ (програмско решење) који се одабере, постоји могућност да резултати нису очекивани, па је неретко неопходно извршити модификацију изворног кода како би се произвели жељени резултати. Такође, проблем могу бити и улазни подаци, јер се у зависности од изабраног приступа захтева другачији вид улазних параметара. На пример, приступ дефинисан у [198] као улазне параметре захтева онтологије у N-Triples формату, док су решења [131,205,212] захтевала онтологије у OWL формату. Такође, да би се у неким приступима [131,205,212] омогућило поравнање инстанци било је неопходно користити *owl:Class* синтаксу у циљу њиховог представљања. Међутим, како већина *endpoint*-а користи *rdf:Description* синтаксу за репрезентацију инстанци, то је читав процес захтевао трансформацију улазних података, што је подразумевало велики утрошак временских ресурса. Улазни параметри (онтолошки фајлови) су углавном креирани

комбинацијом мануелног приступа и применом CONSTRUCT SPARQL упита[116]. У случају Платформе се као улазни параметри прослеђују URI спецификације заједно са њиховим *endpoint*-има, што се сматра побољшањем, јер се од корисника не очекује превелико ангажовање за припрему улазних података.

На основу литературних података представљених у одељку 7.4 дошло се до закључка да су истраживања [131,212,205,198] најпогоднија за поређење са алгоритмом представљеним на Платформи. Софтверско решење PARIS [198] је прикладно за компарацију, с обзиром да је било једноставно припремити улазне податке и извршити процес поређења инстанци без додатног мануелног подешавања. Међутим, ово решење се заснива на приступу вероватноће, а од интереса је извршити компарацију са решењима која подржавају лексичке приступе или приступе векторске репрезентације речи за поравнање инстанци. Због тога су решења [205,212] која подржавају лексичке приступе и истраживање [131] које подржава векторску репрезентацију речи, од великог значаја за процес компарације. Међутим, нека софтверска решења, иако прикладна за процес компарације (због примене лексичког поравнања [213,211] или векторске репрезентације речи [219]) нису применљива због своје недоступности. Такође, неки приступи нуде искључиво псеудо-код [132], што би подразумевало утрошак ресурса у циљу имплементације програмског кода. Табела 8.6 представља преглед резултата рада алгоритма представљеног на Платформи и алгоритама представљених у истраживањима [131,212,205,198]. Прва колона садржи примере улазних параметра, док остале колоне представљају резултате рада појединачних решења.

Табела 8.6 Преглед резултата рада алгоритма за детекцију сличних података представљеног у дисертацији и алгоритама представљених у истраживањима *[131,212,205,198]*

Улазни параметри		**Алгоритам Платформе**	**OntoEmma**	**Hertuda**	**AML**		**PARIS**
					word-matcher	lexical-matcher	
Супстанце (лекови)							
kegg:C07649 drugbank:DB00544	**И**	1	0	0	0	0	0
drugbank:DB00515 pubchem_compound:2767	**И**	1	0	0	0	0	0
drugbank:DB01229 pubchem_compound:36314	**И**	1	0	0	0	0	0
drugbank:DB01229 chembl_molecule:CHEMBL428647	**И**	1	0	0	0	0	1
kegg:C07666 drugbank:DB00250	**И**	0	0	0	0	0	0
pubchem_compound:16213520 pubchem_compound:3385	**С**	1	0	0	0	1	1
pubchem_compound:2767 pubchem_compound:24191118	**С**	1	0	0	0	1	1
pubchem_compound:2767 pubchem_compound:6712951	**С**	1	0	0	0	1	0
pubchem_compound:2767 pubchem_compound:5702198	**С**	1	0	0	0	1	1
pubchem_compound:2767 pubchem_compound:5460033	**С**	1	0	0	0	1	1
pubchem_compound:2767 pubchem_compound:84691	**С**	1	0	0	0	1	1
pubchem_compound:2767 pubchem_compound:441203	**С**	1	0	0	0	1	1
chembl_molecule:CHEMBL428111 pubchem_compound:441212	**Н**	1	Нема резултата	Нема резултата	Нема резултата		Нема резултата
pubchem_compound:2768 pubchem_compound:421213	**Н**	1	Нема резултата	Нема резултата	0	1	Нема резултата
kegg:C07122 drugbank:DB00466	**Н**	1	0	0	0	0	0
Биолошке мете							
uniprot:P04818 chembl_target:CHEMBL3160	**И**	1	0	0	1	1	1
uniprot:P04818 drugbank:BE0000324	**И**	1	0	0	0	1	1
uniprot:P04818 chembl_target:CHEMBL1952	**И**	1	0	0	1	1	1
uniprot:P49137 uniprot:P49138	**С**	1	0	0	1	1	1

116 Пример једног таквог фајла дат је на адреси *https://figshare.com/articles/Kegg_Instance/7667792*.

uniprot:P04818 uniprot:Q53Y97	**C**	1	1	1	1	1	1
kegg:K00560 uniprot:P04818	**C**	1	0	0	0	1	0
kegg:K00560 chembl_target:CHEMBL3587	**H**	1	0	0	0	0	0
chembl_target:CHEMBL2208 chembl_target:CHEMBL3587 chembl_target:CHEMBL4040 chembl_target:CHEMBL614245	**H**	1	Нема резултата	Нема резултата	Нема резултата		1
uniprot:P04818 chembl_target:CHEMBL22080	**H**	1	0	0	0	0	0
Ћелијске линије							
chembl_cell_line:CHEMBL3307945 chembl_cell_line:CHEMBL3308372	**C**	1	1	1	0	0	1
chembl_cell_line:CHEMBL3307710 chembl_cell_line:CHEMBL3307960	**H**	1	0	0	0	0	1
chembl_cell_line:CHEMBL3308723 chembl_cell_line:CHEMBL3308729	**H**	1	0	0	0	0	1

У компарацији са приступом које нуди истраживање [131] (OntoEmma пројекат), алгоритам Платформе пружа квалитетније резултате. Главни разлог је тај што OntoEmma утврђује искључиво еквивалентност између ентитета. Овај приступ користи *idf* меру како би се извршила селекција ентитета који ће користи за процес поравнања. Слично томе, алгоритам представљен на Платформи користи *tf-idf* меру за процес селекције предиката који ће даље утицати на процес онтолошког поравнања. У случају компарације ентитета OntoEmma алгоритам је ограничен на унапред одређене предикате, док алгоритам Платформе настоји да одреди и додатне предикате, који могу утицати на поузданији и квалитетнији процес поравнања. Кључна разлика је и у примени модела векторског простора. OntoEmma користи *word2vec* модел, док алгоритам Платформе користи *bag-of-words* принцип. Генерално, резултати OntoEmma приступа указују само на детекцију еквивалентних (идентичних) инстанци, а не и на њихову сличност, чиме се потврђује надмоћност алгоритма на Платформи.

У поређењу са приступом које нуди софтверско решење Hertuda [212], алгоритам на Платформи има знатно боље учинке. За то постоје два суштинска оправдања. Прво оправдање је то што Hertuda пореди искључиво објектне вредности *rdfs:label* и *rdfs:comment* предиката, док алгоритам на Платформи поред стандардизованих предиката (*rdfs:label* и *dc:tittle*), тежи да додатно прошири листу потенцијалних предиката за поређење. На пример, објектне вредности предиката *rdfs:label* инстанци *drugbank:DB00515* и *pubchem_compound:2767* су респективно, *Cisplatin* и *2767*. Поређењем ових објектних вредности, ниједан од датих система није дао позитиван резултат. Међутим, алгоритам Платформе користи и предикат *pubchem*:*synonyms* чиме се проширује листа кандидата за процес поравнања. Једна од објектних вредности овог предиката јесте вредност *cisplatin*, чиме се аутоматски успоставља релација сличности између инстанци *drugbank:DB00515* и *pubchem_compound:2767.* Друго важно оправдање је то што Hertuda има способност детекције искључиво еквивалентних (идентичних) ентитета што је одређено прагом сличности - вредност 1.0, чиме се аутоматски одбацују ентитети код којих је уочена „слабија“ сличност. На Платформи то није случај, јер је праг сличности прописан вредношћу 0.52, чиме се обезбеђује потенцијал детекције не само идентичних, већ и сличних ентитета. Hertuda је генерално произвела позитивне резултате уколико су се поредиле инстанце истог именског простора и уколико су објектне вредности предиката *rdfs:label* представљале искључиво називе инстанци, а не идентификаторе (ID ознаке). На пример, инстанце *uniprot:P04818* и *uniprot:Q53Y97* деле исту објектну вредност (TYMS) подразумеваног предиката (*rdfs:label*) због чега је регистрована сличност, док сличне инстанце *pubchem_compound:16213520* и *pubchem_compound:3385* користе идентификаторе (16213520 и 3385) због чега није утврђена сличност. Иначе, Hertuda се може применити само над паровима онтологија, тако да није могуће поредити три или више онтологија истовремено. На Платформи је овај приступ знатно лакши, јер је дозвољен произвољан број улазних параметара (URI спецификација).

У случају AML [205] софтверског решења тестиране су *word-matcher* и *lexical-matcher* технике за праг сличности са вредношћу од 0.52. Генерално, AML приступ је обезбедио квалитетније резултате у случају *lexical-matcher* технике, а слабије у случају *word-matcher* технике. Успех прве технике је свакако био очекиван с обзиром да су неки стрингови делили групе истих речи (нпр. *thyA, TYMS; thymidylate synthase [EC:2.1.1.45] [kegg:K00560]* и *TYMS*) или су постојали одговарајући акроними који су могли да се

мечирају са пуним називима (нпр. *TYMS* и *Thymidylate synthase*). Интересантно да је ова техника продуковала позитивне резултате и у случају да је поредила парове идентификатора, што се може оправдати применом тежинске JM. Техника *word-matcher* је произвела позитивне резултате уколико су објектне вредности предиката *rdfs:label* биле идентичне. У поређењу са алгоритмом Платформе, обе технике постижу слабије резултате. То се може образложити чињеницом да *word-matcher* техника користи искључиво *rdfs:label* предикате, док алгоритам на Платформи, осим *rdfs:label* и *dc:tittle* предиката, користи и екстерне предикате који могу побољшати процес поравнања. Други узрок може бити и тај да је тежинска JM техника произвела слабије резултате него CSM. Такође, AML технике се могу применити само над паровима онтологија, тако да није било могуће поредити три или више онтологија истовремено.

У поређењу са приступом које нуди PARIS [198], алгоритам Платформе обезбеђује боље резултате. На пример, PARIS није имао позитиван исход за идентичне инстанце *kegg:C07649* и *drugbank:DB00544*, иако се наслућује супротно. Ово се може објаснити чињеницом да онтологије које представљају ове инстанце не користе исте релације за моделовање или су одређена објектна својства за исте релације (предикате) представљена на различите начине. На пример, инстанце *kegg:C07649* и *drugbank:DB00544* деле предикате *rdfs:label* и *dc:tittle*, али су њихове објектне вредности другачије представљене. За прву инстанцу објектне вредности ових предиката су *Fluorouracil [drugbank:DB00544]* и *Fluorouracil*, а за другу *5-FU; Fluorouracil; 5-Fluorouracil [kegg:C07649]* и *5-FU; Fluorouracil; 5-Fluorouracil*. Такође, неке релације су различито моделоване. На пример, *SameAs* релација је у Kegg/Bio2RDF бази података моделована са *kegg:same-as*, а у DrugBank/Bio2RDF са *owl:sameAs*. Са друге стране Платформа не користи овакав вид ограничења, већ пореди објектне (стринг) вредности различитих предиката, без обзира на њихов начин моделовања, што се сматра предношћу.

8.3 Ограничења Платформе

Платформа представљена у дисертацији настоји да кроз резултате својих метода пружи информације које су значајне за планирање будућих истраживања у домену рационалног дизајна лекова. Ипак, корисници се повремено могу суочити са одређеним проблемима приликом њеног коришћења. Један од проблема јесте потреба да се поседује доменско знање приликом анализе резултата. Проблем анализе података делимично је решен комбинацијом одговарајућих приступа за визуелизацију података - увођењем једноставног иницијалног корисничког интерфејса, увођењем табеларних приказа (уз опцију сортирања и претраге) и панела (*accordion*) елемената. Други проблем јесте познавање семантичких технологија (SPARQL синтаксе) приликом додавања кориснички селектованог скупа података. Превазилажење овог недостатка није нимало једноставно и захтева велики напор у циљу екстракције података за унапред одређене типове упита. Приликом истраживања која су обављена за потребе дисертације, није откривено ниједно решење које може аутоматски да генерише упите за специфичне задатке над неком кориснички одабраном базом података, већ се кориснички селектоване базе података најчешће претражују по унетој кључној речи, чиме се претрага каналише у образац облика *? s ? p "key_word"* . Међутим, како овакав приступ није од превеликог значаја за потребе специфичних упита које се изводе на Платформи (јер је превише уопштен), тако да метода додавања кориснички селектоване базе података захтева познавање семантичких технологија. Потреба доменског знања и познавање семантичких технологија генерално су проблеми са којим се суочавају и све апликације које су представљене у прегледу литературе (одељак 6.4).

Платформа се суочава и са одређеним ограничењима која могу утицати на њене перформансе. Један од главних дефицита јесте могућа недоступност или блокираност *remote endpoint*-а. О узроцима овог проблема било је више речи у одељку 3.6.2. Иако је могуће да се овај недостатак превазиђе креирањем локалних копија одговарајућих *endpoint*-а, постоји шанса да се на тај начин корисници лише ажурних података. Са друге стране, креирање локалних копија одговарајућих база је захтеван процес, јер се често величине одређених база података мере у терабајтима. Такође, локални SPARQL сервери се могу суочити и са проблемима складиштења података. На пример, SPARQL сервери (JOSEKI и FUSEKI) имају проблем чувања фајлова који користе N3 синтаксу (која је често коришћена у базама актуелних

репозиторијума), па су неопходне одређене конфигурационе промене. Тежећи да буде у току са актуелним подацима, Платформа је пристала на ризик коришћења *remote endpoint*-а. Зато се као мера предострожности у предефинисаним Federated SPARQL упитима користи кључна реч SILENT, која спречава грешке приликом преузимања података са недоступног *endpoint*-а.

Време извршавања SPARQL упита на Платформи такође је један од потенцијалних проблема. Примећено је да се кроз прву итерацију сваки предефинисани упит увек дуже извршава, а да је свака следећа итерација (извршавање истог упита за исте параметре) бржа, вероватно због кеширања података на претраживачу. Такође, време извршавања упита условљено је и проблемом недоступности *remote endpoint*-а, али може зависити и од брзине корисникове интернет конекције, као и брзине интернет конекције између Платформе и *remote endpoint*-а. Брзина упита би се могла повећати ако би се користиле локалне копије база података, које би биле доступне кроз упит применом резервисане речи FROM. Брзина упита на тај начин директно утиче на време које се улаже за извршавање свих метода на Платформи. Треба напоменути и да време извршавања алгоритма за детекцију сличних података расте линеарно са бројем улазних параметара (с обзиром на учестало извршавање SPARQL упита за одговарајуће инстанце) и бројем векторских (објектних) вредности за процес поравнања. Ипак, ови недостаци немају превелики утицај на квалитет резултата.

8.4 Дискусија постигнутих резултата

На основу приказаних резултата се може закључити да Платформа представљена у дисертацији представља одличну основу за откривање знања неопходног за планирање будућих истраживања у домену биоинформатике, посебно у области рационалног дизајна лекова. Представљени резултати указују да свака од имплементираних метода има одговарајућу улогу у процесу откривања релевантних информација, али и да доприноси популаризацији база података које нису познате биоинформатичкој јавности. Иако Платформа не покрива све доступне и јавне базе података, ипак можемо рећи да подржава најзначајније базе (репозиторијуме) у овом домену, које су представљене у одељку 4.5.1, а које су такође подржане у другим софтверским решењима представљеним у одељку 6.4. Иако је у домену биоинформатике показала велику ефикасност, остаје отворено питање да ли би и на који начин Платформа била од користи у другим областима које подржавају онтолошко представљање података.

Алгоритам за детекцију сличних података показао је добре резултате за овај значајан биоинформатички проблем. Приказани резултати указују да примена семантичких технологија у комбинацији са одређеним математичким приступима, може унапредити методе онтолошких поравнања. Такође, корак селекције предиката је од велике важности за побољшање перформанси алгоритма. Његовом применом било је могуће извршити неку врсту кластеризације података, која је допринела ефикаснијем раду алгоритма. Дати алгоритам се може применити и у било којој другој области која покрива онтолошко моделовање података. Такође, може се закључити да успешност алгоритма зависи од тога колико „богато“ су одговарајуће инстанце представљене својим подацима. Инстанце које немају довољно предиката и објектних вредности које их описују, могу се окарактерисати као „сиромашне“ и као такве немају карактеристику сличних, иако је реално супротно. Ово нарочито важи за инстанце које су у фази истраживања или за оне чији су подаци из неког разлога заштићени за јавност.

Од доступности података на *remote endpoint*-има зависи ефикасност целе Платформе. У случају да су *endpoint*-и недоступни, методе које Платформа нуди не могу бити искоришћене. Такође, показало се да је за потпуну примену неких метода Платформе, неопходно доменско знање и познавање семантичких технологија, чиме се огледа несавршеност Платформе. Ипак, примена одговарајућих приступа за визуелизацију података, настоји да направи баланс и пружи подршку у откривању релевантних података. Ова комбинација била је довољна да повољно утиче на ефикасност Платформе.

9 Закључак

Непосредна улога биоинформатике за људску популацију и квалитетан, здрав живот, постала је импресивно значајна последњих година. Достигнућа биоинформатике имају за циљ да потпомогну побољшаној дијагностици болести, да откривају генетске предиспозиције болести, да утичу на рационални дизајн лекова итд. Међутим, кључ успеха у пољу биоинформатике у нарочитој мери зависи од доступности и расположивости података. Данас се многе научно-истраживачке институције и лабораторије (попут Лабораторије представљене у одељку 2.3) суочавају са проблемом проналажења адекватних података јер су они често доступни преко хетерогених база података, које користе различите формате, конвенције и речнике. Овај проблем је постао још изазовнији и зрелији са брзом акумулацијом података. Подаци се непрекидно концентришу, акумулирају и ажурирају и зато су неадекватна имплементација података, дефицит интегрисаних решења и отежана претрага података, витални проблеми биоинформатичке заједнице. С обзиром на природу проблема, свој велики потенцијал у циљу решавања истог пружају технологије семантичког веба. У дисертацији је постављена хипотеза да технологије семантичког веба:

- нуде богате и добро дефинисане моделе за представљање и интеграцију података;
- омогућавају агрегацију хетерогених података коришћењем експлицитне семантике;
- олакшавају претрагу података;
- омогућавају поновну употребу података у циљу извођења закључака неопходних за даља биоинформатичка истраживања.

Многа софтверска решења, представљена у оквиру прегледа литературе (одељак 6.4), су својим присуством на биоинформатичкој сцени допринела истраживачкој заједници, водећи се хипотезама представљеним у дисертацији. Биоинформатичка платформа (Платформа), која представља главни предмет истраживања ове дисертације, један је од примера софтверских решења која је имплементирана у складу са циљевима представљеним у одељку 1.1, а која уз помоћ сопствених метода и истраживањима која су претходила њеном развоју, доказује дате хипотезе. Платформа пружа подршку за извођење Federated SPARQL упита над више иницијалних извора података (CPCTAS, Bio2RDF, Chem2Bio2RDF, EMBL-EBI) уз могућност додавања кориснички селектованог скупа података. Резултати упита имају за циљ откривање релевантних и комплементарних информација, који би утицали на даље одлуке у спровођењу експерименталних истраживања у домену рационалног дизајна лекова и преклиничких тестирања активних супстанци. Једна од метода Платформе јесте и детекција сличних података, која се примењује над подацима добијеним као резултат реализације предефинисаних упита. Алгоритам, који је развијен за потребе ове методе, представља оригиналан и нов приступ за откривање сличних података, који се може експлоатисати не само у домену биоинформатике, него и било којој другој области која подржава онтолошко моделовање података.

У оквиру дисертације су спроведена и истраживања која су претходила развоју Платформе: креирање семантичког модела за представљање „Експеримент“ концепта - PIBAS онтологија; развој онтолошке базе података - CPCTAS база података; представљање прототипа софтвера *ExperimentSearch* за претрагу података CPCTAS базе; интеграција PIBAS модела са актуелним онтолошким моделима популарних репозиторијумима (Bio2RDF). Кренувши од анализе постојећих онтологија које се баве дефинисањем „Експеримент“ термина, развијена је специфична доменска PIBAS онтологија за потребе Лабораторије. Она се састоји од концепата који презентују прецизне стручне термине, а који могу бити изнова коришћени и дељени. Основна карактеристика PIBAS онтологије јесте њена проширивост. Ова предност јој омогућава лаку надоградњу као и интеграцију са другим онтолошким моделима. На овај начин је доказан део хипотезе који се односи на процес представљања и интеграције података. База података CPCTAS, поред PIBAS онтологије, садржи и податке појединачних експеримената који се изводе у Лабораторији. Ови подаци се могу одразити на одлуке приликом реализације будућих експеримената. Како би се олакшала претрага дате базе развијен је и прототип софтвера *ExperimentSearch*, који се сматра претечом Платформе представљене у дисертацији. Преостали, главни део истраживања је усмерен на

Платформу.

Платформа, као главни предмет истраживања дисертације, обезбеђује методе извршавања предефинисаних упита, динамичког филтрирања резултата упита, додавања кориснички селектоване базе података и детекцију сличних података. Метода извршавања предефинисаних упита има за циљ да применом Federated SPARQL упита спроводи специфичне претраге одређених репозиторијума, да комбинује обрасце над различитим онтолошким базама, како би се открили релевантни и комплементарни подаци. Овим је доказан део хипотезе који се односи на агрегацију података. Исто правило важи и за било коју другу базу података, која се може интегрисати кроз методу додавања кориснички селектоване базе. На овај начин се многе мање лабораторије мотивишу да укажу на свој значај у домену биоинформатике. Ово свакако отвара пут ка сарадњи лабораторија широм света. С обзиром на анализу резултата спроведену у претходном поглављу, закључује се да су подаци до којих се долази применом основних метода, од значаја за планирање будућих истраживања. На тај начин је потврђен последњи део хипотезе.

Метода детекције сличних података претендује да применом семантичких технологија уз одређене технике процесирања текстуалних података и математичке приступе, допринесе решавању једног од значајних проблема у домену биоинформатике. Откривање сличних података је круцијалан фактор за планирање будућих експеримената и базира се на чињеници да сличне структуре деле сличне физичко-хемијске особине и биолошке активности. На тај начин се могу уштедети ресурси и избећи експериментисање са структурама који могу довести до потенцијално сличних резултата. Алгоритам има за циљ да детектује сличност између инстанци онтолошких база података, које су добијене као резултат извршавања предефинисаних упита. Предложени алгоритам се генерално заснива на примени екстензијских техника онтолошког поравнања, претпроцесирању текстуалних података, представљању текстуалних вредности у форми вектора (применом модела векторског простора) и одређивању угла косинуса између њих (применом мере косинусне сличности). На крају, децизивни фактор представља и улога експерата, који дају своју коначну процену о успешности рада алгоритма. Предложени алгоритам се може експлоатисати и у било ком другом домену који покрива онтолошко моделовање података.

На основу свега изложеног закључује се да Платформа представљена у дисертацији означава изврстан приступ за откривање релевантних информација и знања, неопходних за решавање широког спектра биоинформатичких проблема. Платформа се потенцијално може применити и у било ком другом домену који пружа подршку онтолошком моделовању. Иако су циљеви дисертације испуњени, простор за будућа истраживања и побољшања Платформе засигурно постоји, и то у неколико праваца:

- Проширење Платформе новим методама које би олакшале рад биоинформатичке истраживачке заједнице;
- Коришћење VoID[117] [104] речника за успостављање лакше интеграције између база података приликом примене методе додавања кориснички селектоване базе;
- Побољшање перформанси алгоритма за детекцију сличних података:
  - применом техника паралелизације на процес обраде инстанци и поређења објектних (векторских) вредности;
  - имплементација опције мануелног подешавања прага сличности, како би се уочиле разлике у резултатима рада алгоритма;
  - проширење листе непожељних предиката (*non_selected_predicates*) методом кластеровања (неком другом DM методом) или применом IE процеса;
  - коришћење неке од метода машинског учења за корак селекције предиката, чиме би се проблем детекције сличних података свео на проблем предвиђања сличности.

[117] VoID је речник RDFS-а за представљање метаподатака о RDF базама података, који презентује спону између власника базе и корисника.

Додатак

У овом сегменту су представљени примери стандардизованих и корисничких именских простора који се користе у оквиру дисертације.

Стандардизовани (подразумевани) или уграђени:

- *rdf :<http://www.w3.org/1999/02/22-rdf-syntax-ns#>*
- *rdfs:<http://www.w3.org/2000/01/rdf-schema#>*
- *xsd:<http://www.w3.org/2001/XMLSchema#>*
- *owl:<http://www.w3.org/2002/07/owl#>*
- *foaf:<http://xmlns.com/foaf/0.1/>*
- *dc:<http://purl.org/dc/elements/1.1/>, (dc ≡ dcterms)*
- *skos:<http://www.w3.org/2004/02/skos/core#>*

Кориснички*:*

- *pibas:<http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>*
- *testOntology:< http://147.91.205.66:2020/Tests/TestOntology#>*
- *expo:<http://www.hozo.jp/owl/EXPOApr19.xml/>*
- *bio2rdf_vocabulary:<http://bio2rdf.org/bio2rdf_vocabulary:>*
- *drugbank:<http://bio2rdf.org/drugbank:>*
- *drugbank_drug:<http://chem2bio2rdf.org/drugbank/resource/drugbank_drug/>*
- *drugbank_vocabulary:<http://bio2rdf.org/drugbank_vocabulary:>*
- *kegg:<http://bio2rdf.org/kegg:>*
- *kegg_vocabulary:<http://bio2rdf.org/kegg_vocabulary:>*
- *pubmed_vocabulary:<http://bio2rdf.org/pubmed_vocabulary:>*
- *lodd:<http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/>*
- *chembl:<http://rdf.ebi.ac.uk/terms/chembl#>* (*chembl ≡ cco*)
- *chembl_molecule:<http://rdf.ebi.ac.uk/resource/chembl/molecule/>*
- *chembl_cell_line:<http://rdf.ebi.ac.uk/resource/chembl/cell_line/>*
- *chembl_target:<http://rdf.ebi.ac.uk/resource/chembl/target/>*
- *chembl_assay:<http://rdf.ebi.ac.uk/resource/chembl/assay/>*
- *sio:<http://semanticscience.org/resource/>*
- *uniprot:<http://purl.uniprot.org/uniprot/>*
- *pubchem:<http://chem2bio2rdf.org/pubchem/resource/>*
- *pubchem_compound:<http://chem2bio2rdf.org/pubchem/resource/pubchem_compound/>*
- *pubchem_bioassay:<http://chem2bio2rdf.org/pubchem/resource/pubchem_bioassay_interaction/>*
- *bindingdb:<http://chem2bio2rdf.org/bindingdb/resource/>*
- *bibo:<http://purl.org/ontology/bibo/>*

Библиографија

1. Perović V. Development of multifunctional bioinformatics platform based on electron-ion interaction potential of biological molecules. PhD Thesis. University of Belgrade, Faculty of Mathematics; 2013.

2. Nagarajan P. An Over View of Bioinformatics. Trends in Biomaterials & Artificial Organs. 2004; 17(2): p. 4-8.

3. Bioinformatika. [Online] wikipedia.org.; 2018 [cited 2019 Januar 23]. Available from: https://bs.wikipedia.org/wiki/Bioinformatika.

4. Stephens S, LaVigna D, DiLascio M, Luciano J. Aggregation of bioinformatics data using Semantic Web technology. Web Semantics: Science, services and agents on the world wide web. 2006 September; 4(3): p. 216-221.

5. Masseroli M, Mons B, Bongcam-Rudloff E, Ceri S, Kel A, Rechenmann F, et al. Integrated Bio-Search: challenges and trends for the integration, search and comprehensive processing of biological information. BMC bioinformatics. 2014 January; 15(1): p. S2.

6. Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific American. 2001 May 17; 284(5): p. 28-37.

7. Djokic-Petrovic M, Cvjetkovic V, Yang J, Zivanvic M, Wild DJ. PIBAS FedSPARQL: a web-based platform for integration and exploration of bioinformatics datasets. Journal of Biomedical Semantics. 2017 December; 8(1): p. 42.

8. IC50. [Online] wikipedia.org.; 2018 [cited 2019 February 19]. Available from: https://sr.wikipedia.org/sr-ec/IC50.

9. Berners-Lee T, Fielding R, Masinter L. Uniform Resource Identifier (URI): Generic Syntax. Standards Track. Network Working Group; 2005.

10. Baker CJ, Cheung KH, editors. Semantic web: Revolutionizing knowledge discovery in the life sciences: Springer Science & Business Media; 2007.

11. Resource Description Framework (RDF) Model and Syntax Specification. [Online] www.w3.org.; 1999 [cited 2019 February 10]. Available from: https://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

12. Maedche A, Staab S. Ontology learning for the semantic web. IEEE Intelligent systems. 2001 March; 16(2): p. 72-79.

13. SPARQL 1.1 Overview. [Online] www.w3.org.; 2013 [cited 2019 February 10]. Available from: https://www.w3.org/TR/sparql11-overview/.

14. Bioinformaticsweb.co.nr: Open Access Bioinformatics Education Resource. [Online] bioinformaticsweb.net.; 2005 [cited 2019 February 10]. Available from: http://bioinformaticsweb.net.

15. Luscombe NM, Greenbaum D, Mark G. What is bioinformatics? A proposed definition and overview of the field. Methods of information in medicine. 2001; 40(04): p. 346-358.

16. Achuthsankar NS. Computational biology & bioinformatics: a gentle overview. Communications of the Computer Society of India. 2007; 2.

17. Stoesser G, Tuli MA, Lopez R, Sterk P. The EMBL nucleotide sequence database. Nucleic Acids Research. 1999 January 1; 27(1): p. 18-24.

18. Burks C, Cassidy M, Cinkosky MJ, Cumella KE, Gilna P, Hayden JE, et al. GenBank. Nucleic Acids Research. 1991 April 25; 19(suppl): p. 2221-2225.

19. George DG, Barker WC, Hunt LT. The protein identification resource (PIR). Nucleic acids research. 1986 January 10; 14(1): p. 11-15.

20. Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank. Nucleic acids research. 1991 April 25; 19(Suppl): p. 2247.

21. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000 January 1; 28(1): p. 27-30.

22. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Research. 2019 January 8; 47(D1): p. D1102-D1109.

23. Wheeler D, Chappey C, Lash AE, Leipe DD, Madden TL, Schuler GD, et al. Database resources of the national center for biotechnology information. Nucleic acids research. 2000 January 1; 28(1): p. 10-14.

24. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. The universal protein resource (UniProt). Nucleic acids research. 2005; 33(suppl_1): p. D154-D159.

25. Schuler GD, Epstein JA, Ohkawa H, Kans JA. [10] Entrez: Molecular biology database and retrieval system. Methods in enzymology. 1996 January 1; 266: p. 141-162.

26. Fujibuchi W, Goto S, Migimatsu H, Uchiyama I, Ogiwara A, Akiyama Y, et al. DBGET/LinkDB: an integrated database retrieval system. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 1998; 98: p. 683-694.

27. Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The Influenza Virus Resource at the National Center for Biotechnology Information. Journal of Virology. 2008 January 15; 82(2): p. 596-601.

28. Hughes JP, Rees S, Kalindjian BS, Philpott KL. Principles of early drug discovery. British journal of pharmacology. 2011 March 1; 162(6): p. 1239-1249.

29. Drug design. [Online] wikipedia.org.; 2019 [cited 2019 February 10]. Available from: http://en.wikipedia.org/wiki/Drug_design.

30. Gashaw I, Ellinghaus P, Sommer A, Asadullah K. What makes a good drug target. Drug discovery today. 2011 December 1; 16(23-24): p. 1037-1043.

31. Jović D. Synthesis, characterisation and biological activity investigation of fullerenol/doxorubicin nanocomposite. PhD Thesis. University of Novi Sad, Faculty of Science; 2018.

32. Jakimov D. Uticaj modifikovanih steroidnih jedinjenja na ćelijski ciklus, indukciju apoptoze i nastanak genetskih oštećenja u humanim tumorskim ćelijama. Doktorska disertacija. Univerzitet u Novom Sadu, Prirodno-matematički fakultet, Departman za hemiju, biohemiju i zaštitu zivotne

sredine; 2016.

33. Riss TL, Moravec RA, Niles AL, Duellman S, Benink HA, Worzella TJ, et al. Cell viability assays: Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2016.

34. Riss T, Moravec R, Niles A. Selecting cell-based assays for drug discovery screening. Cell Notes. 2015; 13: p. 16-21.

35. CPCTAS-LCMB, Faculty of Science, University of Kragujevac, Serbia. [Online] cpctas-lcmb.pmf.kg.ac.rs.; 2015 [cited 2019 February 10]. Available from: http://cpctas-lcmb.pmf.kg.ac.rs/lcmb/.

36. Cvjetković V, Marija Đ, Branko A, Milena Ć. The ontology supported intelligent system for experiment search in the scientific research center. Kragujevac Journal of Science. 2014;(36): p. 95-110.

37. G. de Brevern A, Meyniel JP, Fairhead C, Neuvéglise C, Malpertuy A. Trends in IT innovation to build a next generation bioinformatics solution to manage and analyse biological big data produced by NGS technologies. BioMed research international. 2015; 2015.

38. Stroil KB, Dorić S, Lukić BL, Pojskić N. Aplikativna bioinformatika-Praktikum. Sarajevo: Univerzitet u Sarajevu, Institut za genetičko inženjerstvo i biotehnologiju; 2018.

39. Bin C, Dong X, Jiao D, Wang H, Zhu Q, Ding Y, et al. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. BMC bioinformatics. 2010 December; 11(1): p. 255.

40. Wild DJ. Mining large heterogeneous data sets in drug discovery. Expert Opinion on Drug Discovery. 2009 October 1; 4(10): p. 995-1004.

41. Slater T, Bouton C, Huang ES. Beyond data integration. Drug discovery today. 2008 July 1; 13(13-14): p. 584-589.

42. Berners-Lee T. The World Wide Web: A very short personal history. [Online] www.w3.org.; 1998 [cited 2019 February 22]. Available from: https://www.w3.org/People/Berners-Lee/ShortHistory.html.

43. Petrušić D. Semantičko modelovanje i ontološka integracija informacionih sistema Otvorene vlade. Doktorska disertacija. Novi Sad: Univerzitet u Novom Sadu, Fakultet tehničkih nauka; 2016.

44. Cannata N, Schröder M, Marangoni R, Romano P. A Semantic Web for bioinformatics: goals, tools, systems, applications. BMC Bioinformatics. 2008 April 25; 9(Suppl 4 :S1).

45. Grigoris A, Van Harmelen F. A semantic web primer: MIT press; 2004.

46. Bray T, Paoli J, Sperberg-McQueen MC, Maler E, Yergeau F. Extensible markup language (XML). World Wide Web Journal. 1997 December; 2(4): p. 27-66.

47. Consortium TU. The Unicode Standard, Version 2.0: Addison-Wesley Longman Publishing Co., Inc.; 1997.

48. Coates T, Connolly D, Dack D, Daigle LL, Denenberg R, Dürst MJ, et al. Uris, urls, and urns: Clarifications and recommendations 1.0. World Wide Web Consortium, Note NOTE-uri-clarification-20010921. 2001 September.

49. Namespaces in XML 1.0 (Third Edition). [Online] www.w3.org.; 1999 [cited 2019 February 22]. Available from: https://www.w3.org/TR/xml-names/.

50. Ranogajec A. Diplomski rad. Sveučilište u Zagrebu, Fakultet strojarstva i brodogradnje; 2011.

51. Roy J, Ramanujan A. XML schema language: taking XML to the next level. IT professional. 2001 March; 3(2): p. 37-40.

52. RDF Schema 1.1. [Online] www.w3.org.; 2014 [cited 2019 February 22]. Available from: https://www.w3.org/TR/rdf-schema/.

53. Pavić K. Semantički web. Seminarski rad. Zagreb: Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva.

54. Brewster C, O'Hara K. Knowledge representation with ontologies: the present and future. IEEE Intelligent Systems. 2004 January; 19(1): p. 72-81.

55. Stevens R, Goble CA, Bechhofer S. Ontology-based knowledge representation for bioinformatics. Briefings in bioinformatics. 2000 November; 1(4): p. 398-414.

56. Guarino N. Understanding, building and using ontologies. International Journal of Human-Computer Studies. 1997 February 1; 46(2-3): p. 293-310.

57. Gómez-Pérez A, Corcho O. Ontology languages for the semantic web. IEEE Intelligent systems. 2002 January; 17(1): p. 54-60.

58. Uschold M, King M. Towards a methodology for building ontologies. In Workshop on Basic Ontological Issues in Knowledge Sharing; 1995; Edinburgh: Citeseer. p. 13.

59. Corcho O, Fernández-López M, Gómez-Pérez A. Methodologies, tools and languages for building ontologies. Where is their meeting point? Data & knowledge engineering. 2003 July 1; 46(1): p. 41-64.

60. Ismail MA, Yaacob M, Kareem AK. Ontology Construction: An Overview, University of Malaya. University of Malaya; 2006.

61. Musen MA. The Protégé project: A look back and a look forward. AI Matters. 2015 Jun; 1(4).

62. Weiten M. Ontostudio® as a ontology engineering environment. In Semantic knowledge management.: Springer, Berlin, Heidelberg; 2009. p. 51-60.

63. McGuinness DL, van Harmelen F. OWL web ontology language overview. W3C recommendation. 2004 February 10; 10(10).

64. Bogdanović M. Semantički veb – primena u automatskom upravljanju uređajima. Master rad. Beograd: Univerzitet u Beogradu, Matematički fakultet; 2011.

65. Paunović L, Stokić A. Uticaj ontologija na funkcionalnost Web-a. In XI međunarodni naučno-stručni simpozijum, INFOTEH-JAHORINA; 2012. p. 920-924.

66. SPARQL 1.1 Federated Query. [Online] www.w3.org.; 2013 [cited 2019 February 20]. Available from: https://www.w3.org/TR/sparql11-federated-query/.

67. Soldatova LN, King RD. An ontology of scientific experiments. Journal of the Royal Society Interface. 2006 Juny 6; 3(11): p. 795-803.

68. Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, et al. The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics. 2006 January 21; 22(7): p. 866-873.

69. Mayer G, Montecchi-Palazzi L, Ovelleiro D, Jones AR, Binz PA, Deutsch EW, et al. The HUPO proteomics standards initiative-mass spectrometry controlled vocabulary. Database. 2013 January 1; 2013.

70. Ontology WG. [Online] msi-workgroups.sourceforge.net.; 2015 [cited 2019 March 11]. Available from: http://msi-ontology.sourceforge.net/.

71. Chemical Methods Ontology. [Online] www.obofoundry.org.; 2019 [cited 2019 March 28]. Available from: http://www.obofoundry.org/ontology/chmo.html.

72. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The ontology for biomedical investigations. PLoS ONE. 2016; 11(4): p. e0154556.

73. Arsic B, Djokic M, Cvjetkovic V, Spalevic P, Zivanovic M, Mladenovic M. Integration of bioactive substances data for preclinical testing with Cheminformatics and Bioinformatics resources. In Proceedings of the 23nd International Electrotechnical and Computer Science Conference. ERK; 2014; Portorož, Slovenia. p. 146-149.

74. Cvjetković V, Đokić M, Arsić B. Ontology Visualization. In Proceedings of the 1st Virtual International Conference on Advanced Research in Scientific Areas (ARSA-2012); 2012. p. 1999-2004.

75. Arsic B, Djokic M, Cvjetkovic V, Spalevic P, Ilic S. Semantic search framework for distributed semantically based cheminformatics and bioinformatics datasets. In Proceedings of the 5th International Conference on Information Society and Technology (ICIST 2015); 2015; Serbia: Society for Information Systems and Computer Networks. p. 518-522.

76. Cvjetković V, Đokić M, Arsić B. Semantically based customized search on local web site. In Proceedings of the 2nd Virtual International Conference on Advanced Research in Scientific Areas (ARSA-2013); 2013; Slovakia. p. 453-458.

77. Cvjetković V, Marković S, Arsić B, Žižić J, Đokić M. Semantički bazirana kastomizovana pretraga na lokalnom veb sajtu: http://cpctas-lcmb.pmf.kg.ac.rs. Univerzitet u Kragujevcu, Prirodno-matematički fakultet; 2013.

78. Abdulganiyu AY, Zahraddeen S, Kabir YM, Abubakar US. Comparison Of Popular Bioinformatics Databases. International Journal of Applied and Advanced Scientific Research. 2016; 1(1): p. 19-28.

79. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics. 2008 October 1; 41(5): p. 706-716.

80. Bizer C, Heath T, Berners-Lee T. Linked data: The story so far. In Semantic services, interoperability and web applications: emerging concepts.: IGI Global; 2011. p. 205-227.

81. Jentzsch A, Zhao J, Hassanzadeh O, Cheung KH, Samwald M, Andersson B. Linking open drug data. In Proceedings of the 5th International Conference on Semantic Systems, I-SEMANTICS; 2009; Graz, Austria.

82. Smith AK, Cheung KH, Yip KY, Schultz M, Gerstein MB. LinkHub: a semantic web system that facilitates cross-database queries and information retrieval in proteomics. BMC bioinformatics. 2007; 8(Suppl 3): p. S5.

83. Antezana E, Blondé W, Egaña M, Rutherford A, Stevens R, De Baets B, et al. BioGateway: a semantic systems biology tool for the life sciences. BMC bioinformatics. 2009; 10(10): p. S11.

84. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic acids research. 2015 April 6; 43(W1): p. W580-W584.

85. ChEMBL documentation. [Online] www.ebi.ac.uk.; 2019 [cited 2019 March 24]. Available from: https://www.ebi.ac.uk/rdf/documentation/chembl/.

86. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. Clinical Pharmacology & Therapeutics. 2016 March 1; 99(3): p. 285-297.

87. Drug Targets. [Online] www.horizondiscovery.com.; 2019 [cited 2019 March 28]. Available from: https://www.horizondiscovery.com/cell-lines/all-products/explore-by-your-research-area/drug-targets.

88. Patel MN, Halling-Brown MD, Tym JE, Workman P, Al-Lazikani B. Objective assessment of cancer genes for drug discovery. Nature reviews Drug discovery. 2013; 12(1): p. 35.

89. Zlatović M, Petrović D. Osnovi molekulskog modelovanja. Praktikum. Univerzitet u Beogradu, Hemijski fakultet; 2016.

90. 5 tips for choosing the right cell line for your experiment. [Online] blog.horizondiscovery.com.; 2016 [cited 2019 March 22]. Available from: https://blog.horizondiscovery.com/5-tips-for-choosing-the-right-cell-line-for-your-experiment.

91. Cvjetkovic V, Djokic M. Semantic web based organization of scientific bibliography references. In Proceedings of the 3rd International Virtual Conference on Advanced Scientific Results (SCIECONF-2015); 2015; Slovakia: EDIS - Publishing Institution of the University of Zilina. p. 25-29.

92. Djokic-Petrovic M, Pritchard D, Ivanovic M, Cvjetkovic V. IMI Python: Upgraded CS Circles web-based Python course. Computer Applications in Engineering Education. 2016; 24(3): p. 464-480.

93. Gunaratna K, Lalithsena S, Sheth A. Alignment and dataset identification of linked data in semantic web. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014 March; 4(2): p. 139-151.

94. Saleem M. Efficient Source Selection for SPARQL Endpoint Query Federation. PhD Thesis. Leipzig: University of Leipzig, Faculty of Mathematics and Computer Science; 2016.

95. Caliusco ML, Stegmayer G. Semantic web technologies and artificial neural networks for intelligent web knowledge source discovery. In Emergent web intelligence: Advanced semantic technologies.: Springer, London; 2010. p. 17-36.

96. Arsić B, Đokić-Petrović , Marija , Spalević P, Milentijević I, Rančić D, et al. SpecINT: A framework for data integration over cheminformatics and bioinformatics RDF repositories. Semantic Web - Interoperability, Usability, Applicability. 2018;(Preprint).

97. Quilitz B, Leser U. Querying distributed RDF data sources with SPARQL. In European semantic web conference; 2008: Springer, Berlin, Heidelberg. p. 524-538.

98. Buttler D, Coleman M, Critchlow T, Fileto R, Han W, Pu C, et al. Querying multiple bioinformatics information sources: can semantic web research help? ACM SIGMOD Record. 2002; 31(4): p. 59-64.

99. Euzenat J, Shvaiko P. Ontology matching: Heidelberg: Springer; 2007.

100. Juričić K, Meštrović A. Pregled tehnika i postupaka poravnavanja ontologija. Zbornik Veleučilišta u Rijeci; 2013.

101. Hartig O, Bizer C, Freytag JC. Executing SPARQL queries over the web of linked data. In International Semantic Web Conference; 2009: Springer, Berlin, Heidelberg. p. 293-309.

102. Ladwig G, Tran T. Linked data query processing strategies. In International Semantic Web Conference; 2010: Springer, Berlin, Heidelberg. p. 453-469.

103. Görlitz O, Staab S. Splendid: Sparql endpoint federation exploiting void descriptions. In Proceedings of the Second International Conference on Consuming Linked Data-Volume 782. CEUR-WS. org; 2011: CEUR-WS. org. p. 13-24.

104. Describing Linked Datasets with the VoID Vocabulary. [Online] www.w3.org.; 2017 [cited 2019 March 15]. Available from: https://www.w3.org/TR/void/.

105. Schwarte A, Haase P, Hose K, Schenkel R, Schmidt M. Fedx: Optimization techniques for federated query processing on linked data. In International semantic web conference; 2011: Springer, Berlin, Heidelberg. p. 601-616.

106. de Oliveira HR, Tavares AT, Lóscio BF. Feedback-based data set recommendation for building linked data applications. In Proceedings of the 8th International Conference on Semantic Systems; 2012: ACM. p. 49-55.

107. Nikolov A, d'Aquin M, Motta E. What should I link to? Identifying relevant sources and classes for data linking? In Joint International Semantic Technology Conference; 2011: Springer, Berlin, Heidelberg. p. 284-299.

108. Cvjetković V, Djokić M, Arsić B. Wikipedia Browsing With DBpedia. In Proceedings in EIIC-The 2nd Electronic International Interdisciplinary Conference; 2013.

109. Aparício AS, Farias OL, dos Santos N. Applying ontologies in the integration of heterogeneous relational databases. In Proceedings of the 2005 Australasian Ontology Workshop; 2005: Australian Computer Society, Inc. p. 11-16.

110. Seneviratne O, Sealfon R. QueryMed: An intuitive federated SPARQL query builder for

biomedical RDF data. ; 2010.

111. Hu W, Qiu H, Huang J, Dumontier M. BioSearch: a semantic search engine for Bio2RDF. Database. 2017 January 1; 2017.

112. Longley DB, Harkin PD, Johnston PG. 5-fluorouracil: mechanisms of action and clinical strategies. Nature reviews cancer. 2003 May; 3(5): p. 330.

113. Lin LM, Liu GC, Wang Y, Lu W. Star-shaped SPARQL Query Optimization on Column-family Overlapping Storage. In Current Trends in Computer Science and Mechanical Automation.: Sciendo Migration; 2017. p. 67-73.

114. Schweiger D, Trajanoski Z, Pabinger S. SPARQLGraph: a web-based platform for graphically querying biological semantic web databases. BMC bioinformatics. 2014; 15(1): p. 279.

115. Löffler F, Opasjumruskit K, Karam N, Fichtmüller D, Schindler U, Klan F, et al. Honey bee versus apis mellifera: A semantic search for biological data. In European Semantic Web Conference; 2017: Springer, Cham. p. 98-103.

116. Zaki N, Tennakoon C. BioCarian: search engine for exploratory searches in heterogeneous biological databases. BMC bioinformatics. 2017 December; 18(1): p. 435.

117. García-Godoy MJ, Navas-Delgado I, Aldana-Montes J. Bioqueries: a social community sharing experiences while querying biological linked data. In Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences; 2011: ACM. p. 24-31.

118. Groth P, Loizou A, Gray AJG, Goble C, Harland L, Pettifer S. API-centric linked data integration: the open PHACTS discovery platform case study. Web Semantics: Science, Services and Agents on the World Wide Web. 2014 December 31; 29: p. 12-18.

119. Chen Q, Zobel J, Verspoor K. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. Database. 2017; 2017.

120. Johnson MA, Maggiora GM, editors. Concepts and Applications of Molecular Similarity: Wiley; 1990.

121. Cao Y, Jiang T, Girke T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. Bioinformatics. 2008 Jul 1; 24(13): p. i366-i374.

122. Ferreira JD, Couto FM. Semantic similarity for automatic classification of chemical compounds. PLoS computational biology. 2010 September 23; 6(9): p. e1000937.

123. Le SQ, Ho TB, Phan TH. A novel graph-based similarity measure for 2D chemical structures. Genome Informatics. 2004; 15(2): p. 82-91.

124. Doniger S, Hofmann T, Yeh J. Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. Journal of computational biology. 2002 December 1; 9(6): p. 849-864.

125. Dacić G. Primena klasterovanja u rešavanju problema grupisanja genoma. In INFOTEH-JAHORINA; 2008. p. 451-454.

126. Dunham MH. Data mining: Introductory and advanced topics: Pearson Education India; 2006.

127. Sheridan RP, Kearsley SK. Why do we need so many chemical similarity search methods? Drug discovery today. 2002 September 1; 7(17): p. 903-911.

128. Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics. 2006 July 31; 22(19): p. 2405-2412.

129. Holub M, Proksa O, Bieliková M. Detecting identical entities in the semantic web data. In International Conference on Current Trends in Theory and Practice of Informatics; 2015: Springer, Berlin, Heidelberg. p. 519-530.

130. Paliouras G, Spyropoulos CD, Tsatsaronis G, editors. Knowledge-driven multimedia information extraction and ontology evolution: bridging the semantic gap: Springer Science & Business Media; 2011.

131. Wang LL, Bhagavatula C, Neumann M, Lo K, Wilhelm C, Ammar W. Ontology alignment in the biomedical domain using entity definitions and context. arXiv preprint arXiv:1806.07976. 2018 Jun 20.

132. Zhang Y, Wang X, Lai S, He S, Liu K, Zhao J, et al. Ontology matching with word embeddings. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data.: Springer, Cham; 2014. p. 34-45.

133. Ferrara A, Nikolov A, Scharffe F. Data linking for the semantic web. International Journal on Semantic Web and Information Systems (IJSWIS). 2011 July 1; 7(3): p. 46-76.

134. Halpin H, Hayes PJ, McCusker JP, McGuinness DL, Thompson HS. When owl: sameas isn't the same: An analysis of identity in linked data. In International Semantic Web Conference; 2010: Springer, Berlin, Heidelberg. p. 305-320.

135. Miles A, Bechhofer S. SKOS simple knowledge organization system reference. W3C recommendation. 2009; 18: p. W3C.

136. Spasic I, Ananiadou S, McNaught J, Kumar A. Text mining and ontologies in biomedicine: making sense of raw text. Briefings in bioinformatics. 2005 September 1; 6(3): p. 239-251.

137. The Dublin Core Element Set Version 1.1. [Online] www.dublincore.org.; 1999 [cited 2019 February 11]. Available from: http://www.dublincore.org/specifications/dublin-core/dces/1999-07-02/.

138. Ehrig M. Ontology Alignment – Bridging the Semantic Gap: Springer Science & Business Media; 2006.

139. Shvaiko P, Euzenat J. Ontology matching: state of the art and future challenges. IEEE Transactions on knowledge and data engineering. 2013 January; 25(1): p. 158-176.

140. Romero MM, Naya JMV, Loureiro JP, Ezquerra N. Ontology alignment techniques. In Encyclopedia of Artificial Intelligence.: IGI Global; 2009. p. 1290-1295.

141. Gali N, Mariescu-Istodor R, Fränti P. Similarity measures for title matching. In 2016 23rd International Conference on Pattern Recognition (ICPR); 2016: IEEE. p. 1548-1553.

142. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. In Soviet physics doklady.; 1966. p. 707-710.

143. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association. 1989 June 1; 84(406): p. 414-420.

144. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology. 1970 March 28; 48(3): p. 443-453.

145. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of molecular biology. 1981 March 25; 147(1): p. 195-197.

146. Cohen W, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records. In Kdd workshop on data cleaning and object consolidation.; 2003. p. 73-78.

147. Gusfield D. Algorithms on strings, trees, and sequences: computer science and computational biology: Cambridge university press; 1997.

148. Kukich K. Techniques for automatically correcting words in text. Acm Computing Surveys (CSUR). 1992 December 1; 24(4): p. 377-439.

149. Brew C, McKelvie D. Word-pair extraction for lexicography. In Proceedings of the 2nd International Conference on New Methods in Language Processing; 1996. p. 45-55.

150. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat. 1901; 37: p. 547-579.

151. Salton G. Automatic text processing: The transformation, analysis, and retrieval of. Reading: Addison-Wesley. 1989; 169.

152. The Soundex Indexing System. [Online] www.archives.gov.; 2007 [cited 2019 February 11]. Available from: https://www.archives.gov/research/census/soundex.html.

153. Philips L. Hanging on the Metaphone. Computer Language. 1990 December; 7(12): p. 39-44.

154. Aggarwal CC, Zhai C, editors. Mining text data: Springer Science & Business Media; 2012.

155. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research. 2004 January 1; 32(suppl_1): p. D267-D270.

156. Miller GA. WordNet: a lexical database for English. Communications of the ACM. 1995 November 1; 38(11): p. 39-41.

157. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. Journal of machine learning research. 2011 November; 12(Aug): p. 2493-2537.

158. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics. 2007 Jun 1; 40(3): p. 288-299.

159. Garla VN, Brandt C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC bioinformatics. 2012 December; 13(1): p. 261.

160. Rybinski M, Aldana-Montes JF. Calculating semantic relatedness for biomedical use in a knowledge-poor environment. BMC bioinformatics. 2014 December; 15(14): p. S2.

161. What's the difference between semantic relatedness and semantic similarity, and how to calculate them? [Online] www.quora.com.; 2015 [cited 2019 February 11]. Available from: https://www.quora.com/Whats-the-difference-between-semantic-relatedness-and-semantic-similarity-and-how-to-calculate-them.

162. Semantic similarity. [Online] wikipedia.org.; 2019 [cited 2019 February 11]. Available from: https://en.wikipedia.org/wiki/Semantic_similarity.

163. Feng Y, Bagheri E, Ensan F, Jovanovic J. The state of the art in semantic relatedness: A framework for comparison. The Knowledge Engineering Review. 2017; 32.

164. Slimani T. Description and evaluation of semantic similarity measures approaches. arXiv preprint arXiv:1310.8059. 2013 October 30.

165. Leacock C. Filling in a sparse training space for word sense identification. PhD Thesis. Macquarie University; 1994.

166. Wu Z, Palmer M. Verbs semantics and lexical selection. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics; 1994: Association for Computational Linguistics. p. 133-138.

167. Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of artificial intelligence research. 1999 July 1; 11: p. 95-130.

168. Lin D. Principle-based parsing without overgeneration. In 31st annual meeting of the association for computational linguistics; 1993: Association for Computational Linguistics. p. 112-120.

169. Ghannay S, Favre B, Esteve Y, Camelin N. Word embedding evaluation and combination. In LREC; 2016. p. 300-305.

170. Word embedding. [Online] wikipedia.org.; 2019 [cited 2019 February 20]. Available from: https://en.wikipedia.org/wiki/Word_embedding.

171. Lai S, Liu K, He S, Zhao J. How to generate a good word embedding. IEEE Intelligent Systems. 2016 November; 31(6): p. 5-14.

172. A Beginner's Guide to Word2Vec and Neural Word Embeddings. [Online] skymind.ai.; 2019 [cited 2019 March 11]. Available from: https://skymind.ai/wiki/word2vec.

173. Wallach HM. Topic modeling: beyond bag-of-words. In Proceedings of the 23rd international conference on Machine learning; 2006: ACM. p. 977-984.

174. Minarro-Giménez JA, Marin-Alonso O, Samwald M. Exploring the application of deep learning techniques on medical text corpora. Studies in health technology and informatics. 2014; 205: p. 584-588.

175. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM. 1975 November 1; 18(11): p. 613-620.

176. Stolić P, Stolić S, Milosavljević A. Some of text analytics applications in higher education institutions. In Conference Proceedings of the 6th International Conference Technics and Informatics in Education; 2016; Čačak, Srbija: Faculty of technical sciences. p. 211-217.

177. Becker KG, Hosack DA, Dennis G, Lempicki RA, Bright TJ, Cheadle C, et al. PubMatrix: a tool for multiplex literature mining. BMC bioinformatics. 2003 December; 4(1): p. 61.

178. Mima H, Ananiadou S, Nenadic G, Tsujii J. A methodology for terminology-based knowledge acquisition and integration. In Proceedings of the 19th international conference on Computational linguistics; 2002: Association for Computational Linguistics. p. 1-7.

179. Müller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS biology. 2004 September 21; 2(1): p. e309.

180. Kijevčanin V, Gračanin Š. Data Mining. Research work. Kragujevac: Univeristy of Kragujevac; 2009.

181. Piateski G, Frawley W. Knowledge discovery in databases: MIT press; 1991.

182. Jovanović N. Alati za Data Mining. Master rad. Beograd: Univerzitet Singidunum, Departmant za poslediplomske studije; 2011.

183. Mao M. Ontology Mapping: An Information Retrieval and Interactive Activation Network Based Approach. In The Semantic Web.: Springer, Berlin, Heidelberg; 2007. p. 931-935.

184. Borbinha JL, Kapidakis S, Papatheodorou C, Tsakonas G. Research and Advanced Technology for Digital Libraries. In 13th European Conference. ECDL 2009, Corfu, Greece, September 27 - October 2, 2009, Proceedings.: Springer-Verlag Berlin Heidelberg; 2009. p. XIX, 497.

185. Salton G, McGill MJ. Introduction to modern information retrieval New York, NY, USA: McGraw-Hill, Inc.; 1986.

186. Maynard D, Yaoyong L, Peters W. Ontology Learning and Population: Bridging the Gap between Text and Knowledge Buitelaar P, Philipp C, editors.: Ios Press; 2008.

187. Hariri BB, Sayyadi H, Abolhassani H, Esmaili KS. Combining Ontology Alignment Metrics Using the Data Mining Techniques. In ECAI International Workshop on Context and Ontologies; 2006. p. 65-67.

188. Webster JJ, Kit C. Tokenization as the initial phase in NLP. The 15th International Conference on Computational Linguistics. 1992; 4: p. 1106-1110.

189. Stanojević M. Određivanje sličnosti između naučnih radova primenom metoda mašinskog učenja. Master rad. Beograd: Univerzitet u Beogradu, Elektrotehnički fakultet; 2017.

190. Manning C, Raghavan P, Schütze H. Introduction To Information Retrieval. Natural Language Engineering. 2010; 16(1): p. 100-103.

191. Brill E. A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing; 1992: Association for Computational Linguistics. p. 152-155.

192. Màrquez L, Padro L, Rodriguez H. A Machine Learning Approach to POS Tagging. Machine

Learning. 2000 April 1; 39(1): p. 59-91.

193. All About Stop Words for Text Mining and Information Retrieval. [Online] text-analytics101.rxnlp.com.; 2014 [cited 2019 February 11]. Available from: http://text-analytics101.rxnlp.com/2014/10/all-about-stop-words-for-text-mining.html.

194. Porter MF. An algorithm for suffix stripping. Program. 1980 March 1; 14(3): p. 130-137.

195. Lovins JB. Development of a stemming algorithm. Mech. Translat. & Comp. Linguistics. 1986 March 11; 11(1-2): p. 22-31.

196. Bleiholder J, Naumann F. Data fusion. ACM Computing Surveys (CSUR). 2009 January 15; 41(1): p. 1.

197. Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Transactions on knowledge and data engineering. 2007 January; 19(1): p. 1-16.

198. Suchanek FM, Abiteboul S, Senellart P. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. In Proceedings of the VLDB Endowment; 2011: VLDB Endowment. p. 157-168.

199. Wang C, Lu J, Zhang G. Integration of ontology data through learning instance matching. In IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06); 2006: IEEE. p. 536-539.

200. Castano S, Ferrara A, Lorusso D, Montanelli S. On the ontology instance matching problem. In 19th International Workshop on Database and Expert Systems Applications; 2008: IEEE. p. 180-184.

201. Daskalaki E, Plexousakis D. OtO matching system: a multi-strategy approach to instance matching. In International Conference on Advanced Information Systems Engineering; 2012: Springer, Berlin, Heidelberg. p. 286-300.

202. Otero-Cerdeira L, Rodríguez-Martínez FJ, Gómez-Rodríguez A. Ontology matching: A literature review. Expert Systems with Applications. 2015 February 1; 42(2): p. 949-971.

203. Gracia J, d'Aquin M, Mena E. Large scale integration of senses for the semantic web. In Proceedings of the 18th international conference on World wide web; 2009: ACM. p. 611-620.

204. Jean-Mary YR, Shironoshita EP, Kabuka MR. Ontology matching with semantic verification. Journal of Web Semantics. 2009 September 1; 7(3): p. 235-251.

205. Faria D, Pesquita C, Santos E, Palmonari M, Cruz IF, Couto FM. The agreementmakerlight ontology matching system. In OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"; 2013: Springer, Berlin, Heidelberg. p. 527-541.

206. Noessner J, Niepert M, Meilicke C, Stuckenschmidt H. Leveraging terminological structure for object reconciliation. In Extended Semantic Web Conference; 2010: Springer, Berlin, Heidelberg. p. 334-348.

207. Saıs F, Pernelle N, Rousset MC. L2r: A logical method for reference reconciliation. In Proc. AAAI; 2007. p. 329-334.

208. Bhattacharya I, Getoor L. Collective entity resolution in relational data. ACM Transactions on

Knowledge Discovery from Data (TKDD). 2007 March 1; 1(1): p. 5.

209. Saïs F, Pernelle N, Rousset MC. Combining a logical and a numerical method for data reconciliation. In Journal on Data Semantics XII.: Springer, Berlin, Heidelberg; 2009. p. 66-94.

210. Flouris G, Manakanatas D, Kondylakis H, Plexousakis D, Antoniou G. Ontology Change: Classification and Survey. The Knowledge Engineering Review. 2008 June; 23(2): p. 117-152.

211. Hassen W. Medley results for OAEI 2012. In Proceedings of the 7th International Conference on Ontology Matching; 2012: CEUR-WS. org. p. 168-172.

212. Hertling S. Hertuda results for OEAI 2012. In Proceedings of the 7th International Conference on Ontology Matching: CEUR-WS. org. p. 141-144.

213. Behkamal B, Naghibzadeh M, Moghadam RA. Using pattern detection techniques and refactoring to improve the performance of ASMOV. In 2010 5th International Symposium on Telecommunications; 2010: IEEE. p. 979-984.

214. Damerau FJ. A technique for computer detection and correction of spelling errors. Communications of the ACM. 1964 March 1; 7(3): p. 171-176.

215. Lin D. An Information-Theoretic Definition of Similarity. In Proceedings of the Fifteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1998. p. 296-304.

216. Lin F, Sandkuhl K. A survey of exploiting wordnet in ontology matching. In IFIP International Conference on Artificial Intelligence in Theory and Practice; 2008: Springer, Boston, MA. p. 341-350.

217. Socher R, Lin CC, Manning C, Ng AY. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th international conference on machine learning (ICML-11); 2011. p. 129-136.

218. Mikolov T, Zweig G. Context dependent recurrent neural network language model. In IEEE Spoken Language Technology Workshop (SLT); 2012: IEEE. p. 234-239.

219. Wang Z, Zhang X, Hou L, Zhao Y, Li J, Qi Y, et al. RiMOM results for OAEI 2010. Ontology Matching. 2010 November 7; 195.

220. Large BioMed Track (largebio). [Online] www.cs.ox.ac.uk.; 2019 [cited 2019 March 1]. Available from: http://www.cs.ox.ac.uk/isg/projects/SEALS/oaei/.

221. Mathur S, Dinakarpandian D. Finding disease similarity based on implicit semantic similarity. Journal of biomedical informatics. 2012 April 1; 45(2): p. 363-371.

222. Zhang R, Pakhomov S, McInnes BT, Melton GB. Evaluating measures of redundancy in clinical texts. In AMIA Annual Symposium Proceedings; 2011: American Medical Informatics Association. p. 1612.

223. Thada V, Jaglan V. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. International Journal of Innovations in Engineering and Technology. 2013 August; 2(4): p. 202-205.

224. Sidorov G, Gelbukh A, Gómez-Adorno H, Pint D. Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas. 2014 September; 18(3): p. 491-504.

225. Strehl A, Ghosh J, Mooney R. Impact of similarity measures on web-page clustering. In Workshop on artificial intelligence for web search (AAAI 2000); 2000. p. 64.

226. Veljković D. Tekst mining i model vektorskog prostora u funkciji klasifikacije bezbednosno interesantnih blogova. Bezbednost. 2016; 58(2): p. 223-242.

227. Manning CD, Raghavan P, Schütze H. Scoring, term weighting and the vector space model. In Introduction to Information Retrieval.: Cambridge University Press; 2008. p. 100-123.

228. Huang A. Similarity Measures for Text Document Clustering. In Proceedings of the 6th New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand; 2008. p. 49-56.

229. Berger FG, Berger SH. Thymidylate synthase as a chemotherapeutic drug target: where are we after fifty years? Cancer biology & therapy. 2006 September 1; 5(9): p. 1238-1241.

230. Uysal AK, Gunal S. The impact of preprocessing on text classification. Information Processing & Management. 2014 January 1; 50(1): p. 104-112.

231. Jiménez-Ruiz E, Grau BC. Logmap: Logic-based and scalable ontology matching. In International Semantic Web Conference; 2011: Springer, Berlin, Heidelberg. p. 273-288.

232. Tous R, Delgado J. A vector space model for semantic similarity calculation and OWL ontology alignment. In International Conference on Database and Expert Systems Applications; 2006: Springer, Berlin, Heidelberg. p. 307-316.

233. Savić-Pavićević D, Matić G. Molekularna biologija 1: NNK INTERNATIONAL; 2011.

Биографија

Марија В. Ђокић Петровић је рођена 11. фебруара 1986. године у Крагујевцу. Основну школу „Карађорђе“ завршила је у Рачи 2001. године, као носилац Вукове дипломе. Другу крагујевачку гимназију, општи смер, завршила је 2005. године као ученик генерације и носилац Вукове дипломе. На Природно-математички факултет у Крагујевцу, група Математика, смер математика-информатика, уписала се школске 2005/06, где је и дипломирала маја 2010. године са просечном оценом 8,92. Током студија учествовала је на EDIT летњој школи фирме *ComTrade ITSS* (Београд, Србија), а затим је у истој фирми обавила стручну праксу у трајању од два месеца, радећи на тестирању софтвера. Током студија била је корисник студентског кредита Министарства просвете и спорта, а последње године студија и стипендиста општине Рача. Школске 2010/2011 године уписује докторске студије из области рачунарских наука на Природно-математичком факултету у Крагујевцу.

Од маја 2011. године до јула 2017. године радила је као истраживач-приправник и истраживач-сарадник за ужу научну област Програмирање на Институту за Математику и информатику, Природно-математичког факултета, Универзитета у Крагујевцу. Била је ангажована на реализацији вежби из предмета „Основи програмирања” на студијском програму за стицање стручног/академског звања Дипломирани математичар и Дипломирани физичар, школске 2010/2011.

Своју професионалну каријеру наставила је у аустријским фирмама *Virtual World Services GmbH* из Беча и *mything GmbH* из Граца, преко којих била ангажована на пројекту Техничког Универзитета у Грацу, у Аустрији. Тренутно је као истраживач запослена на истраживачком центру *Virtual Vehicle* Техничког Универзитета у Грацу, у Аустрији. Рад на овом престижном Универзитету омогућио јој је усавршавање знања из области програмирања и наставак научно-истраживачког рада.

Члан је уредништва сајта *austria-forum.org*, који својим квалитетним садржајем, у виду многих биографија и публикација, представља значајан допринос аустријском друштву.

Члан је управног одбора аустријског удружења *NEU Graz* (Verein für Neoösterreicher und EU-Bürger Graz), где својом информатичком подршком доприноси развоју веб платформе удружења.

Активни је члан Друштва информатичара Србије. Њено ангажовање на Техничком Универзитету у Грацу у Аустрији покренуло је сарадњу овог Универзитета са Друштвом информатичара Србије.

Као члан управног одбора удружења *NEU Graz* активно се бави хуманитарним радом. Често помаже удружења и активисте који се боре за заштиту животиња.

Списак објављених радова

1. Cvjetkovic, V., **Djokic, M.**, Arsic, B., and Curcic, M. (2014). *The ontology supported intelligent system for experiment search in the scientific research center*. Kragujevac Journal of Science, (36), pp. 95-110. ISSN: 1450-9636.

2. **Djokic-Petrovic, M.**, Pritchard, D., Ivanovic, M. and Cvjetkovic, V. (2016), *IMI Python: Upgraded CS Circles web-based Python course.* Computer Applications in Engineering Education, 24(3), pp. 464 - 480, ISSN: 1061-3773, doi:10.1002/cae.21724.

3. **Djokic-Petrovic, M.**, Cvjetkovic, V., Yang, J., Zivanovic, M., & Wild, D. J. (2017). *PIBAS FedSPARQL: a web-based platform for integration and exploration of bioinformatics datasets.* Journal of biomedical semantics, 8(1), pp. 42. ISSN: 2041-1480.

4. Arsić B, **Đokić-Petrović M**, Spalević P, Milentijević I, Rančić D, Živanović M. (2018) *SpecINT: A framework for data integration over cheminformatics and bioinformatics RDF repositories*. Semantic Web - Interoperability, Usability, Applicability. Pre-press(Pre-press), pp. 1-19. DOI: 10.3233/SW-180327. ISSN: 1570-0844.

САОПШТЕЊА НА КОНФЕРЕНЦИЈАМА ШТАМПАНА У ИЗВОДУ ИЛИ У ПОТПУНОСТИ

5. V. M. Cvjetkovic, **M. Djokic**, B. Arsic, *Owl based modelling and visualisation of arbitrary semantic data structure*, Proceedings of the 5th International Conference "Science and Higher Education in Function of Sustainable Development'", Business Technical College, pp. 2:13-19, Uzice, October 04-05, 2012, http://sed.vpts.edu.rs/ (ISBN 978-86-83573-26-4)

6. V. Cvjetkovic, **M. Djokic**, B. Arsic, *Ontology Visualization*, Proceedings of the 1st Virtual International Conference on Advanced Research in Scientific Areas (ARSA-2012), vol. 1, issue 1, pp. 1999-2004, Slovakia, December 3 - 7, 2012, http://www.arsa-conf.com (ISBN: 978-80-554-0606-0, ISSN: 1338-9831)

7. V. Cvjetkovic, **M. Djokic**, B. Arsic, *Wikipedia Browsing with DBpedia*, Proceedings in EIIC - The 2nd Electronic International Interdisciplinary Conference, vol. 2, issue 1, pp. 470-475, 2013, http://www.eiic.cz/ (ISBN: 978-80-554-0762-3, ISSN: 1338-7871)

8. V. Cvjetković, **M. Djokic**, B. Arsić, *Semantically based customized search on local web site*, Proceedings of the 2nd Virtual International Conference on Advanced Research in Scientific Areas (ARSA-2013) vol. 2, issue 1, pp. 453--458, Slovakia, December 3 - 7, 2013, http://www.arsa-conf.com (ISBN: 978-80-554-0825-5, ISSN: 1338-9831)

9. **M. Djokic**, N. Stefanovic, *Application of Semantic Web in tourism information systems*, Proceedings of the 3rd International Conference on Information Society Technology (ICIST-2013), pp. 3:130-135 Kopaonik, March 3 - 6, 2013, http://e-drustvo.org/icist/2013/default.html (ISBN: 978-86-85525-12-4)

10. B. Arsic, **M. Djokic**, N. Stefanovic, *Mapping ebXML standards to ontology*, Proceedings of the 4th International Conference on Information Society Technology (ICIST-2014), vol. 1, pp. 198-203, Kopaonik, Serbia, March 9-13, 2014 (ISBN: 978-86-85525-14-8)

11. B. Arsic, **M. Djokic**, V. Cvjetkovic, P. Spalevic, M. Zivanovic and M. Mladenovic, *Integration of bioactive substances data for preclinical testing with Cheminformatics and Bioinformatics resources*, Proceedings of the 23nd International Electrotechnical and Computer Science Conference, vol. 1, issue 1, pp. 146-149, ERK, September 22-24, 2014, Portorož, Slovenia (ISSN:1581-4572)

12. B. Arsic, **M. Djokic**, V. Cvjetkovic, P. Spalevic and S. Ilic, *Semantic search framework for distributed semantically based cheminformatics and bioinformatics datasets*, 5th International Conference on Information Society and Technology (ICIST 2015), Society for Information Systems and Computer Networks, pp. 518-522, Serbia, 8-11. March, 2015 (ISBN: 978-86-85525-16-2)

13. V. Cvjetkovic, **M. Djokic**, *Semantic web based organization of scientific bibliography references*, The 3rd International Virtual Conference on Advanced Scientific Results (SCIECONF-2015), EDIS - Publishing Institution of the University of Zilina, vol. 1, no. 3, pp. 230-235, Slovakia, 25-29. May, 2015, pp. 230-235, 2015. (ISBN: 978-80-554-0891-0, ISSN: 1339-9071)

Техничко решење

14. V. Cvjetkovic, S. Markovic, B. Arsic, J. Zizic, **M. Djokic**, *Семантички базирана кастомизована претрага на локалном веб сајту: http://cpctas-lcmb.pmf.kg.ac.rs*, Универзитет у Крагујевцу, Природно-математички факултет, 2013.

Постер презентација

15. **M. Djokic-Petrovic**, D. Becejski-Vujaklija, A. Pajic-Simovic. (3-7 September 2018). *The status of women in the Serbian IT sector*. Poster session presented at the European Women in Mathematics General Meeting 2018, Karl-Franzens Univeristy Graz, Austria.

Учешће на пројектима

1. 2011-2017: Преклиничка Испитивања БиоАктивних Супстанци (евиденциони број пројекта III41010, пројекат Министарства републике Србије)

2. 2017-2018: Demabu (Технички универзитет Грац, Аустрија)
3. 2018- LDMA: Virtual Vehicle (Истраживачки центар, Технички универзитет Грац, Аустрија)
4. 2019- EVOLVE: Horizon 2020 (партнери: IBM, BMW AG, AVL List GmbH, Virtual Vehicle)
5. 2019- ContextEng: K2 Digital Mobility project (партнери: Audi AG, AVL List GmbH, Virtual Vehicle)

Djokic-Petrovic et al. Journal of Biomedical Semantics (2017) 8:42
DOI 10.1186/s13326-017-0151-z

Journal of Biomedical Semantics

SOFTWARE Open Access

PIBAS FedSPARQL: a web-based platform for integration and exploration of bioinformatics datasets

Marija Djokic-Petrovic[1,2*], Vladimir Cvjetkovic[2], Jeremy Yang[3,4], Marko Zivanovic[5] and David J. Wild[3]

Abstract

Background: There are a huge variety of data sources relevant to chemical, biological and pharmacological research, but these data sources are highly siloed and cannot be queried together in a straightforward way. Semantic technologies offer the ability to create links and mappings across datasets and manage them as a single, linked network so that searching can be carried out across datasets, independently of the source. We have developed an application called PIBAS FedSPARQL that uses semantic technologies to allow researchers to carry out such searching across a vast array of data sources.

Results: PIBAS FedSPARQL is a web-based query builder and result set visualizer of bioinformatics data. As an advanced feature, our system can detect similar data items identified by different Uniform Resource Identifiers (URIs), using a text-mining algorithm based on the processing of named entities to be used in Vector Space Model and Cosine Similarity Measures. According to our knowledge, PIBAS FedSPARQL was unique among the systems that we found in that it allows detecting of similar data items. As a query builder, our system allows researchers to intuitively construct and run Federated SPARQL queries across multiple data sources, including global initiatives, such as Bio2RDF, Chem2Bio2RDF, EMBL-EBI, and one local initiative called CPCTAS, as well as additional user-specified data source. From the input topic, subtopic, template and keyword, a corresponding initial Federated SPARQL query is created and executed. Based on the data obtained, end users have the ability to choose the most appropriate data sources in their area of interest and exploit their Resource Description Framework (RDF) structure, which allows users to select certain properties of data to enhance query results.

Conclusions: The developed system is flexible and allows intuitive creation and execution of queries for an extensive range of bioinformatics topics. Also, the novel "similar data items detection" algorithm can be particularly useful for suggesting new data sources and cost optimization for new experiments. PIBAS FedSPARQL can be expanded with new topics, subtopics and templates on demand, rendering information retrieval more robust.

Keywords: Federated SPARQL query, Bioinformatics, Data integration, Ontologies, Data mining and information retrieval

Background

Motivation

Nowadays, large amounts of bioinformatics data are publicly available to researchers of the life science community. These data and associated annotations are accessible through heterogeneous databases hosted as part of many independent and highly specialized resources and represented in different formats, conventions, vocabularies and ontologies. Still, modern research in bioinformatics greatly depends on the availability and efficient use of these data. Scientific research often requires access to various data points across scattered and highly distributed sources. This makes finding relevant data for scientific research projects a difficult and laborious task. With the rapid accumulation of bioinformatics data, this issue has only become more important and challenging.

* Correspondence: m.djokic@kg.ac.rs
[1]Virtual World Services GmbH, Asperner Heldenplatz 6, 1220 Wien, Austria
[2]Department of Mathematics and Informatics, Faculty of Science, University of Kragujevac, Radoja Domanovica 12, Kragujevac 34000, Serbia
Full list of author information is available at the end of the article

Kragujevac J. Sci. 36 (2014) 95-110. UDC 004.651:519.6:57.087:577

THE ONTOLOGY SUPPORTED INTELLIGENT SYSTEM FOR EXPERIMENT SEARCH IN THE SCIENTIFIC RESEARCH CENTER

Vladimir Cvjetković[1], Marija Đokić[1], Branko Arsić[1] and Milena Ćurčić[2]

[1]*Department of Mathematics and Informatics,* [2]*Department of Biology and Ecology,*
Faculty of science, University of Kragujevac,
Radoja Domanovića 12, 34000 Kragujevac, Republic of Serbia
E-mail: vladimir@kg.ac.rs

(Received January 24, 2014)

ABSTRACT. Ontologies and corresponding knowledge bases can be quite successfully used for many tasks that rely on domain knowledge and semantic structures, which should be available for machine processing and sharing. Using SPARQL queries for retrieval of required elements from ontologies and knowledge bases, can significantly simplify modeling of arbitrary structures of concepts and data, and implementation of required functionalities. This paper describes developed ontology for support of Research Centre for testing of active substances that conducts scientific experiments. According to created ontology corresponding knowledge base was made and populated with real experimental data. Developed ontology and knowledge base are directly used for an intelligent system of experiment search which is based on many criteria from ontology. Proposed system gets the desired search result, which is actually an experiment in the form of a written report. Presented solution and implementation are very flexible and adaptable, and can be used as kind of a template by similar information system dealing with biological or similar complex system.

Key words: Ontology, knowledge base, bio-experiments, active substances, SPARQL.

INTRODUCTION

The main role of information system is to support the operation of some real system which is mainly some enterprise or organization. Real system that is supported in this paper is the Research Center (RC) [3] for testing of active substances. Active substances are candidates for medicaments that are tested in laboratory, prior to being approved or not, for medical treatments. The RC is also the leader of the large Project [2] financed by the Ministry that consists of many institutions, departments, equipment and staff working on the project.

The subject of various analysis that are carried out at the RC includes monitoring of *in vitro* effects of active substances in the cell lines of different origin, primarily cancer cell lines and primary cells isolated from different tissues. Tests include cytotoxic active substances in human cancer cell lines, while monitoring includes the type of cell death, the mechanisms of apoptosis, migration and angiogenesis and prooxidant-antioxidant mechanisms which underlie the regulation of these processes. Tests are based on protocols such as MTT cytotoxicity test, AO/EtBr staining of cells for examination of the type of cell death, Western blot technique for examining proteins, Multiplex PCR, etc.

IMI Python: Upgraded CS Circles Web-Based Python Course

MARIJA DJOKIC-PETROVIC,[1] DAVID PRITCHARD,[2] MILOS IVANOVIC,[1] VLADIMIR CVJETKOVIC[1]

[1]*Faculty of Science, Department of Mathematics and Informatics, University of Kragujevac, Kragujevac, Serbia*

[2]*Google, Los Angeles, California*

Received 17 October 2015; accepted 28 January 2016

ABSTRACT: The rapid growth of student demand for flexible education and learning alternatives has caused a significant increase in web-based programming course offerings. In order to ensure easy and enjoyable ways of acquiring knowledge, many web-based solutions have customized the design and content to student needs. This paper introduces a project of the Institute for Mathematics and Informatics (IMI) called IMI *Python*, an interactive online course. It is based on the open-source Computer Science Circles (CS Circles) project. IMI Python aims to assist the target audience, primarily students, learn a spectrum of Python knowledge. The benefits of this enhanced system are multiple, both for students and their teachers. The course content is structured and divided by levels: basic, medium, and advanced. Flexible navigation through the different levels of difficulty and lesson units allows students to easily review any forgotten material and adopt new knowledge. Teachers have the ability to follow the progress of individual students or all students in a level, and communicate with them about their work. Teachers and students can communicate within the system to discuss individual exercises through a simple user interface. The system is enhanced with the possibility of testing students' knowledge through quizzes. Quizzes are visible at assigned time intervals and are worth a certain number of points. By tracking students' results, teachers can determine whether the site has enough quality material and what can contribute to its improvement. © 2016 Wiley Periodicals, Inc. Comput Appl Eng Educ 24:464–480, 2016; View this article online at wileyonlinelibrary.com/journal/cae; DOI 10.1002/cae.21724

Keywords: web-based course; Python

INTRODUCTION

In the last decade, web-based programming courses became a major trend in many institutions of higher education [1]. Faced with the rapid growth of demand, most courses are traditional frameworks, either just static information on a webpage, or exercises without feedback. Many online approaches try to innovate in the educational sense; they are improving their demonstrations of programming techniques by including examples, exercises with feedback, and environments that adapt to users' needs. But, some problems are still common [2]. Students can benefit from a richer approach that includes interaction between students and teachers. Following the course is difficult when there is a separation between lesson content and exercises; following all the details is quite hard and laborious work and requires constant student attention in the flood of information. Re-reading prior content before writing solutions to exercises also diverts student concentration, and is an obstacle to wholly understanding the programming language.

In order to overcome these deficiencies, and minimize the difficulties in the learning process, we created an improved web-based programming course called IMI Python (named after our Institute for Mathematics and Informatics). It is publicly available for free,[1] and we also use it in our classes. The system software is based on Computer Science Circles (CS Circles) [3], an open-source interactive platform for learning Python. We have extended and built upon its infrastructure. Previously, in the IMI, Python was taught in a workshop[2] aimed at Python beginners. The

Correspondence to M. Djokic-Petrovic (marija.djokic795@gmail.com, m.djokic@kg.ac.rs).

[1]http://147.91.205.71/wordpress
[2]http://imi.pmf.kg.ac.rs/moodle/course/view.php?id=288

SpecINT: A framework for data integration over cheminformatics and bioinformatics RDF repositories

Branko Arsić [a,*], Marija Đokić-Petrović [a,b], Petar Spalević [c], Ivan Milentijević [d], Dejan Rančić [d], Marko Živanović [a]

[a] *Faculty of Science, University of Kragujevac, Serbia*
E-mails: brankoarsic@kg.ac.rs, marija.djokic@virtualworldservices.at, zivanovicm@kg.ac.rs
[b] *Virtual World Services GmbH, Austria*
E-mail: marija.djokic@virtualworldservices.at
[c] *Faculty of Technical Sciences, University of Priština, Serbia*
E-mail: petar.spalevic@pr.ac.rs
[d] *Faculty of Electronic Engineering, University of Niš, Serbia*
E-mails: ivan.milentijevic@elfak.ni.ac.rs, dejan.rancic@elfak.ni.ac.rs

Editor: Michel Dumontier, Maastricht University, The Netherlands
Solicited reviews: Alasdair J G Gray, Heriot-Watt University, UK; Two anonymous reviewers

Abstract. Many research centers and medical institutions have been accumulating a vast amount of various biological and chemical data over the past decade and this trend continues. Based on Linked Data vision, many semantic applications for distributed access to these heterogeneous RDF (Resource Description Framework) data sources have been developed. Their improvements have brought about a decrease of intermediate results and optimizing query execution plans. But still many requests are unsuccessful and they time out without producing any answer. Also, the applications which operate over repositories taking into consideration their specificities and inter-connections are not available. In this paper, the SpecINT is proposed as a comprehensive hybrid framework for data integration and federation in semantic data query processing over repositories. The SpecINT framework represents a trade-off solution between automatic and user-guided approaches, since it can create queries which return relevant results, while not being dependent on human work. The innovativeness of the approach lays in the fact that the coordinates of graph eigenvectors are used for the automatic sub-queries joining over the most relevant data sources within repositories. In this way searching can be effected without a common ontology between resources. In experiments, we demonstrate the potential of our framework on a set of heterogeneous and distributed cheminformatics and bioinformatics data sources.

Keywords: Federated SPARQL query, Data Integration, Matrix Eigenvectors

1. Introduction

New data about chemical compounds, the influence they have on cancer cell-lines, genes and proteins, genetic variations and cell pathways have been emerging at a staggeringly rapid pace in recent chemical and biological experiments. Research centers and laboratories work independently storing data in different data formats with different vocabularies. The very abundance of heterogenic data sources prevents the life science community reaching its maximum. In this information vortex scientists need to put effort into finding and pairing relevant information over heterogeneous data within different data sources and consoli-

*Corresponding author. E-mail: brankoarsic@kg.ac.rs.

Образац 1

ИЗЈАВА АУТОРА О ОРИГИНАЛНОСТИ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Ја, Марија Ђокић Петровић, изјављујем да докторска дисертација под насловом:

Биоинформатичка платформа за извршавање Federated SPARQL упита над онтолошким базама података и детектовање сличних података утврђивањем њихове семантичке повезаности

која је одбрањена на Природно-математичком факултету Универзитета у Крагујевцу представља *оригинално ауторско дело* настало као резултат *сопственог истраживачког рада*.

Овом Изјавом такође потврђујем:

- да сам *једини аутор* наведене докторске дисертације,
- да у наведеној докторској дисертацији *нисам извршио/ла повреду* ауторског нити другог права интелектуалне својине других лица,
- да умножени примерак докторске дисертације у штампаној и електронској форми у чијем се прилогу налази ова Изјава садржи докторску дисертацију истоветну одбрањеној докторској дисертацији.

У Крагујевцу, 6.9.2019. године,

Марија Ђокић Петровић

потпис аутора

Образац 2

ИЗЈАВА АУТОРА О ИСКОРИШЋАВАЊУ ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

Ја, Марија Ђокић Петровић ,

☑ дозвољавам

☐ не дозвољавам

Универзитетској библиотеци у Крагујевцу да начини два трајна умножена примерка у електронској форми докторске дисертације под насловом:

Биоинформатичка платформа за извршавање Federated SPARQL упита над онтолошким базама података и детектовање сличних података утврђивањем њихове семантичке повезаности

која је одбрањена на Природно-математичком факултету Универзитета у Крагујевцу, и то у целини, као и да по један примерак тако умножене докторске дисертације учини трајно доступним јавности путем дигиталног репозиторијума Универзитета у Крагујевцу и централног репозиторијума надлежног министарства, тако да припадници јавности могу начинити трајне умножене примерке у електронској форми наведене докторске дисертације путем *преузимања*.

Овом Изјавом такође

☑ дозвољавам

☐ не дозвољавам[1]

[1] Уколико аутор изабере да не дозволи припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од *Creative Commons* лиценци, то не искључује право припадника јавности да наведену докторску дисертацију користе у складу са одредбама Закона о ауторском и сродним правима.

припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од следећих *Creative Commons* лиценци:

1) Ауторство

2) Ауторство - делити под истим условима

3) Ауторство - без прерада

4) Ауторство - некомерцијално

5) Ауторство - некомерцијално - делити под истим условима

6) Ауторство - некомерцијално - без прерада[2]

У Крагујевцу, 6.9.2019. године,

Марија Јакић Петровић

потпис аутора

[2] Молимо ауторе који су изабрали да дозволе припадницима јавности да тако доступну докторску дисертацију користе под условима утврђеним једном од *Creative Commons* лиценци да заокруже једну од понуђених лиценци. Детаљан садржај наведених лиценци доступан је на: http://creativecommons.org.rs/