УНИВЕРЗИТЕТ У НИШУ
ЕЛЕКТРОНСКИ ФАКУЛТЕТ

Бранко Ј. Арсић

СКАЛАБИЛНА СОФТВЕРСКА ПЛАТФОРМА ЗА ПРЕТРАЖИВАЊЕ ХЕМИЈСКИХ И БИОЛОШКИХ РЕПОЗИТОРИЈУМА

ДОКТОРСКА ДИСЕРТАЦИЈА

Ниш, 2020.

UNIVERSITY OF NIŠ
FACULTY OF ELECTRONIC ENGINEERING

Branko J. Arsić

A SCALABLE SOFTWARE PLATFORM FOR EXPLORING CHEMICAL AND BIOLOGICAL REPOSITORIES

DOCTORAL DISSERTATION

Niš, 2020.

Подаци о докторској дисертацији

Ментор:	Проф. др Иван З. Милентијевић, редовни професор, Универзитет у Нишу, Електронски факултет
Наслов:	Скалабилна софтверска платформа за претраживање хемијских и биолошких репозиторијума
Резиме:	У овој дисертацији је представљена *SpecINT* (*Spectral Integration*) скалабилна софтверска платформа која користи предности које пружа Семантички Веб и резултате спектралне теорије графова за интеграцију и претраживање семантички базираних репозиторијума. Како би се уштедело време и ресурси, све институције имају потребу за свеобухватним прегледом релевантних и последње објављеним подацима на глобалном нивоу. Статистика каже да од више хиљада испитаних супстанци, тек једна задовољава све критеријуме задате правилима преклиничких и клиничких тестирања и користи се као лек. У лабораторијским условима рада, потребно је пуно времена и ресурса да се испита ефекат овако великог броја супстанци, зато је неопходна примена *in silico* модела. Методологија која је коришћена за постизање овог циља се базира на координатама сопствених вектора графа које се користе за аутоматско повезивање подупита (patterns) у Federated SPARQL упит, при чему се у обзир узимају само најрелевантнији скупови података унутар репозиторијума. Овакав приступ омогућава смањивање броја дупликата у враћеним резултатима, али и добијање резултата који имају употребну вредност за истраживаче. На овај начин се интеграција репозиторијума постиже без заједничке онтологије између њих и тако ствара утисак о постојању централног, виртуелног складишта података које се претражује. Платформа је развијена у сарадњи са истраживачима Лабораторије за ћелијску и молекуларну биологију, Природно-математичког факултета у Крагујевцу. Ипак, методологија се може применити и шире, јер се базира на опште прихваћеним стандардима и концепту „Отворених података".
Научна област:	Рачунарске науке
Научна дисциплина:	Семантички Веб
Кључне речи:	Интеграција репозиторијума, Семантички Веб, софтверска платформа, спектрална кластеризација, теорија графова
УДК:	004.62/.65:519.177+004.822:576/577](043.3)
CERIF класификација:	P176: Вештачка интелигенција
Тип лиценце Креативне заједнице:	**CC BY-NC-ND**

Data on Doctoral Dissertation

Doctoral Supervisor:	Ivan Z. Milentijević, Ph.D., full professor at Faculty of Electronic Engineering, University of Niš
Title:	A scalable software platform for exploring chemical and biological repositories
Abstract:	This dissertation is about the *SpecINT* (*Spectral Integration*), a scalable software platform, which benefits from the Semantic Web and from the results of spectral graph theory for data integration and exploring semantic-based repositories. In order to save time and resources, institutions are in need of a comprehensive overview of relevant and most recently published data globally. According to statistics, only one substance out of thousands satisfies the preclinical and clinical tests' criteria and can be used as a medicament. Laboratory conditions require a lot of time and resources to test the effect of large number of substances, and that is why application of *in silico* models is found necessary. The methodology applied to achieve this goal is based on the coordinates of graph eigenvectors used for automatic join of sub-queries in Federated SPARQL query out of which only the most relevant data sources within repositories are taken into consideration. Such an approach enables reduction of number of duplicates in the results obtained, but also provides useful results for the researchers. In this way the integration of repositories can be effected without a common ontology between them, leaving an impression there exists a searchable central and virtual storage. The platform is developed in collaboration with the Laboratory for Cell and Molecular Biology of the Faculty of Science, University of Kragujevac. However, the methodology can be applied more broadly, since it is based on the „Open Data" standards and concepts.
Scientific Field:	Computer science
Scientific Discipline:	Semantic Web
Key Words:	Data integration, Semantic Web, software platform, spectral clustering, graph theory
UDC:	004.62/.65:519.177+004.822:576/577](043.3)
CERIF Classification:	P176: Artificial intelligence
Creative Commons License Type:	**CC BY-NC-ND**

Захвалница

Захвалница, текст који се обично пише на крају, а ипак стоји на почетку свега, тамо где му је и место. Текст који би требало да буде написан на формални начин и који треба да покаже оно што је немогуће показати, а то је бескрајна захвалност коју дугујем одређеним људима. Да ли је икако могуће ставити на тас све изречене речи подршке, сав предани рад, мотивацију и савете, стечене заслуге? Дрзнућу се да ово питање које прожима живот сваког од нас назовем Филозофским, јер одговор на њега немам.

Током израде докторске дисертације свих ових година, оно што сам видео, а можда нисам желео да себи признам, је чињеница да је породица она која највише губи и која је највише запостављена. Пре свега желим да се захвалим својој супрузи Јелени, својим родитељима, Југославу и Добрили, и брату Николи, што су ми били узор и веровали у мене, те пружили безрезервну подршку и помоћ да истрајем. Тешко би ми било да не поменем своју ћерку Нину која ми је, иако не зна нити једно слово, пружила несебичну помоћ у писању последњих реченица ове дисертације. Њено рођење је улило додатну снагу мом духу и пружило унутрашњи мир.

Ова докторска дисертација је резултат вишегодишњег рада и истраживања под менторством професора Ивана Милентијевића, редовног професора на Електронском факултету, Универзитета у Нишу. Професору Милентијевићу дугујем највећу захвалност за конструктивне сугестије и подршку да истрајем на научном плану. Његови савети су били велики мотив да своје оружје не положим на земљу, већ да са још већом снагом кренем напред. Посебну захвалност изражавам професорима Дејану Ранчићу и Петру Спалевићу што су ми пружили шансу и били уз мене од самог почетка студија у Нишу. Њихови пријатељски савети и смернице су у великој мери утицали на формирање мог истраживачког ја. Посебно се захваљујем колегиници Марији Ђокић-Петровић, јер је ова дисертација њеним највећим делом настала као резултат вишегодишњег заједничког и преданог рада. Као круне нашег рада стоје две докторске дисертације и пријатељство које траје.

Резултати који су приказани у овој дисертацији настали су као плод успешне сарадње Института за математику и информатику и Института за биологију и екологију, Природно-математичког факултета у Крагујевцу. Захваљујем се др Снежани Марковић, управници Лабораторије за ћелијску и молекуларну биологију Природно-математичког факултета у Крагујевцу и руководиоцу пројекта под називом „Преклиничка испитивања биоактивних супстанци (ПИБАС)”, а који финансира Министарство за просвету, науку и технолошки развој, Републике Србије (ИИИ41010). Захваљујем се свим члановима Лабораторије, јер су били отворени за сваки вид сарадње, што су ми омогућили коришћење њихових података за испитивање предложене методологије и што су свом снагом учествовали у интерпретацији и провери резултата. Такође се захваљујем

и Марку Живановићу, тада члану Лабораторије, за успешну сарадњу и конструктивне предлоге. Посебну захвалност дугујем и професору Владимиру Цвјетковићу који је поставио темеље информационог система Лабораторије, те самим тим омогућио потенцијалну интеграцију наших података и резултата са резултатима великих фармацеутских кућа и иницијатива.

Додатно желим да се захвалим академику Драгошу Цветковићу који ме је увео у свет спектралне теорије графова и колеги и пријатељу Милану Башићу који ми је помогао да теорију графова и област машинског учења спојим у једну целину, и да њихову примену сагледам кроз другачију призму. На крају желим да се захвалим свим колегама са Института за математику и информатику Природно-математичког факултета у Крагујевцу који су ми све ове године били ослонац и пружали безрезервну подршку.

Бранко Арсић

Садржај

Списак слика

Списак табела

Списак скраћеница

ANN	Artificial Neural Networks.
API	Application programming interface.
CPCTAS	Centre for PreClinical Testing of Active Substances.
DAWG	Data Access Working Group.
EBI	The European Bioinformatics Institute.
EM	Expectation-Maximization.
GO	Gene Ontology.
HTML	Hypertext Markup Language.
InChI	International Chemical Identifier.
IRI	International Resource Identifier.
IUPAC	International Union of Pure and Applied Chemistry.
KEGG	Kyoto Encyclopedia of Genes and Genomes.
LLD	Linked Life Data.
LOD	Linking Open Data.
LODD	Linked Open Drug Data.
Open PHACTS	Open Pharmacological Concept Triple Store.
OWL	Web Ontology Language.
PIBAS	Preclinical Investigation of BioActive Substances.
QSAR	Quantitative Structure–Activity Relationship.
RDBMS	Relational Database Management System.
RDF	Resource Description Framework.
RDFS	Resource Description Framework Schema.
SIO	Semanticscience Integrated Ontology.
SMILES	Simplified molecular input line entry specification.
SPARQL	SPARQL Protocol and RDF Query Language.
SWEO	W3C Semantic Web Education and Outreach Interest Group.
UniProt	Universal Protein Resource.
UPGMA	Unweighted Pair Group Method with Arithmetic Mean.
URI	Uniform Resource Identifier.
URL	Uniform Resource Locator.
W3C	World Wide Web Consortium.
WWW	World Wide Web.

Сажетак

Циљеви многих лабораторија широм света се базирају на идентификацији нових потенцијалних хемиотерапеутика за третирање канцера. Уопштено говорећи, лабораторије за хемијске синтезе обезбеђују новосинтетисане хемијске супстанце, које се даље испитују биолошким методама. Међутим, статистика каже да од 10 000 овако испитаних супстанци тек једна задовољава све критеријуме задате правилима преклиничких и клиничких тестирања и користи се као лек. У лабораторијским условима рада потребно је пуно времена и ресурса да се испита овако велики број супстанци на само једној ћелијској линији. Олакшавајућа околност је чињеница да су резултати многих експеримената спроведених у лабораторијама широм света постали доступни истраживачкој заједници и фармацеутским компанијама. Анализа онлајн доступних података може доста да олакша овај задатак и уштеди време и ресурсе. Претрагом Веба се може доћи до информација какву цитотоксичност одређене хемијске структуре показују на одређеним ћелијама канцера. Касније, ови се подаци могу употребити за креирање модела квантитативног односа структуре и активности (енг. *Quantitative Structure–Activity Relationship*) који су још познати под називом QSAR модели, а који се већ увелико користе у компанијама као што су Novartis, Bayer, Procter & Gamble итд.

У последњој деценији многи истраживачки центри и медицинске институције су током рада створиле и акумулирале огромну количину различитих билошких и хемијских података и овај тренд се наставља. Базирајући се на визији отворених повезаних података (*Linked Open Data*) током времена су се развиле многе апликације које обезбеђују дистрибуирани приступ хетерогеним RDF (*Resource Description Framework*) ресурсима. Ове апликације са собом доносе побољшања која се односе на смањивање међурезултата и оптимизацију плана извршавања упита. Ипак, многи захтеви се завршавају неуспешно и након завршетка не враћају никакве резултате. Такође, данас не постоје апликације које раде са више репозиторијума истовремено и узимају у обзир њихове специфичности и међусобне везе. У овој докторској дисертацији је описана софтверска платформа SpecINT (*Spectral Integration*) која представља компромисно решење (trade-off) између аутоматских и ручно навођених приступа, пошто је њена основна функција да креира упите који могу да обезбеде релевантне резулатате, независно од људског утицаја. Иновативност овог приступа који ће бити описан у наредним секцијама лежи у чињеници да се координате сопствених вектора графа искористе за повезивање подупита тако да се у обзир узму само релевантни подаци из различитих репозиторијума. На овај начин претраживање може бити изведено без заједничке онтологије која повезује репозиторијуме и коју је тешко направити. Кроз експерименте биће демонстриран потенцијал софтверске платформе на скупу хетерогених и дистибуираних хемијски и биолошки оријентисаних скупова података. Додатно, за потребе платфор-

ме је развијена посебна онтологија која се користи за фаворизацију одређених скупова података у зависности од корисничког питања које је прослеђено на улазу. Валидација свих добијених резултата је спроведена од стране истраживача из Лабораторије за ћелијску и молекуларну биологију Природно-математичког факултета у Крагујевцу за чије је потребе, у недостатку других решења, платформа и развијена.

Кључне речи: Семантички Веб, интеграција репозиторијума, спектрална теорија графова, сопствени вектори графова, Page Rank алгоритам, SPARQL упити

Глава 1

Увод

Нове информације о хемијским супстанцама и утицају које оне имају на ћелијске линије канцера, затим подаци о генима и протеинима, генетским варијацијама и ћелијским путањама почеле су страховито брзо да се нагомилавају након што се број изведених хемијских и биолошких експеримената драстично повећао у последњих двадесет година. Истраживачки центри и лабораторије своја истраживања углавном врше независно једни од других чувајући податке у различитим форматима и користећи различите речнике. Велика количина хетерогених скупова података, који су још додатно и дистрибуирани, спречава истраживачку заједницу (посебно у области природних наука) да ефикасније и ефективније постигне још бољи учинак. У циљу претраге велике количине информација, научници треба да уложе доста времена и труда за проналажење и упаривање релевантних појмова из хетерогених скупова података и репозиторијума, креираних са различитом сврхом и на различите начине. За постизање успешних перформанси биомедицинских истраживања и уштеде времена, процес интеграције података прераста у важан предуслов за превазилажење постојећих разлика у ресурсима. У раду [3] аутори су навели значај интеграције података у хемијској и био-информатици.

Како би се уштедели ресурси, све институције имају потребу за свеобухватним прегледом релевантних и последње објављених података на глобалном нивоу. На пример, ако се унапред зна какво ће понашање супстанца да испољи приликом хемијске реакције, на основу праћења резултата сличног или истог експеримента, лабораторијама се може омогућити да избегну скупе синтезе хемијских једињења или хемијских комплекса из биљака (комплекса), као и праћење нових хемијских реакција. Истраживања која ће бити представљена у овој докторској дисертацији обухватиће развој скалабилне софтверске платформе (у даљем тексту Платформа) за интеграцију и претраживање биоинформатичких и хемијско-информатичких RDF репозиторијума. Софтверска платформа је развијена у сарадњи са Лабораторијом за ћелијску и молекуларну биологију[1] (у даљем тексту Лабораторија) која је део Института за биологију и екологију на Природно-математичком факултету у Крагујевцу. Истраживања у Лабораторији се реализују у оквиру научно-истраживачког пројекта Министарства просвете, науке и технолошког развоја Републике Србије (Преклиничка испитивања биоактивних супстанци – ПИБАС, ИИИ 41010). Главна активност Лабораторије обухвата испитива-

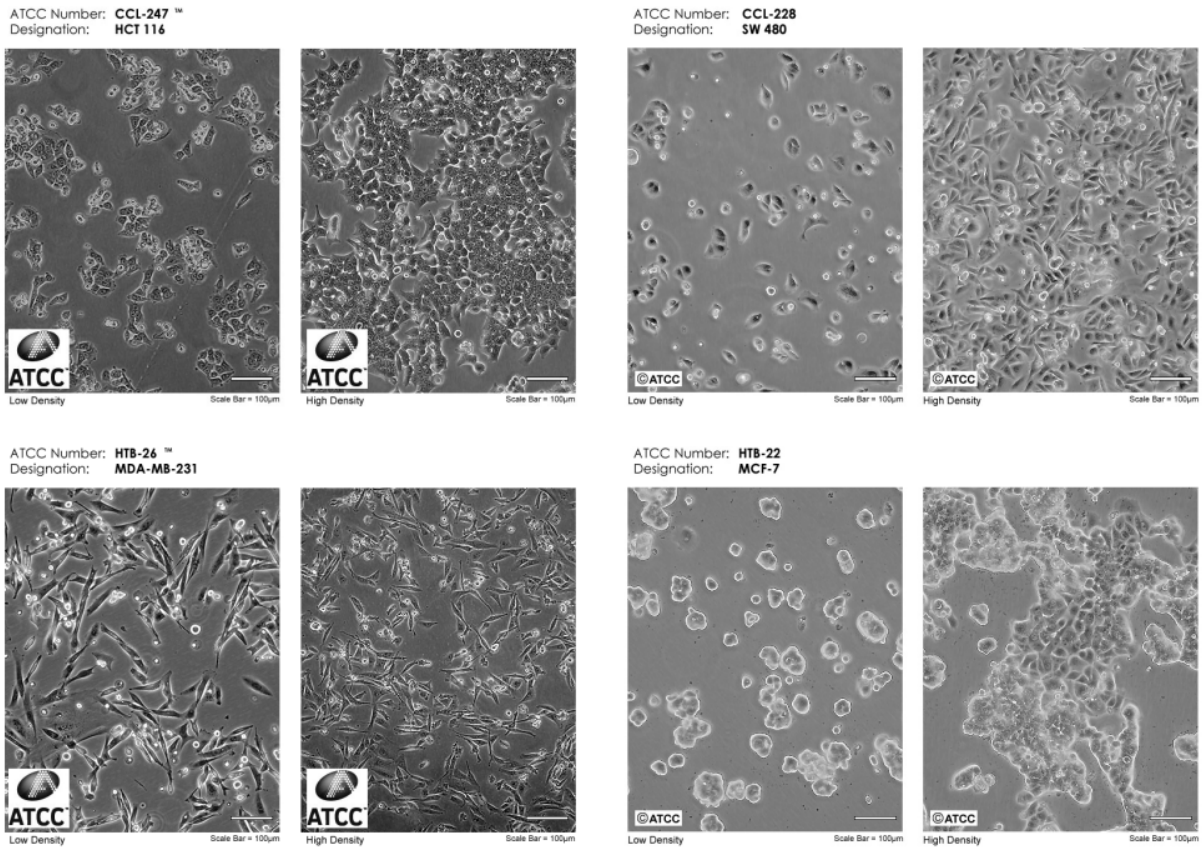
[1] CPCTAS-LCMB, Serbia, http://cpctas-lcmb.pmf.kg.ac.rs

ње цитотоксичности различитих супстанци на HCT-116 и SW-480 ћелијским линијама канцера колона помоћу MTT теста ћелијске вијабилности. *MTT тест* је колориметријска метода за одређивање вијабилности ћелија и процену цитотоксичности третмана у односу на задате ћелијске линије (изоловане популације ћелија способне за гајење у лабораторијским условима). *Federated SPARQL упит* на којима се Платформа базира биће дизајнирани у складу са потребама Лабораторије, али на такав начин да се покаже и њихова генерална важност и за друга биоинформатичка истраживања. *SPARQL* представља упитни језик који се извршава над подацима представљеним помоћу технологија Семантичког Веба, а претходно поменута реч *Federated* означава да упит истовремено приступа подацима који су смештени у дистрибуираном окружењу. У зависности од даљих потреба истраживача радиће се на унапређењу комплетног решења, генерисању нових упита и интеграцији нових скупова података.

Како бисмо што боље објаснили проблематику и изазове са којима се суочавају истраживачи у лабораторијама широм света најпре ће, као пример, бити објашњени сврха и протокол експеримената који се изводе у Лабораторији у Крагујевцу. У исто време биће уведена и битна терминологија која се користи кроз читаву дисертацију, а која ће бити од помоћи у разумевању предложене методологије за интеграцију репозиторијума.

1.1 Лабораторија за ћелијску и молекуларну биологију

Циљеви Лабораторије за ћелијску и молекуларну биологију базирани су првенствено на идентификацији нових потенцијалних хемиотерапеутика за третирање канцера. Лабораторија функционише на принципу мултидисциплинарног приступа истраживањима, јер лабораторије за хемијске синтезе обезбеђују новосинтетисане хемијске супстанце које се даље испитују биолошким методама. Преклиничка пракса испитивања нове супстанце, потенцијалног хемиотерапеутика, подразумева рад на модел системима, тзв. комерцијалним ћелијским линијама у *in vitro* условима. Ћелијска линија може се дефинисати као култура ћелија униформно изолована из ћелијске популације која потиче из реалног, обично хомогеног извора ткива (као што је орган). Данас су комерцијално доступне дефинисане, одређене и класификоване ћелијске линије свих органа у својим нативним (здравим) и патолошким (болесним) стањима. Лабораторија користи модел системе ћелијских линија канцера и њихових здравих аналога, а пре свега ћелијске линије канцера колона (HCT-116, SW-480) и дојке (MDA-MB-231, MCF-7) (за више детаља погледати Слику 1.1). За испитивање модел система ћелијских линија потребно је одржавати услове одгајања ћелија на оптималан начин, који симулира физиолошке услове у највећој мери, колико је то могуће. Ћелије се одгајају у стерилним условима у инкубатору у специјалним посудама за одгајање, уз додатак хранљивог медијума, одржавање температуре на 37 °C, влажности ваздуха и концентрације угљен диоксида од 5%, уз додатак антибиотика за спречавање могуће контаминације бактеријама.



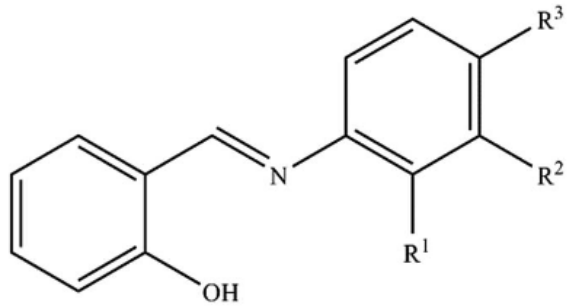
Слика 1.1: HCT-116, SW-480, MDA-MB-231 и MCF-7 ћелијске линије.

1.1.1 Експерименти у Лабораторији

Први експеримент који одређује биолошки утицај нове супстанце на испитивани модел систем јесте тест цитотоксичности. Данас постоји више тестова који одређују утицај хемикалије на вијабилност и цитотоксичност, али је најчешће примењивана тзв. MTT метода. MTT метода се заснива на бојеној реакцији супстанце (3-(4,5-диметилтиазол-2-ил)-2,5-дифенил тетразолијум бромид) са ензимом живих ћелија, митохондријалном дехидрогеназом. Сама супстанца је жуте боје, али у реакцији са ензимом даје љубичасто обојени кристал формазана. Овако створени љубичасти формазан одређује се спектрофотометријски, тј. одређује се интензитет његове боје квантификацијом абсорбанце на таласној дужини примењеног светла од 550 nm. Експеримент се изводи на популацији од најчешће 10 000 ћелија по узорку (у више поновљених експеримената због статистичке обраде). Што је већи број живих ћелија, већи је интензитет абсорбанце и обрнуто. Другим речима, уколико је нека супстанца токсична, она ће консеквентно смањити број ћелија, а самим тим и количину формазана, тј. мерену абсорбанцу. Како је метода заснована на релативном поређењу са контролним – нетретираним ћелијама, ми заправо добијамо релативан однос третирана/нетретирана ћелија. Конкретно, како би се цитотоксично дејство неке испитиване супстанце одредило, дели се измерена абсорбанца третираних ћелија са абсорбанцом негативне контроле (нетретиране ћелије) и множи се са 100, добијајући заправо проценте преживелих ћелија. Што је овај однос већи, већи је ефекат цитотоксичности. Пример конкретног експеримента представља третман ћелија у 6 концентрација у опсегу од 0.1 μM до 500 μM у 6 поновака. Уз то на исти начин се припреми и 6 поновака нетретираних ћелија, тј. негативне контроле. Након завршеног експеримента, на овај начин добије се по 6 процентуално изражених вредности вијабилности за сваку концентрацију и за контролу. Резултати се изражавају као вредност $\pm$ стандардна грешка за сваку концентрацију. Статистичка значајност се одређује Студентовим t-тестом и ANOVA тестом за поређења. p-вредност мања од 0.05 сматра се значајном. За добијање потпуне слике утицаја испитиване хемијске супстанце на вијабилност ћелијске линије, потребно је одредити додатно и тзв. IC_{50} вредност. IC_{50} вредност представља потенцијал супстанце да инхибира, у овом случају, ћелијску пролиферацију и дефинише се као половина максималне инхибиторне концентрације. За одређивање IC_{50} вредности користе се специјализовани рачунарски програми (нпр. *CalcuSyn*) или on-line платформе. Што је IC_{50} вредност нижа, то је цитотоксични ефекат примењене хемијске супстанце већи.

Статистика каже да од више хиљада овако испитаних супстанци, тек једна задовољава све критеријуме задате правилима преклиничких и клиничких тестирања и користи се као лек. У лабораторијским условима рада, потребно је пуно времена и ресурса уопште да се испита овако велики број супстанци на само једној ћелијској линији. Како су ова испитивања у својој основи временски и ресурсно захтевна, примена *in silico* модела је неопходна. Наиме, анализа литературних података у многобројним радовима доступних на интернету могу олакшати овај задатак. Правилном анализом релације хемијска структура-биолошки ефекат (преко IC_{50} вредности) може се доћи до претпоставке која би хемијска структура била што приближнија оптималној у циљу сузбијања раста одређене ћелијске линије. Ако посматрамо одређене ћелијске линије, употребом интелигентних алата можемо истражити у исто време много различитих хемијских супстанци са дефинисаним хемијским структурама у смислу испитивања

ефеката на задатим ћелијским линијама (пример хемијских структура је дат на Слици 1.2). Што је анализа детаљнија, тј. што је више хемијских супстанци испитивано на одређеним ћелијским линијама, расте вероватноћа да постоји одређени хемијско-структурни образац који најоптималније утиче на смањење вијабилности ћелијске линије. У том смислу, полазећи са краја процеса испитивања, *in silico* анализа нам обезбеђује повратну информацију о биолошком ефекту, која даље повратном спрегом даје потребну информацију лабораторији за хемијску синтезу. Оваква каскада повратних информација упућује нас ка синтези тачно специфичних хемијских једињења, која ће највероватније имати задовољавајући биолошки ефекат. На овај начин, постижу се значајне уштеде у времену и ресурсима са највећом вероватноћом успешних резултата. Коначно, овакав приступ утиче на смањење горе поменутог односа $1 : 10000$ и омогућава нам да у краћем временском року дођемо до потенцијалног лека. Сличан приступ може се употребљавати и за многе друге биолошке и хемијске експерименте који нису предмет рада Лабораторије.



1 R^1, R^2=H, R^3=OH
2 R^1, R^2=H, R^3=NO_2
3 R^1, R^2=H, R^3=CH_3
4 R^1, R^2, R^3=H
5 R^1, R^2=H, R^3=F
6 R^1=H, R^2=OH, R^3=H
7 R^1=OH, R^2, R^3=H

Слика 1.2: Schiff базе за супстанце. Преузето из [1].

У наставку су кроз два случаја коришћења описани начини на које се Платформа може употребити за унапређење истраживачких активности. На тај начин се оправдава неопходност постојања Платформе и показује мотивација за њено даље унапређивање.

1.2 Мотивација

Подаци добијени коришћењем софтверске платформе могу бити од велике користи хемичарима и биолозима, јер на овај начин истраживачи имају увид у анти-туморска својства хемијских комплекса који могу да открију потенцијалну стратегију у дизајнирању лекова базираних на металима. Истраживачи могу, на пример, да користе Платформу за повезивање биолошких (нпр. протеинске структуре и њихове путање) и хемијских података (лекови, интеракције са протеинима), као и за њихово приказивање на једном месту. Такође, на овај начин се може открити какав утицај имају изабране супстанце на одређене ћелијске линије канцера (тј. IC_{50} вредност за процену

и квантификацију цитотоксичности) и добити информације о генима и протеинима, формирајући тако кохерентну слику резултата и комплементарних података. Оно што Платформа пружа је увид у истраживачке трендове на глобалном нивоу у последњим годинама који ће послужити као основа за будућа истраживања. На пример, један од главних циљева модерног истраживања у биоорганској и медицинској хемији је развој нових лекова базираних на металима са фармацеутском активношћу која се разликује од терапија заснованим на платини [4]. Међу металским комплексима који не садрже плантину као метал за третирање канцера, деривати palladium(II) се доста користе због сличне структуре као Pt(II) комплекси који показују добру анти-туморску активност и мање нежељених реакција. Недавно су Петровић и др. [1] показали да избор одговарајућих лиганда може да доведе до тога да palladium(II) комплекси постану екстремно цитотоксични према ћелијама канцера. Платформа треба да омогући увид у комплексе који су показали добру антитуморску активност и тако утиче на даљи ток истраживања како се не би трошило време на комплексе од којих је тешко очекивати добре ефекте.

У Лабораторији је показано да Pt(IV), Pd(II), и Rh(III) комплекси индукују оксидативни стрес и цитотоксичност у HCT-116 ћелијским линијама канцера [5]. Такође, Живановић и др. [6] су испитивали биолошки ефекат хемијске структуре bicyclic selenohydantoin $(HidSe)$ и њеног palladium(II) комплекса са ознаком $((HidSe)2Pd)$ на $HCT-116$ ћелијским линијама дебелог црева човека и $MDA-MB-231$ ћелијским линијама груди. Открили су да $HidSe$ и $(HidSe)2Pd$ показују прооксидативни цитотоксични карактер, и јако анти-миграторни потенцијал на метастазе $MDA-MB-231$ ћелија. У овом случају Платформа треба да омогући не само проверу добијених резултата, већ и приступ резултатима других сличних експеримената како би истраживачи могли да планирају следеће кораке у истраживању. Скуп ових података треба касније да омогући прављење QSAR модела који ће на основу структуре супстанце моћи да предвиди њено дејство без извођења експеримената. Такође, ови резултати преко Платформе постају интегрисани са другим резултатима и доступни трећим странама.

1.3 Циљеви

Главни циљ овог научног истраживања је развој јавно доступне и функционалне софтверске платформе која користећи резултате спектралне теорије графова треба да формира виртуелно, централно складиште података које обједињује више репозиторијума и тако допуњава постојећа решења. Платформа ће најпре бити тестирана на реалним примерима, а њене функционалности и враћени резултати проверени од стране истраживача којима је она и намењена. Након тога, Платформа ће бити инсталирана на серверу и доступна истраживачкој заједници 24/7, као и сви резултати добијени у току израде дисертације. На овај начин ће истраживачима бити омогућен увид у резултате хемијских реакција које су спроведене у свету, а у којима су полазне супстанце управо оне које се разматрају за будућа испитивања. Једна од очекиваних функционалности Платформе је да у сваком тренутку можемо да проверимо која супстанца је добар инхибитор за одређене ћелијске линије канцера тј. чија IC_{50} вредност (концентрација која инхибира 50% ћелија) задовољава прописана ограничења у пре-клиничкој фази испитивања. Такође, поред увида у тако интегрисане податке, подаци се потен-

цијално могу искористити за прављење математичких модела који нам омогућавају да предвидимо IC_{50} вредност (регресија) за нову супстанцу, да сличне супстанце поделимо у кластере према структури или резултатима (кластеризација), или пак да се одређене супстанце доделе унапред задатим класама (класификација).

Циљеви који су постигнути израдом ове докторска дисертација и који ће бити детаљно описани у наредним секцијама су следећи:

(I) Креирана софтверска платформа[2], под називом SpecINT (*Spectral Integration*), за интеграцију RDF репозиторијума која се базира на на математичком концепту графа где се сопствени вектори графа се користе за избор релевантних скупова података и надовезивање шаблона (*patterns*). Сврха платформе је да унапреди рад истраживачких тимова у области природних наука, а пре свега биолошких и хемијских наука.

(II) У оквиру платформе је развијена процедура за конструисање графа чији су чворови различите репрезентације исте (сличне) супстанце у репозиторијумима, док су гране дефинисане постојећим триплетима у RDF графу који повезују те исте репрезентације супстанце. Сваки типлет је облика (субјекат, предикат, објекат), где је субјекат назив чвора од којег креће грана, објекат назив чвора у којем се завршава грана, док предикат означава везу између њих.

(III) Развијен алгоритам за аутоматско креирање Federated SPARQL упита који обухватају више репозиторијума на основу изабране путање у претходно добијеном графу. На основу координата одређених вектора графа, алгоритам омогућава препознавање скупова података у којима се релевантни подаци највероватније налазе, прикупљајући тако нове и комплементарне податке о супстанцама у реалном времену (on-the-fly). Константни статистички прорачуни и праћење новонасталих промена су избегнути.

(IV) Креирана база података са одговарајућим подупитима (*patterns*) која омогућава креирање валидних и сврсисходних Federated SPARQL упита. Додатно је развијена онтологија *RepoIntegration.owl* са крађим описом сваког скупа података која нам омогућава фаворизовање одређених скупова у зависности од постављеног питања.

(V) Скалабилна архитектура која омогућава лаку интеграцију нових скупова података. На овај начин друге институције могу веома једноставно да презентују и учине јавно доступним своја истраживања. Додавање нових и брисање старих скупова података је веома једноставно и не утиче на рад Платформе, нити је потребно радити додатно конфигурисање.

(VI) Тестиране функционалности платформе за различите супстанце од стране истраживача у Лабораторији (доменских експерата). Урађена провера релевантности враћених резултата и њихова анализа.

[2]http://147.91.203.161/specint

(VII) Извршено поређење Платформе са другим доступним платформама. За поређење враћених резултата биће коришћена Open PHACTS платформа на којој тренутно ради 27 партнера. Циљ овог поређења није да се покаже која је платформа боља, већ да се укаже на постојање комплементарних података и да су даља сарадња и интеграција, у светлу све веће експанзије података на Вебу, неопходни кораци.

(VIII) Подаци и резултати су постали доступни за целокупну истраживачку заједницу. Комплетан код је јавно доступан и може се репродуковати[3].

Пре него што било шта кажемо о SpecINT софтверској платформи која је развијена, битно је да најпре обратимо пажњу на начин на који су подацу смештени и повезани, као и на инфраструктуру који се користи за приступ тим подацима. Наредних неколико поглавља биће усмерено управо на преглед технологија Семантичког Веба како бисмо стекли представу о формату у којем се подаци налазе, о репозиторијумима са подацима у таквом облику, као и изазовима са којима се суочавају они који те податке желе да прочитају и интегришу. Након тога биће представљене функционалности платформе, математичка позадина која се користи у њеној основи и бенефити које њено коришћење доноси.

1.4 Преглед садржаја

Платформа која је развијена у оквиру ове докторске дисертације је веома комплексна и заснива се на широком спектру знања које је требало интегрисати у функционалну целину која доноси корист истраживачкој заједници. За почетак је неопходно упознати се са појмом „Отворених података", технологијама и начином на који се подаци представљају на Вебу. Затим треба препознати предности које овакав приступ доноси у односу на неке раније приступе, али и уочити њихове недостатке које не треба испуштати из видокруга током рада. Како би подаци постали доступнији и јаснији они се групишу у скупове података, а скупови у репозиторијуме, стварајући тако базу знања за одређене области на Вебу. Обично репозиторијуми настају као резултат постојања одређених питања из неке области, где се подаци повезују на такав начин да се приликом њиховог претраживања може лако доћи до одговора. Упознавање сваког репозиторијума подразумева комплетну његову анализу, детектовање потенцијалних уских грла на које се може наићи током рада, као и изналажење алтернатива за њихово превазилажење. Такође, битно је уочити разлике које постоје између самих скупова података у оквиру једног репозиторијума, а онда и разлике у начину предствљања и повезивања података између репозиторијума.

Посебна пажња у овој докторској дисертацији је посвећена интеграцији нових скупова података са већ постојећим великим иницијативама. Мале лабораторије морају да се прилагоде правилима игре које прописују велики играчи на Вебу уколико желе да им се прикључе и своје податке учине јавно доступним широј истраживачкој заједници, али и да добију приступ другим подацима на глобалном нивоу. У једном делу

[3] https://github.com/malibanekg/SpecINT

дисертације биће, на реалном примеру, описана процедура за једно могуће представљање података у облику „Отворених података" и начин на који се нови скуп података може повезати са другим скуповима.

Након интеграције података битно је упознати се и са начинима њиховог претраживања, зато су овде представљена тренутно доступна решења за претраживање семантички базираних скупова података. За свако решење су наведене све њихове предности и мане, побројани изазови на које се може наићи приликом њиховог коришћења, као и дати предлози за њихово превазилажење. Сама чињеница да не постоји слична платформа која се бави „онлајн" интеграцијом репозиторијума и њиховим претраживањем је нит која прожима ову дисертацију.

Иновативност приступа, који ће бити описан кроз наредне секције, се састоји у томе да се координате сопствених вектора графа искористе за повезивање подупита. Повезивање се врши на такав начин тако да се у обзир узимају само релевантни подаци из различитих репозиторијума. На овај начин претраживање може бити изведено без заједничке онтологије која повезује репозиторијуме и коју је тешко направити. Решење је имплементирано на основу одређених резултата из спектралне теорије графова, зато ће непосредно пре представљања архитектуре Платформе бити дат преглед математичких основа које су битне за њено разумевање. Након што сва неопходна знања буду детаљно описана, биће приказана архитектура Платформе и сви њени делови.

Друго поглавље описује технологије Семантичког Веба са посебним освртом на RDF (*Resource Description Framework*), OWL (*Web Ontology Language*) и *SPARQL* упитни језик. RDF је битан, јер приказује начин на који су подаци представљени и међусобно повезани што је од значаја за њихову каснију интерпретацију и претрагу. Онтологије се користе за проширивање семантике RDF-а и најчешће за представљање нових података са додатном семантиком, док се као излаз из софтверске платформе добијају Federated SPARQL упити. Након извршавања ових упита корисник добија приказ релевантних података за хемијску структуру која је задата као улаз платформи.

Треће поглавље уводи концепт повезаних података на Вебу кроз *LOD Cloud* и приказује тренутно доступне, семантички базиране, репозиторијуме чији је домен везан за хемијске структуре и њихове биолошке активности (један део облака). Репозиторијуми повезују одређене скупове података формирајући тако свој именски простор. Одређени репозиторијуми су представљени са више детаља, јер ће се провера предложене методологије радити управо над тим подацима.

Четврто поглавље приказује један начин на који се скуп података може представити помоћу технологија Семантичког Веба на примеру базе података која потиче из Лабораторије за ћелијску и молекуларну биологију у Крагујевцу.

Пето поглавље представља постојећа решења за претраживање RDF скупова података (репозиторијума). Свако решење је фокусирано само на одређени скуп података, јер је већина њих направљена са сврхом тестирања брзине упита, а не добијање употребљивих података на основу којих се могу донети ваљани закључци и планирати наредна истраживања. Посебан осврт ће бити направљен на аутоматске генераторе упита и кориснички оријентисане приступе, јер предложена софтверска платформа представља решење које узима добре стране ова два приступа.

Шесто поглавље приказује математички апарат који се користи за имплементацију логике платформе. У овом делу биће дате дефиниције и теоријски резултати из теорије графова. Један део овог поглавља је посвећен одређеним сопственим вредностима и сопственим векторима графова који носе информације о структури графа.

У **седмом поглављу** је представљена архитектура софтверске платформе која за унету хемијску структуру (InChIKey је улаз) генерише SPARQL упит (излаз) који претражује податке са више репозиторијумима, извршава га и кориснику приказује резултате претраге. Архитектура подразумева и позив процедуре на основу које се, коришћењем UniChem API-ја, креирају графови који одговарају репозиторијума, и позив алгоритма који платформа користи за избор најрелевантнијих скупова података за постављено питање и креирање финалног упита.

Осмо поглавље је одређено за представљање добијених резултата на основу провере коју су урадили истраживачи из Лабораторије. У овом делу биће приказане хемијске структуре које су коришћене за процес евалуације резултата, хеуристике које су тестиране за побољшање тачности резултата, као и начин избора иницијалних чворова у алгоритму. Затим ће бити објашњена процедура за валидацију и проверу резултата, и на крају приказани добијени резултати за сваку хемијску структуру, за сваку хеуристику посебно.

Девето поглавље садржи критичку дискусију која наводи све предности, али и недостатке представљене платформе, а упоредо са тим и предлоге за њено даље унапређење. Додатно су описане још неке могуће варијације платформе које се односе на начине повезивања графова репозиторијума, а које могу да доведу до одређених мањих промена у платформи.

Након описаних поглавља у закључку (**Поглавље 10**) је дат преглед резултата докторске дисертације и предлози за даље унапређење платформе.

На крају дисертације се налаз и један додатак у којем су дати математички докази неких теорема које могу бити значајне за унапређење истраживања.

Глава 2

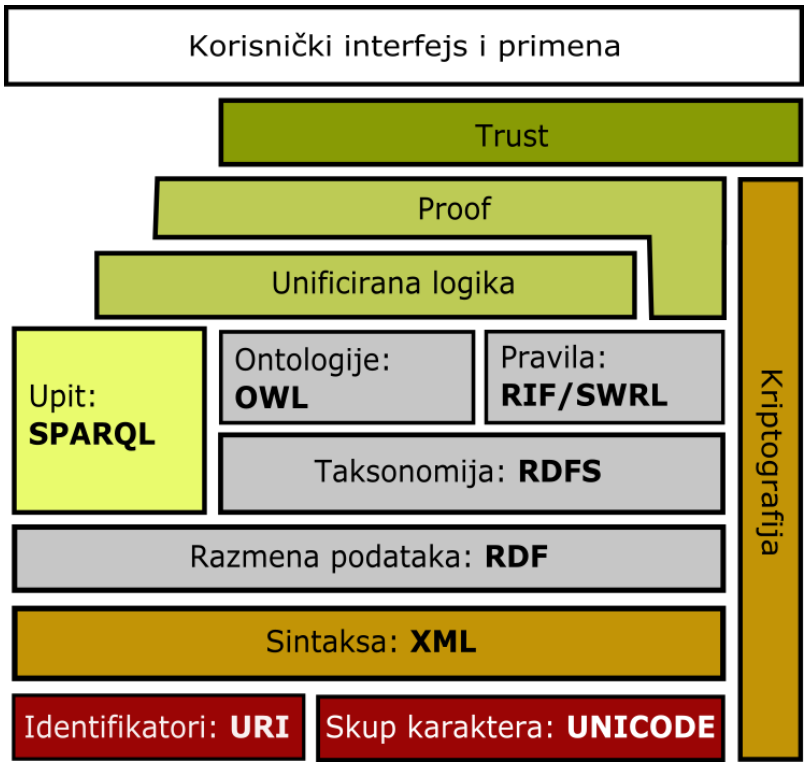
Технологије Семантичког Веба

Како бисмо боље разумели начин на који се подаци чувају у семантички базираним складиштима података, како су концепти међусобно повезани и инфрастуктуру која се користи за њихово претраживање, неопходно је да се најпре упознамо са готово свим технологијама Семантичког Веба. У првом делу поглавља дајемо кратак увод у парадигму која се зове Семантички Веб, његове концепте и намену, као и мотивацију за његово увођење. Након тога се упознајемо са основним градивним блоковима који дефинишу његове концепте, као и њихове формалне дефиниције и нотације.

2.1 Семантички Веб

Пројекат *World Wide Web (WWW)* је покренуо Tim Berners-Lee током раних 80-их, а сам пројекат је основан као систем међусобно повезаних ресурса који формирају мрежу којој се може приступити преко Интернета [7]. Документима који су смештени на удаљеним серверима се може приступити помоћу глобалне јединствене адресе - *Uniform Resource Locator (URL)*, која пружа локацију преко које се може приступити ресурсу на Интернету. У већини случајева ови документи се састоје од хипертекста [8] који је написан на основу спецификације за *Hypertext Markup Language (HTML)*. Како је Интернет постајао бржи и доступнији тако је и расла његова популарност. Људи све више времена проводе на Интернету живећи у исто време у "паралелном свету" који постаје њихов извор најновијих информација. Данас се многи вебсајтови користе за друштвено умрежавање као што су Twitter[1], Facebook[2] и LinkedIn[3] итд; за смештање видео снимака као што су dailymotion[4], YouTube[5]; затим за праћење оцена на Интернет бази података за филмове (IMDb[6]); читање онлајн енциклопедија као што је WIKIPEDIA[7] итд. Све набројани сајтови су заправо постали прави сервиси за размену података и интеракцију са кориснички генерисаним садржајем у кориснички оријентисаном окружењу

[1] https://twitter.com/
[2] https://facebook.com/
[3] https://linkedin.com/
[4] http://www.dailymotion.com/
[5] https://youtube.com/
[6] http://www.imdb.com/
[7] https://www.wikipedia.org/



Слика 2.1: Слојеви Семантичког Веба.

[7]. Међутим, тренутни Веб садржи огромну количину информација која је генерално разумљива само људима, а то постаје велики проблем у ери велике количине података (*Big Data*). Људи више не могу да прегледају толику количину генерисаног садржаја који се акумулирао у последњој деценији и тај тренд експоненцијално расте.

Општији контекст од WWW је Семантички Веб који представља Веб података (*Web of data*) који се лако могу прочитати и обрадити од стране машина. Главна сврха постојања Семантичког Веба је да дозволи и људима и машинама да уређују податке који се могу наћи на Вебу. На овај начин они имају могућност да донесу закључак о значењу података из информација и тако асистирају кориснику у току његових активности (”Где се налази најближе такси стајалиште?”). Треба имати у виду да Семантички Веб није замена за Веб 2.0, већ само његова екстензија што потврђује и Tim Berners-Lee, идејни творац Семантичког Веба, који је први увео овај термин 2001. године [9]:

”Семантички Веб није одвојен од Веба, већ представља његову екстензију у којој су информацијама дата добро дефинисана значења, омогућавајући бољу комуникацију између људи и рачунара.”

Сви циљеви Семантичког Веба су усмерени ка представљању заједничког оквира који омогућава да се подаци користе и деле између апликација и заједница. Да би се постигли ови циљеви потребно је направити неколико корака у стандардизацији. Радна група у оквиру World Wide Web Конзорцијума (W3C)[8] је направила неколико препорука и стандарда који описују делове Семантичког Веба од којих је он изграђен. Семантички Веб се најчешће неформално представља као скуп наслаганих слојева (ви-

[8]https://www.w3.org/

дети Слику 2.1) који је креирао Tim Berners-Lee, а сваки од њих је детаљно описан у [10]. Ова репрезентација приказује основне концепте Семантичког Веба као и њихову хијерархијску структуру, јер сваки слој користи функционалности оних слојева испод њега. На слици се могу приметити три основна блока при чему су на дну слојеви који представљају базу Семантичког Веба URI [11], Unicode [12] и XML [13]. У средини се налазе слојеви који чине стандарди који се користе за креирање апликација (према W3C): RDF [14], SPARQL [15, 16], RDFS [17], OWL [18] и RIF [19], док се на врху налази лепеза технологија које још увек нису стандардизоване: криптографија (*Cryptography*), "Trust"-слојеви и кориснички интерфејс. У наставку овог поглавља биће описани скоро сви стандарди које је развио W3C што ће допринети бољем разумевању формата у којем се подаци чувају и међусобно повезују, а посебна пажња биће посвећена начину на који се подаци претражују.

2.2 Resource Description Framework (RDF)

Resource Description Framework (RDF) је језик који је стандардизован од стране W3C који треба да структурне информације на Вебу представи у облику графова [14], са посебним акцентом на метаподатке о ресурсима. Оно што је, можда и битније, у овом тренутку је чињеница да се постојећи Веб садржај (Web 2.0) може проширити како би се описале ствари које се већ могу идентификовати на њему, као што је опис истраживача, или везе између њих. RDF је више намењен за аутоматску обраду података, пре него да буде пријемчив људима, што се може закључити по самој контрукцији, пошто његов формализам уопште није кориснички оријентисан ("user-friendly"). Штавише, фрејмворк је и дизаниран за дељење и размену података између апликација и скупова података.

У основи, RDF означава ствари и концепте коришћењем *Uniform Resource Identifiers (URIs)*, концепт који ће у Подсекцији 2.2.3 бити детаљније описан, а везе између њих описује помоћу једноставних својстава. Тренутна RDF спецификација [14] је подељена на шест W3C препорука. Најважнији документ је RDF Primer који уводи RDF концепте и садржи основне информације које су потребне за успешно коришћење RDF-а. Додатно, RDF Primer описује како да се дефинишу речници коришћењем *RDF Vocabulary Description Language* (познатог још као *RDF Schema*).

2.2.1 RDF мотивација и циљеви

Развој RDF-а је био мотивисан следећим случајевима коришћења:

- **Веб метаподаци:** Пружање информација о Веб ресурсима и системима који их користе у случају да они желе да те информације о себи буду доступне (опис садржаја, приватност, аутори, датуме креирања или модификације, да додатно опишу могућности које пружају и слично).

- **Флексибилност:** Апликације које захтевају отворене информације, пре него да те информације буду ограничене (уговорене активности у распоредима, организациони процеси,...).

- **Независност:** Аутори могу да дозволе да се подаци обрађују ван окружења у којем су креирани.
- **Повезивање:** Комбиновање подата из различитих ресурса да би се добиле нове информације.
- **Лингва франка (Lingua Franca):** Софтверски агенти могу да обрађују податке аутоматски и тако креирају глобалну мрежу процеса који сарађују директно уместо да морају да разумеју садржај који је само људима јасан.

Узимајући све ово у обзир, RDF је дизајниран да представи информације на минимално ограничен и флексибилан начин. Заиста, RDF се може користити за изоловане апликације, или између неколико различитих, или чак у раличитим пројектима који не деле исте циљеве. Вредност самих информација је побољшана коришћењем RDF-а, јер подаци постају доступни преко Интернета. Још прецизније, RDF је дизајниран да постигне следеће циљеве:

- представља једноставан начин за представљање података како би се лакше користили у апликацијама.
- поседује формалну семантику за основно закључивање (енг. *reasoning*); поседује правила за закључивање у RDF подацима.
- поседује речник који се може проширити захваљујући коришћењу URI-ја (*Uniform Resource Identifiers*) тако да може бити коришћен за именовање свих врста ствари у RDF-у.
- пружа синтаксу за енкодирање која се базира на XML-у, дељење и размену скупова података између апликација.
- дозвољава свакоме да доноси закључке/даје изјаве о било којем ресурсу.

2.2.2 RDF концепти

Као што је претходно објашњено, циљ RDF-а је презентовање изјава о ресурсима на Вебу. Како би се увела основна идеја, проста чињеница може, на пример, бити исказана на следећи начин: особа по имену Јелена је развила алгоритам за структурну регресију. Таква чињеница може бити исказана на српском језику у следећој реченици:

Пример 1: Алгоритам ”GCRF” има аутора, а то је Јелена.

Лако се могу приметити три основна (наглашена) елемента у овој реченици:

(a) елемент који реченица описује (Алгоритам ”GCRF”),

(b) одређено својство које нас се тиче (аутор) и

(c) елемент који је вредност својста (Јелена).

Користећи исти тип реченице, алгоритам можемо додатно да опишемо на следећи начин:

Примери 2 и 3:

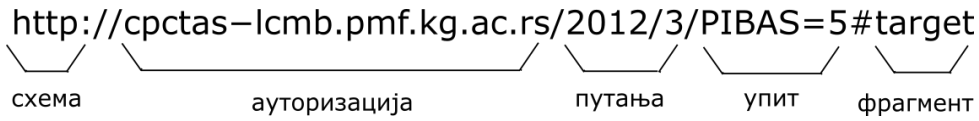
> Алгоритам ”GCRF” припада (има) области машинског учења која се назива структурна регресија, а која је подобласт алгоритама базираних на графовима.
>
> Алгоритам ”GCRF” има годину развоја која је 2019-та.

RDF изјаве могу бити конструисане коришћењем сличних структура као оне које су представљене у претходна три примера, јер RDF креће од претпоставке да ствари које се описују имају својства, која пак имају своје вредности. У исто време, RDF уводи посебну терминологију за одређивање различитих делова изјаве, управо оних делова који су били наглашени (као што је то случај у природним језицима): *субјекат* је нешто о чему је изјава, *предикат* се односи на својство субјекта које га ближе одређује, а *објекат* је његова вредност. Да би био у широкој употреби, RDF треба да се обради помоћу машина, што повлачи да реченице морају бити недвосмислене (што је иначе случај са природним језицима), а изјаве представљене једноставном синтаксом због машина. Веб већ има начин да идентификује странице на Интернету помоћу URL-ова (Uniform Resource Locators), стрингова који представљају механизам за приступ ресурсима на Вебу (обично кроз мрежу). Ипак, веома је важно да постоје референце на ствари и објекте који немају Веб странице, па самим тим ни URL. Из тог разлога, RDF користи генералнији коцепт за идентификацију - *Uniform Resource Identifiers (URIs)* који обухвата URL. URI има својство да није ограничен приступом преко Веба, па тако може да референцира било шта, нпр. програмере, математичаре, дрвеће у улици, апстрактне концепте као што су боја, догађаји итд. RDF користи сличну синтаксу као језици за означавање приликом употребе URI-ја.

2.2.3 URI базирани RDF модел

URI (Uniform Resource Identifier) представља основу Семантичког Веба (видети Слику 2.2), јер RDF користи URI референце за идентификацију концепата. URI се може посматрати као стринг који прати предефинисани скуп синтаксних правила и који недвосмислено идентификује одређени ресурс. На пример, он се може користити као замена за ентитет или концепт (особа, место, историјски догађаји, врста животиња, итд.) који се даље повезују са другим концептима и ентитетима. Коришћење URI-ја недвосмислено дефинише и референцира апстрактне и конкретне концепте на глобалном нивоу. RFC 3986[9] стандард дефинише генеричку синтаксу за URI-је која садржи хијерархијски низ компоненти као што су схема, ауторитет, путања, упит и фрагмент који се могу видети на Слици 2.2. Шема и хијерархијски организовани делови су обавезни, при чему су упит и фрагмент опциони. У контексту Семантичког Веба, URI се користи да дефинише концепте као што су места, организације или особе, али исто тако и за дефинисање веза између концепата, на пример, особа-је-рођена-у-граду, глобално и недвосмислено.

[9] http://tools.ietf.org/html/rfc3986



Слика 2.2: Пример URI-ја и његових компоненти према RFC 3986.

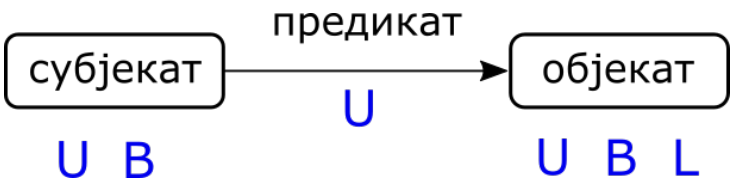
Пример 1 из претходне подсекције се сада на основу нашег именског простора и Dublin Core Metadata Initiative (спецификација за одређене метаподатке) може написати на следећи начин:

- **субјекат:** http://imi.pmf.kg.ac.rs/algorithm/GCRF
- **предикат:** http://purl.org/dc/elements/1.1/creator
- **објекат:** http://imi.pmf.kg.ac.rs/saradnici/Jelena

У наставку дајемо неколико формалних дефиниција претходно поменутих термина како бисмо прецизно знали шта све може да буде део изјаве. На почетку уводимо два концепта која су директно повезана са RDF стандардом која се односе на различите скупове ентитета који могу да буду RDF изјаве [20, 21].

Дефиниција 2.1 (RDF термини)**.** Означимо са U, L и B међусобно дисјунктне, бесконачне скупове где је U је скуп свих URI-ја, L скуп свих литерала, а B скуп свих празних чворова. Тада се скуп свих RDF термина, у ознаци RDF_T, дефинише као унија скупова U, L и B тј. $RDF_T = U \cup L \cup B$.

Дефиниција 2.2. *RDF тријлет* се дефинише као уређена тројка $(s, p, o) \in (U \cup B) \times U \times (U \cup B \cup L)$, где је s субјекат, p предикат и o објекат уређене тројке (видети Слику 2.3).



Слика 2.3: RDF триплет.

Треба имати у виду да се коришћење празних чворова не препоручује за концепт ”Повезаних података”, јер представљају ресурсе за које се не специфицира име и који се идентификују преко ID-ја који не мора бити јединствен у скупу [22].

У облику триплета претходне примере можемо записати као:

```
(<http://imi.pmf.kg.ac.rs/algorithm/GCRF>,
      <http://purl.org/dc/elements/1.1/creator>,
            <http://imi.pmf.kg.ac.rs/saradnici/Jelena>)

(<http://imi.pmf.kg.ac.rs/algorithm/GCRF>,
      <http://purl.org/dc/elements/1.1/isPartOf>,
            <http://imi.pmf.kg.ac.rs/area/StructuralRegression>)
```

```
(<http://imi.pmf.kg.ac.rs/algorithm/GCRF>,
       <http://purl.org/dc/elements/1.1/date>, "2019")
```

Листинг 2.1: Примери триплета.

Као што се може приметити из претходних примера, правилан запис једног триплета садржи комплетне URI-је за субјекат, предикат и објекат, што резултује у јако дугим записима између угластих заграда "<" и ">". *RDF Primer* уводи скраћени опис триплета помоћу префикса како би запис био јаснији и читљивији. Неки примери префикса су наведени у Табели 2.1, а начин њихове употребе у Листингу 2.2.

Табела 2.1: Примери RDF префикса.

Префикс	**Именски простор**
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
owl:	http://www.w3.org/2002/07/owl#
pibas:	http://cpctas–lcmb.pmf.kg.ac.rs/2012/3/PIBAS#
foaf:	http://xmlns.com/foaf/0.1/

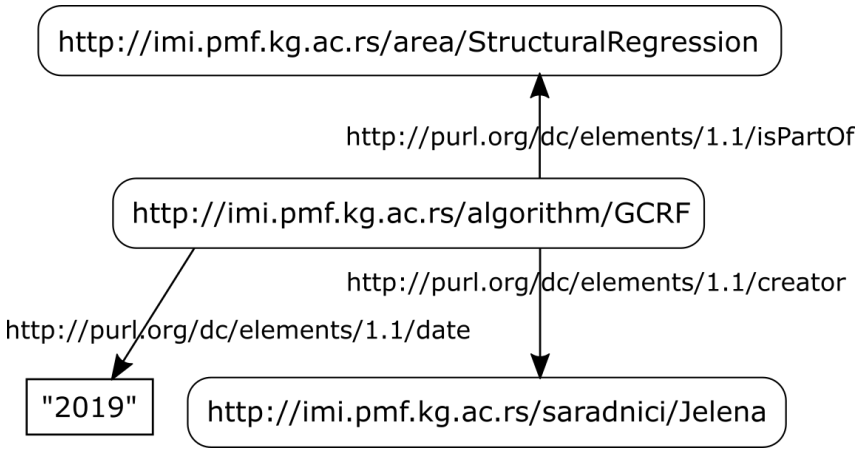
```
PREFIX imi: <http://imi.pmf.kg.ac.rs/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
imi:algorithm/GCRF dc:creator imi:saradnici/Jelena
imi:algorithm/GCRF dc:isPartOf imi:area/StructuralRegression
imi:algorithm/GCRF dc:date "2019"
```

Листинг 2.2: Коришћење префикса.

Дефиниција 2.3. *RDF скуп података* је скуп RDF триплета и може се дефинисати као $D = \{(s_1, p_1, o_1), (s_2, p_2, o_2), \ldots, (s_n, p_n, o_n)\}$.

Алтернативни начин за приказивање триплета су RDF графови (пример је приказан на Слици 2.4). RDF граф је мултиграф у коме су URI-ји повезани гранама које су исто означене URI-јима преузетих из предефинисаног скупа. Ове ознаке су такође познате и као предикати.

Дефиниција 2.4. *RDF граф* G је оријентисани мултиграф (N, D, E) где је N коначни скуп чворова такав да је $N \subset U \cup L$, D је коначан скуп предиката такав да је $D \subset U$, а E коначан скуп обележених тежинских грана у форми $\langle\langle s, p, o\rangle, c\rangle$ тако да је извор (субјекат) гране чвор $s \in N \cap U$, а понор (објекат) чвор $o \in N$, ознака гране $p \in D$ и тежина гране c је ненегативан број.



Слика 2.4: Пример RDF графа.

2.3 RDF Schema (RDFS) и Web Ontology Language (OWL)

Једна од главних снага RDF-а је да може да се користи као језик за моделовање мета података (енг. *meta-modeling*). *World Wide Web Consortium (W3C)* је представио речник који је познат као RDF-Schema, скраћено RDFS[10], који пружа основни језик за онтологије и који је намењен за додатну семантику RDF скупова података. RDFS речник дефинише додатна својства као што су: *Classes*, *SubClasses*, *Properties*, *SubProperties*, *Ranges* и *Domains*. Такође, он пружа стандардни речник за дефинисање типова података, листе (*collections*) и контејнере (*containers*). Такође, постоје и два својства која се доста често користе за повезивање ресурса са описом и именом, *rdfs:comment* и *rdfs:label*, који су пријемчивији човеку.

RDF-Schema (RDFS) је предложена од стране W3C као семантичко проширење RDF-а, са циљем дефинисања речника који се користи у RDF графу, као и да опише везе између ресурса и везе између ресурса и својстава [23]. RDFS је темељ онтолошког закључивања у Семантичком Вебу. Други језици који проширују RDF-Schema су недавно дефинисани, где са посебним осврtом помињемо W3C језик *Web Ontology Language (OWL 2[11])*. Многе имплементације за RDFS и OWL су развијене подржавајући евалуацију и закључивање над RDF подацима [24, 25]. Исто као и RDF граф, *RDF-Schema* је такође мултиграф за предефинисаним скупом ознака за гране. Помоћу RDFS сада је могуће дефинисати класе којима одређени ресурси припадају. На пример, ресурс ”мачка” припада класи ”животиња” или класи ”породица мачака”. Са RDFS је такође могуће дефинисати везе између класа и предиката RDF графа (*rdfs:range, rdfs:domain, rdf:type* итд). У наставку је дата дефиниција онтологије која је повезана са RDF скупом података користећи RDFS речник. Онтологије представљају формализоване речнике термина које често обухватају одређени домен. Оне садрже дефиниције тих термина описујући у исто време и њихове међусобне везе.

Дефиниција 2.5. *Онотолоīија* K је орјентисани граф (N_K, E_K) где сваки чвор из скупа N_K представља класу или својство, а свака грана из E_K има ознаку из скупа $\{rdfs$:

[10] https://www.w3.org/TR/rdf-schema/
[11] *https://www.w3.org/TR/owl2-syntax/*

$subClassOf, rdfs:subPropertyOf, rdfs:domain, rdfs:range\}$ који представља један фрагмент RDFS речника.

Данас је истраживачкој заједници доступан велики број био-онтологија. Све оне су развијене како би се омогућила лакша интеграција података из литературе и јавно доступних биомедицинских база података. Према BioPortal-у, највећем репозиторијуму биомедицинских онтологија [12], све онтологије су сврстане у више од 30 категорија као што су: Генеричке и "Над-онтологије" (енг. *upper ontology*), *Chemical*, *Gene Product*, *Health*, ... Поред прегледа онтологија, репозиторијум даје и статистички преглед њихових класа и својстава. Над-онтологије се фокусирају на опште концепте како би могле да их користе друге онтологије за представљање и дељење знања. Свеобухватни преглед биолошких онотологија и база података се може наћи у [26]. У Поглављу 4 биће приказан пример који показује како се скуп података може приказати помоћу технологија Семантичког Веба и како се над-онтологија може искористити за постизање веће глобалне видљивости скупа. Након што је приказано на који начин се подаци представљају и повезују у семантички базираном окружењу, следећи битан корак се односи на претраживање оваквих података. За претраживање RDF складишта података конструисан је посебан упитни језик, *SPARQL*, о којем ће бити нешто више речи у наредној секцији.

2.4 Simple Protocol and RDF Query Language (SPARQL)

У овој секцији су приказане основе и синтакса SPARQL упитног језика на чијим је основама развијена софтверска платформа. Из тог разлога ћемо SPARQL-у посветити нешто већу пажњу него претходним технологијама. На почетку уводимо неколико концепата чије се дефиниције природно ослањају на оне из Секција 2.2 и 2.3, након чега ће бити приказана генеричка структура SPARQL упита.

SPARQL Protocol и RDF упитни језик (SPARQL) [27] је протокол и упитни језик за приступ и претраживање RDF података. SPARQL упит заправо представља комбинацију уређених тројки (триплета) које се добијају помоћу њихове конјукције, дисјункције и/или скупа опционих шаблона који ће бити посебно описани у наставку. Као резултат рада *Data Access Working Group (DAWG)* групе постао је опште прихваћени стандард за претраживање RDF скупова података. У јануару 2008. године [15] постаје званични W3C стандард, да би у марту 2013. године била представљена његова побољшана верзија под називом SPARQL1.1 [16].

Као што је већ речено, извршавање SPARQL упита се дефинише као низ корака који почиње од SPARQL упита као стринга. Након тога следи његово пребацивање у апстрактну форму, а затим пребацивање апстрактне форме у SPARQL апстрактни упит који обухвата операторе из SPARQL алгебре. Овакав апстрактни упит се затим извршава над RDF скупом података. RDF скуп података представља информације у виду графа који садржи триплете са субјектом, предикатом и објектом. У пракси, RDF складишта података најчешће чувају податке од више RDF графова дозвољавајући апликацијама да креирају упите који обухватају информације од више графова.

[12] https://bioportal.bioontology.org/ontologies

Дакле, SPARQL упит се извршава на RDF скупу података који представља колекцију графова. RDF скуп података обухвата један граф без имена (*default*) и нула или више именованих графова где се сваки од графова идентификује користећи URI или IRI (*International Resource Identifier*). Оно што је битно поменути је да RDF скуп података не мора да има именовани граф, али да увек садржи default граф. SPARQL упит може да упари (енг. *match*) више различитих делова упита са различитим графовима помоћу кључне речи *GRAPH*. У самом упиту се default граф не помиње, док се именовани графови морају навести.

Дефиниција 2.6. *RDF скуī īодатака* се дефинише као скуп:

$$\{G, (< u_1 >, G_1), (< u_2 >, G_2), \ldots, (< u_n >, G_n)\},$$

где су G и G_i графови, а u_i је различити IRI, за $i = 1, \ldots, n$. G се назива основним (default) графом, а парови $(< u_i >, G_i)$ именованим графовима.

Граф над којим се врши претраживање основних графовских шаблона (енг. *basic graph pattern*) се назива активни граф (енг. *active graph*). Дакле, сви упити се извршавају над једним графом, default графом RDF скупа података који представља активни граф. Кључна реч GRAPH се у упиту користи да се један именовани граф у скупу учини активним за неки део упита. У наставку текста дајемо и формалне дефиниције које заокружују претходну причу о упаривању упита са подацима.

2.4.1 Опште дефиниције

Дефиниција 2.7. *Промењиве у уīиту* се дефинишу као елементи скупа V, где је V бесконачни и дисјунктни скуп у односу на скуп RDF_T.

За разлику од Дефиниције 2.2, сада се шаблон триплета који се користи у упиту дефинише на следећи начин:

Дефиниција 2.8. *Шаблон триīлета* (енг. triple pattern) је елемент скупа $(RDF_T \cup V) \times (I \cup V) \times (RDF_T \cup V)$, где је I скуп свих IRI-ја.

Може се приметити да званична дефиниција за шаблон триплета укључује могућност да литерал буде субјекат при чему је то, према RDF-у, претходно било немогуће. Ова дефиниција се сада може прилагодити формалној и написати да је шаблон триплета елемент скупа:

$$(I \cup B \cup V) \times (I \cup V) \times (RDF_T \cup V).$$

Дефиниција 2.9. *Основни шаблон īрафа* (енг. basic graph pattern) је скуп шаблона триплета. Празан шаблон графа је основни шаблон графа који је празан скуп.

Дефиниција 2.10. *Solution Mapping* је мапирање скупа променљивих у скуп RDF термина. Формално се претходна реченица може записати као функција $\mu : V \to RDF_T$.

Дефиниција 2.11. *Solution sequence* је листа решења.

2.4.2 Структура SPARQL упита

SPARQL синтакса доста подсећа на синтаксу коју има SQL упитни језик. На Слици 2.5 може се видети упрошћена верзија синтаксе SPARQL упита који се користе у овој дисертацији. За више детаља читаоца упућујемо на званичну W3C документацију [16], док се нешто краћи преглед делова упита може видети у [28], на чији се садржај наредни текст ослања. SPARQL упит се може поделити на пет делова, при чему ће кроз примере бити детаљно објашњени заглавље и клаузуле упита, док за форму, скупове и модификаторе упита читаоца упућујемо на званичну документацију.

```
SPARQLQuery := [Header*] Form [Dataset] WHERE { Pattern } Modifiers
-------------------------------------------------------------------------------
     Header := PREFIX value value | BASE value
       Form := SELECT [DISTINCT|REDUCED] (joker|var*) | ASK | CONSTRUCT var* | DESCRIBE
    Dataset := FROM value | FROM Named value
  Modifiers := LIMIT value | OFFSET value | ORDER By [ASK|DESC] var*
    Pattern := Pattern . Pattern | {Pattern} UNION {Pattern} | Pattern OPTIONAL {Pattern}
               | (value|var) (value|var) (value|var) | FILTER Constraint
-------------------------------------------------------------------------------
       var := ('?'|'$')value
     joker := '*'
     value ∈ String
```

Слика 2.5: SPARQL синтакса.

2.4.3 Заглавље упита

SPARQL омогућава писање скраћеница (префикса) како би упити били читљивији. Опционо се на почетку упита може додати листа префикса који су корисни у ситуацијама када су URI-ји дугачки. Кључној речи PREFIX се поред ознаке префикса додељује и URI. Назив префикса се састоји од ознаке префикса и локалног дела, који су одвојени знаком ":". Овде наводимо две примера варијација за оригинални URI:

```
# оригинални URI: <http://example.org/book/book1>
# keyword BASE:
BASE <http://example.org/book/>
<book1>

# keyword PREFIX:
PREFIX book: <http://example.org/book/>
book:book1
```

Листинг 2.3: PREFIX у SPARQL упиту.

2.4.4 Клаузуле упита (Pattern)

SPARQL се базира на графовским шаблонима који су упарени са онима који су специфицирани у WHERE делу упита. На овај начин се ради додатно филтрирање траже-

них резултата. Клаузуле упита се користе за креирање сложенијих графовских шаблона комбинујући оне мање на неколико различитих начина:

1. *Основни шаблони īрафа*: скуп шаблона триплета мора бити упарен.
2. *Груīисани шаблони īрафа*: сви графовски шаблони морају бити упарени.
3. *Оīциони шаблони īрафа*: додатни шаблони који могу да прошире решење.
4. *Алтернативни шаблон īрафа*: два или више шаблона могу да се комбинују.
5. *Шаблони на именованим īрафовима*: шаблони се упарују са именованим графовима.

Конјунктивност

Постоје два начина на која се шаблони могу комбиновати помоћу конјукције: **основни шаблони графа** који комбинује шаблоне триплета и **груписани шаблони графа** који комбинују све друге графовске шаблоне.

Основни шаблони īрафа представљају скупове шаблона триплета. Претраживање SPARQL графа помоћу шаблона се може дефинисати као комбиновање резултата добијених претраживањем основних графовских шаблона.

У SPARQL упиту, *īруīисани шаблон īрафа* је одвојен витичастим заградама: {}. На пример, следећи шаблон упита је груписани шаблон графа једног основног графовског шаблона.

```
SELECT ?name ?mbox
WHERE {
    ?x foaf:name ?name .
    ?x foaf:mbox ?mbox .
}
```

Листинг 2.4: Пример груписаног шаблона графа у SPARQL упиту.

Потенцијал

Основни шаблони графа дозвољавају апликацијама да креирају упите где целокупни шаблон упита мора да буде пронађен у графу да бисмо имали решење. За свако решење упита које садржи груписани шаблон графа, са најмање једним основним графовским шаблоном, свака променљива из решења је везана за скуп RDF_T. Ипак, комплетна структура не може бити подразумевана у свим RDF графовима. Корисно је имати могућност да се решењу дода информација која је доступна, а да не буде одбачена из решења ако се неки део шаблона не упари. Опционо упаривање омогућава овакву функционалност тј. ако се опциони део не упари, не прави се спајање, али се решење не елиминише.

Опциони делови графовског шаблона могу се специфицирати користећи кључну реч **OPTIONAL** која је примењена на графовски шаблон:

$$pattern\ \ OPTIONAL\ \{pattern\}$$

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox
WHERE  { ?x foaf:name ?name .
         OPTIONAL {?x foaf:mbox ?mbox}
}
# ------------------------------------------------
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ns: <http://example.org/ns\#>
SELECT ?title ?price
WHERE  { ?x dc:title ?title .
          OPTIONAL {
             ?x ns:price ?price .
             FILTER (?price < 30)
          }
        }
# ------------------------------------------------
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name ?mbox ?hpage
WHERE  { ?x foaf:name ?name .
         OPTIONAL { ?x foaf:mbox ?mbox } .
         OPTIONAL { ?x foaf:homepage ?hpage }
       }
```

Листинг 2.5: OPTIONAL у SPARQL упиту.

Алтернативност

SPARQL омогућава комбиновање графовских шаблона тако да се неколико алтернативних графова могу упарити. Ако се јави више од једног алтернативног упаривања, тада се приказују сва могућа решења што потенцијално може довести до редудантних података.

```
PREFIX dc10: <http://purl.org/dc/elements/1.0/>
PREFIX dc11: <http://purl.org/dc/elements/1.1/>
SELECT ?title ?author
WHERE  {
          { ?book dc10:title ?title .
            ?book dc10:creator ?author }
          UNION
          { ?book dc11:title ?title .
            ?book dc11:creator ?author }
        }
```

Листинг 2.6: UNION у SPARQL упиту.

2.4.5 Federated SPARQL упити

Кључна реч *SERVICE* даје инструкцију процесору упита да део SPARQL упита изврши на удаљеној SPARQL тачки повезивања (енг. *endpoint*). У литератури се термин „endpoint" („ендпоинт") углавном не преводи, тако да ће се исти термин употребљавати и у наставку текста. Следећи пример приказује начин на који се добијени резултати са удаљеног ендпоинта (http://people.example.org/sparql) придружују подацима из локалног RDF скупа података. Као резултат упит враћа имена програмера који су нам неопходни за реализацију пројекта.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name
FROM <http://example.org/myfoaf.rdf>
WHERE
{
   <http://example.org/myfoaf/I> foaf:knows ?person .
   SERVICE <http://people.example.org/sparql> {
     ?person foaf:name ?name .
   }
}
```

Листинг 2.7: SPARQL упит са удаљеним и локалним приступом подацима.

Наредни упит приказује коришћење два ендпоинта. Резултат овог упита су имена програмера, али и информације о њиховим интересовањима и другим програмерима које познају.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?person ?interest ?known
WHERE
{
  SERVICE <http://people.example.org/sparql> {
    ?person foaf:name ?name .
    OPTIONAL {
      ?person foaf:interest ?interest .
      SERVICE <http://people2.example.org/sparql> {
        ?person foaf:knows ?known . }}
  }
}
```

Листинг 2.8: Пример SPARQL упита са коришћењем два ендпоинта.

Извршавање SPARQL упита на удаљеном серверу се може завршити неуспешно из много разлога: сервис није доступан, сервис као резултат враћа грешку и слично. Под оваквим условима, упит који садржи *SERVICE* шаблон се неће извршити тј. неће вратити резултат. Како бисмо превазишли овакве ситуације користи се кључна реч *SILENT* која говори упиту да у случају грешке настави са даљим извршавањем и игнорише тај део упита.

SPARQL верзија 1.1 додатно укључује и коришћење клаузуле *VALUES*. На овај начин је могуће додатно филтрирати податке које су враћени са удаљеног ендпоинта користећи резултате добијене извршавањем других делова упита.

```
PREFIX dbr: <http://dbpedia.org/resource/>
PREFIX dbp: <http://dbpedia.org/property/>
SELECT ?country ?pop
WHERE {
    VALUES ?country {
        dbr:Scotland
        dbr:England
        dbr:Wales
        dbr:Northern_Ireland
        dbr:Ireland
    }
    SERVICE <http://dbpedia.org/sparql> {
        ?country dbp:populationCensus ?pop .
    }
}
```

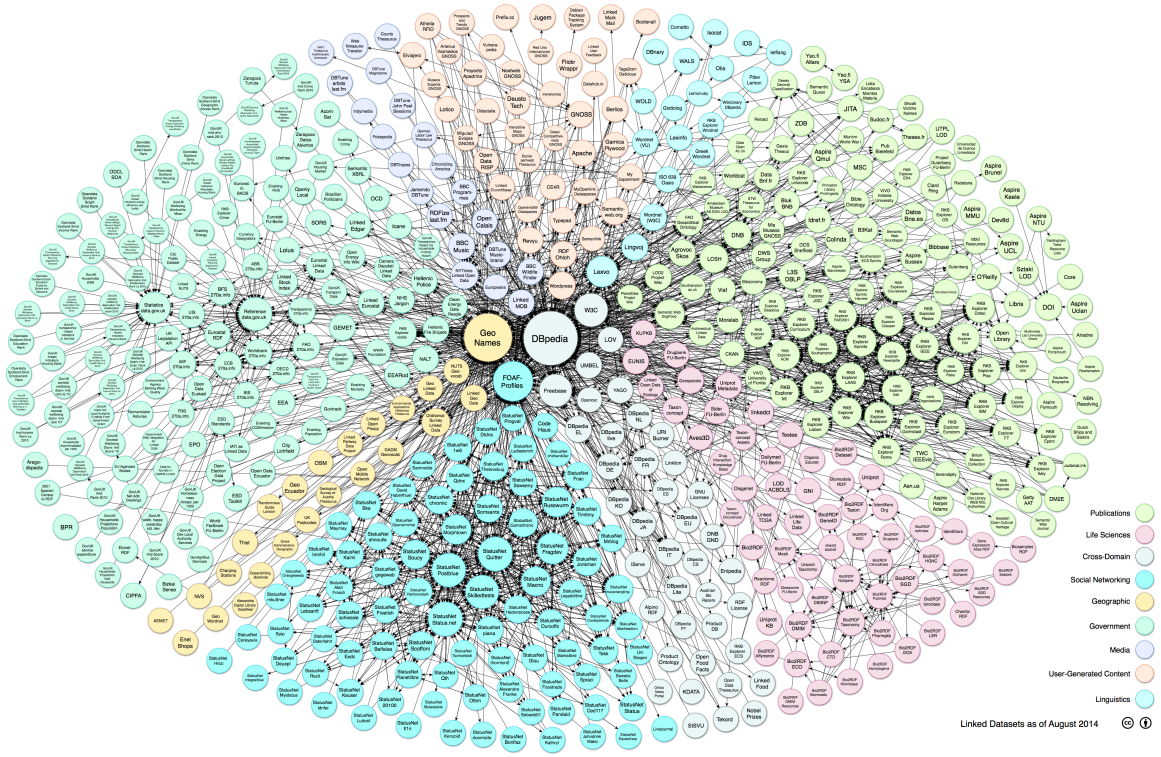
Листинг 2.9: Пример SPARQL упита.

Глава 3

Семантички базирани репозиторијуми

Принципи повезаних података су описани и објављени у *W3C Design Issues* документу [29] у којем је Tim Berners-Lee навео рационалне и практичне примере. Овај документ уопштава најбољу праксу за речнике, и залаже се за употребу широко прихваћених HTTP URI-ја за именовање и додатно укључивање екстерних URI-ја као простог механизма за повезивање података. Ово је свакако покренуло неке почетне напоре за промоцију и подршку алатима као што су претраживач Tabulator [29] који омогућава претрагу RDF скупова података који су публиковани на Вебу као *Linked Data*. У марту 2007 *W3C Semantic Web Education* и *Outreach (SWEO) Interest Group* су најавили нови пројеката под називом *"Interlinking Open Data"*[1] који је касније скраћен на *"Linking Open Data" (LOD)*. Овај пројекат је за циљ имао покретање парадигме Семантичког Веба креирањем, презентовањем и повезивањем RDF података који су добијени из ових "отворених" скупова, као и приказ бенефита које RDF и претходно описане технологије Семантичког Веба доносе широј *Open Data* заједници [30]. Пројекат LOD је постао атрактиван пре свега због превеликог напора који је уложила академска заједница у истраживачким лабораторијама додајући семантику постојећим подацима, као што је на пример DBpedia пројекат [31] који преводи структуиране податке са Википедијиног сајта у семантички допуњене и међусобно повезане податке. Пројекат се касније проширио и на битне корпоративне ентитете као што су BBC, Thompson Reuters, New York Times и различите владине агенције имајући за резултат хетерогени Веб који користи стандарде Семантичког Веба који су побољшани *Linked Data* принципима [30]. Linked Data имају за циљ да податке учине доступним на Вебу у формату који агенти могу лако да открију, приступе, комбинују и користе садржаје из различитих ресурса са вишим нивоом аутоматизације него што би другачије било могуће [32]. Предочени резултати свега тога је "Веб података" (*Web of Data*) тј. Веб структурираних података са богатим семантичким везама где агенти могу да врше претрагу на унифициран начин, преко више ресурса, користећи стандардизоване упитне језике и протоколе. У последњих десет година, стотине база података са огромним бројем чињеница су постале доступне пратећи стандарде Семантичког Веба (користећи RDF као модел података и RDFS и OWL за пружање експлицитне семантике), и *Linked Data* принципе.

[1] http://www.w3.org/blog/SWEO/page-2



Слика 3.1: LOD Cloud облак (преузето из Schmachtenberg et al [2]).

Интеграција биомедицинских података са минималним трошковима и уз минимално уложени напор се у великој мери заснива на примени нових технологија и пракси која на најбољи начин може да искористи структуру Веба за повезивање биомедицинских концепата и идентификатора у скуповима података. Повезивање података се базира на два кључна принципа која су уведена од стране *Linked Data* иницијативе. Први, према коме је сваки ентитет идентификован коришћењем *Universal Resource Identifier (URI)* тако да може лако да се референцира на Вебу, а према другом, линкови између ентитета и концепата из различитих скупова података треба да се установе користећи *Resource Description Framework (RDF)* [29].

Richard Cyganiak и Anja Jentzsch су приказали међусобно повезане скупове података у облику тзв. облак структуре (Слика 3.1) приказујући како су они међусобно повезани и користећи различите палете боја за обележавање различитих домена. Овакав приказ, у коме су скупови класификовани у осам различитих домена, је још познат и под називом *Linked Open Data (LOD) Cloud.* Ти домени се односе на владин сектор, публикације, природне науке, кориснички генерисане скупове, скупове који обухватају више различитих домена, медије, географију и друштвене мреже. Слика приказује податке који су објављени у *Linked Data* формату и који поштују претходно наведене принципе. Сви подаци су представљени у оквиру *Linking Open Data community* пројекта[2], као и они који долазе од стране других индивидуа и организација.

[2]https://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

3.1 Складиште RDF триплета

Складишта триплета (енг. *triplestores*) се користе за складиштење RDF података. Другим речима, triplestore је обично софтвер који има могућност ефикасног складиштења и индексирања RDF података за њихово лако и ефективно претраживање. Концепт triplestore за RDF податке је сличан концепту за релационе базе података, прецизније, релационом систему за управљање базама података (енг. *Relational Database Management System*).

Као што је већ речено, SPARQL је упитни језик за RDF у коме се преклапање графовских шаблона, унија, опционе клаузуле итд. могу користити за постављање упита над RDF знањем (видети Секцију 2.4). Већина складишта за триплете подржава извршавање SPARQL упитног језика за претраживање RDF података. Неколико добро познатих комерцијалних примера за складишта триплета су: Virtuoso[3], Sesame[4], Fuseki[5] и 4Store[6]. Као погодност за агенте, институције које објаве Linked Data скупове обично омогуће и коришћење SPARQL ендпоинта за постављање упита над њиховим локалним садржајем [33]. Стотине јавних ендпоинта је представљено у протеклих 10 година за базе знања различитих величина и тема [33, 34]. Коришћењем ових ендпоинта клијенти могу да добију одговоре за комплексне упите коришћењем само једног захтева упућеном серверу.

Табела 3.1 приказује поређење између четири triplestores (Virtuoso, Sesame, 4store и Fuseki) у односу на програмске језике који се могу користити са овим алатима (нпр. C, C++, Python PHP, Java, Javascript ActionScript, Tcl Perl Ruby, Obj-C, C#) и релевантним технологијама Семантичког Веба (RDF, RDFS, SPARQL, OWL, GRDDL, RDFa, RDB2RDF, R2RML, Direct Mapping). Последњих година Virtuoso се наметнуо као најбоље доступни triplestore који подржава већину програмских језика и технологије Семантичког Веба. За потребе ове докторске дисертације, за семантичку репрезентацију података Лабораторије коришћен је Fuseki, док је за потребе тестирања неких RDF скупова података коришћен Virtuoso triplestore. Том приликом су креирани графови за DrugBank и PubChem[7] скупове података који имају своје јавно доступне ендпоинте.

Табела 3.1: Поређење складишта за триплете.

складиште триплета	програмски језик	технологије СВ	категорије
Virtuoso	C, Python, Ruby, C#, . . .	RDF, RDFS, SPARQL, OWL, GRDDL, RDFa, RDB2RDF, R2RML	Triple Store, Reasoner RDF Generator, SPARQL Endpoint, OWL Reasoner, RDFS Reasoner, RDB2RDF
Sesame	Java, Python, PHP	RDF, RDFS, SPARQL	Triple Store, Programming Environment, Reasoner, Parser, RDFS Reasoner
Fuseki	Java	RDF, SPARQL	Triple Store
4store	Java	RDF, SPARQL	Triple Store

Често се не прави суштинска разлика између графовских и RDF база података, зато ће у тексту који следи бити дата додатна објашњења на ту тему. Треба напоменути да су складишта RDF триплета, као и графовске базе података (Neo4j, RedisGraph, Aran-

[3]http://virtuoso.openlinksw.com/

[4]http://rdf4j.org/sesame/2.8/docs/using+sesame.docbook?view

[5]http://jena.apache.org/documentation/serving_data/

[6]https://www.w3.org/2001/sw/wiki/4store

[7]http://147.91.203.161:8890/conductor/ **user/pass:** dba/dba

goDB и други), дизајнирани тако да складиште повезане податке. Пошто је RDF посебна врста повезаних података који се претражују искључиво коришћењем SPARQL-а, треба рећи да је RDF triplestore заправо врста графовске базе података, али и да међу њима постоје одређене разлике:

- Графовске базе података пружају подршку већем броју упитних језика. Оне могу да подржавају GraphLog, GOOD, SoSQL, BiQL, SNQL, и друге упитне језике, док RDF triplestores користе искључиво SPARQL.
- Графовске базе података могу да складиште различите врсте графова - неоријенисане, тежинске, хиперграфове итд., док RDF triplestores складиште само RDF триплете.
- Графовске базе података су више оријентисане на чворове графа, док су RDF triplestores више окренуте ка гранама графа.
- Графовске базе података су боље оптимизоване за тзв. графовске заобилазнице тј. тражење најкраћих путања у графу.
- RDF triplestores подржавају закључивање док графовске базе података не подржавају.

3.2 Семантички повезани подаци за природне науке

У овој подсекцији биће нешто више речи о самим репозиторијумима из домена природних наука (life science) који су фокусу ове докторске дисертације. Овде је дат детаљнији опис шест значајних биомедицинских скупова који чувају податке о различитим темама које су везане за природне науке, а могу бити од интереса многим истраживачким центрима:

1. ChEMBL
2. Linked Open Drug Data (LODD)
3. Bio2RDF
4. LinkedLifeData
5. Chem2Bio2RDF
6. Open PHACTS

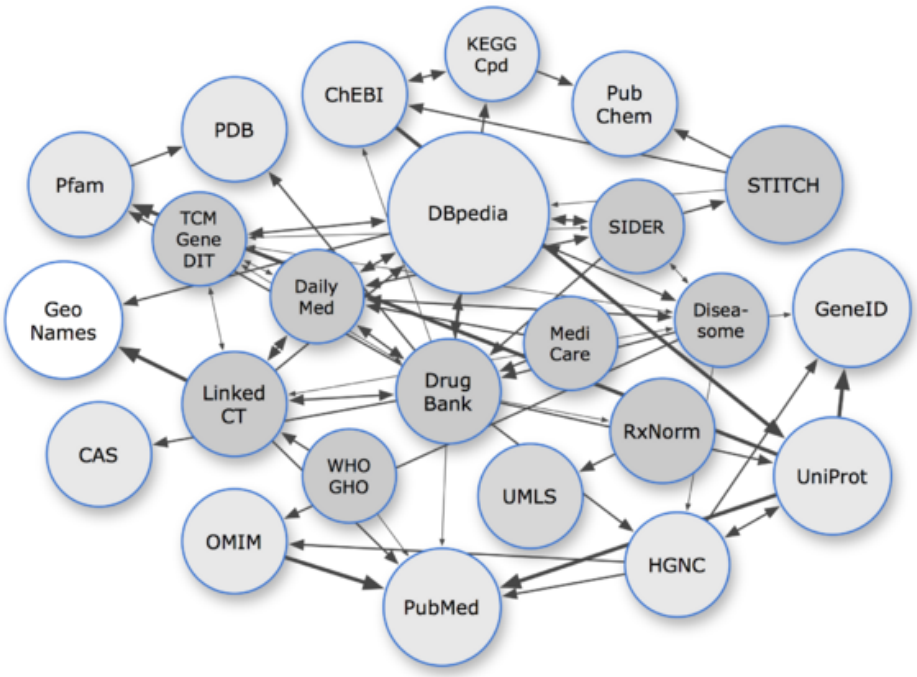
3.2.1 ChEMBL

ChEMBL је ручно припремљена хемијска база података биоактивних молекула са особинама сличним лековима. Ова база података обједињује хемијске, биоактивне и геномске податке како би се унапредило превођење геномских информација у нове

ефикасне лекове. ChEMBL је развијена у оквиру Европског Биоинформатичког Института (ЕБИ) у Великој Британији који је део европске лабораторије за природне науке (EMBL). Октобра 2019. база података садржи: 12 000 таргета, 1.9 милиона супстанци, 11 000 лекова, 4 700 механизама, 1.1 милион биолошких тестова, 655 ткива итд.

3.2.2 Linked Open Drug Data (LODD)

LODD[8] је репозиторијум повезаних скупова података који је креиран да унапреди процес откривања лекова (*Drug Discovery*). Он обухвата неколико скупова као што су Drugbank (766 920 триплета; 4 800 лекова, 2 500 протеинских секвенци), LinkedCT (25 милиона триплета, 106,000 клиничких експеримената), DailyMed (1 604 893 триплета, 36,000 производа), DBpedia (218 милиона триплета; 2 300 лекова, 2 200 протеина), Diseasome (91 182 триплета; 2 600 гена), SIDER (192 515 триплета; 63 000 извештаја о ефектима, 1 737 гена), STITCH (7 500 000 супстанци; 500 000 протеина; 370 организама), Medicare (информације нису доступне), RxNorm (више од 7.7 милиона триплета; 165 806 јединствених лекова), итд. Скупови података које обухвата LODD су приказани на Слици 3.2, а додатни детаљи кроз Табеле 3.2 и 3.3.



Слика 3.2: Неки LODD скупови података (тамно сива), повезани биомедицински скупови (светло сива) и општи скупови (бела), као и њихове међусобне везе.

Оно што је постала свакодневна појава је константна интеграција сличних репозиторијума. Због уштеде ресурса и синхронизованог рада више тимова, две или више иницијатива се спајају у једну унапређујући структуру и семантику података. На тај начин се до одређених података може доћи на једноставнији начин, а истраживачи

[8] http://www.w3.org/wiki/HCLSIG/LODD

врше вишеструку проверу исправности података. Последња битна промена која се догодила је успостављање сарадње LODD иницијативе са Bio2RDF[9] пројектом који је описан у наставку.

Табела 3.2: LODD скупови података.

Скуп	**Тематика**	**Опис**
DrugBank	Лекови	Пружа информације о лековима са детаљним описом таргета на које лек делује
LinkedCT	Клиничка испитивања	Повезани скупови података о урађеним испитивањима са ClinicalTrials.gov
DailyMed	Лекови	Сви FDA-одобрени SPLs и NDF-RT
DBpedia	Лекови/ Болести/ Протеини	Ентитети преузети са Википедије
Diseasome	Болести/ Гени	поремећаји и гени болести повезани за асоцијацијама поремећај-ген
DIKB	Лекови/Интерације лекова (DDIs)	Лекови и DDIs тврђења и докази за механизме лекова
RDF-TCM	Гени/Болести/Медицина	Кинеска медицина, гени, асоцијација и мапирање болести са Extrez Gene IDs
RxNorm	Лекови	Повезивање рецепата лекова, састојака и NDC кроз RXCUI
SIDER	Болести/Контраиндикације	Лекови на тржишту и њихови ефекти
STITCH	Хемијске структуре/Протеини	Хемијске структуре, протеини, и њихове интеракције
Medicare	Формулари здравствене заштитите	Лекари, медицински специјалисти, сервиси
ChEMBL	Огледи (протеини, организми)	Испитивање лекова са информацијама о њиховим активностима у односу на таргете
GHO	Инфекционе болести/Демографија	Инфекције у земљама, на регионалном и глобалном нивоу
UPNR	Лекови/Процедуре/Дијагнозе	800 клиничких забелешки преузетих са Универзитета у Питсбургу

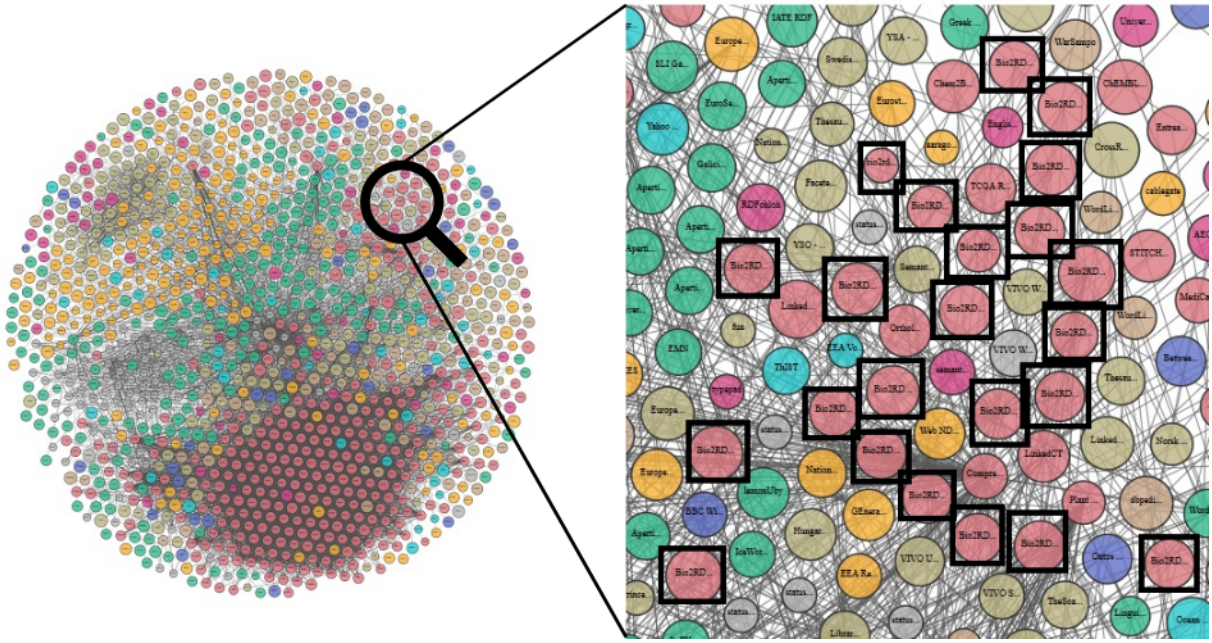
Табела 3.3: LODD ендпоинти.

Скуп	**Година**	**Величина/Покривеност**	**SPARQL ендпоинт**
DrugBank	2010	766 920 триплета; 4 800 лекова	http://www4.wiwiss.fu-berlin.de/drugbank/sparql
LinkedCT	X	25M триплета, 106 000 испитивања	http://data.linkedct.org/sparql
DailyMed	2010	1 604 893 триплета, 36K производа	http://purl.org/net/nlprepository/linkedSPLs
DBpedia	2009	218M триплета; 2 300 лекова, 2 200 протеина	http://dbpedia.org/sparql 2.49 million
Diseasome	2010	91 182 триплета; 2 600 гена	http://www4.wiwiss.fu-berlin.de/diseasome/sparql
DIKB	2011	више од 41K триплета	http://dbmi-icode-01.dbmi.pitt.edu:2020/
RDF-TCM	2009	117 643 триплета	http://www.open-biomed.org.uk/sparql/endpoint/tcm
RxNorm	2011	више од 7.7M триплета	http://link.informatics.stonybrook.edu/sparql/
SIDER	2010	192 515 триплета; 63K ефеката, 1 737 гена	http://www4.wiwiss.fu-berlin.de/sider/sparql
STITCH	2010	7.5M хемијских структура; 0.5M протеина	http://www4.wiwiss.fu-berlin.de/stitch/sparql
Medicare	2010	X	http://www4.wiwiss.fu-berlin.de/medicare/sparql
ChEMBL	2010	130M триплета	http://rdf.farmbio.uu.se/chembl/sparql
GHO	2011	3M триплета	http://gho.aksw.org
UPNR	X	38 664 триплета	http://dbmi-icode-01.dbmi.pitt.edu:8080/sparql

[9]http://bio2rdf.org

3.2.3 Bio2RDF

Bio2RDF (Release 3) представља пројекат који садржи више повезаних биолошких база података укључујући базе података за путање (енг. *pathways databases*), као што су KEGG, DrugBank и PDB, и неколико база података Националног центра за биотехнолошке информације, а који је део америчке националне библиотеке за медицину [35] (Слика 3.3). Bio2RDF је јавно доступан пројекат који користи технологије Семантичког Веба да креира и омогући приступ највећој мрежи повезаних података за природне науке (*Life Sciences*).



Слика 3.3: Bio2RDF у LOD облаку.

Bio2RDF дефинише скуп простих конвенција за креирање RDF репозиторијума из скупа различитих и хетерогених извора података који су добијени од више провајдера. У јуну 2014-те, Bio2RDF Release 3 садржи 11 милијарди триплета који су дислоцирани у 35 скупова: Chembl, linkedSPL, Pathway Commons, Reactome, WikiPathways, OrphaNet, clinicaltrials.gov, PubMed, SIDER, dbSNP, GenAge, WormBase, GenDR,. . . (видети Табеле 3.4 и 3.5). Све класе, својства објеката и анотације тих својстава су мапирани са *Semanticscience Integrated Ontology (SIO)*[10], над-онтологијом која пружа општи речник за Bio2RDF. Приступ подацима је омогућен преко SPARQL 1.1 ендпоинта[11] користећи Virtuoso 7.2.0.

[10]https://bioportal.bioontology.org/ontologies/SIO
[11]http://bio2rdf.org/sparql

Табела 3.4: Опис Bio2RDF скупова података.

Скуп	**Тематика**	**Опис**
Affymetrix	Микронизови	Probesets (групе ДНК секвенци) коришћене у Affymetrix микро-низовима
BioModels	Биолошки и математички модели	Складиштење, претрага и приказивање објављених математичких модела који су од интереса за биологију
BioPortal	Биолошке и биомедицинске онтологије	Репозиторијум биомедицинских онтологија
ChEMBL	Биолошке супстанце	Биоактивне супстанце, квантитативна својства и биоактивности
ClinicalTrials	Клиничка испитивања	Јавне и приватно подржане клиничке студије у којима су учесници људи
CTD	Хемијске ген/протеин интеракције	Хемијске ген/протеин интеракције, хемијска структура-ген-болест
dbSNP	Супституције нуклеотида	Супституција нуклеотида, брисање и уношење полиморфизама
DrugBank	Лекови	Детаљни подаци о лековима са свеобухватним таргетима лекова
GenAge	Гени	Људски и органски гени који су повезани са дуговечношћу и старењем
GenDR	Гени	Гени повезани са ограничењем исхране (дијетом)
GOA	Gene Ontology анотација	Gene Ontology (GO) анотација протеина који се налазе у скуповима UniProtKB и IPI
HGNC	Људски гени	Скуп јединствених и смислених имена за сваки људски ген
HomoloGene	Означени гени	Аутоматско детектовање хомологије међу означеним генима
InterPro	Протеини и геноми	Откривање „потписа" протеина
iProClass	Протеини, путање, гени	UniProtKB и UniParc протеини са линковима до биолошких база података
iRefIndex	Протеини, путање, гени	Интеракције протеина у базама BIND/BioGRID/DIP/HPRD/MPPI/OPHID
KEGG	Гени	16 база података које садрже биолошке и хемијске информације
MGI	Гени	Гени, номенклатура, мапирање, хомологије, фенотипови, алелови
NCBI Gene	Гени	Номенклатура, RefSeqs, мапирање, путање, варијације, фенотипови, локус
NDC	Идентификатори лекова	Лекови намењени људској употреби у САД
OMIM	Mendelian поремећај, гени	Људски гени и генетски фенотипови
Orphanet	Ретке болести и непрофитабилни лекови	Ретке болести и непрофитабилни лекови (Orphan drug). Дијагноза, брига и третмани за ретке болести
PC	Путање	Информације о биолошким путањама које су сакупљене из јавно доступних база података
LinkedSPL	Лекови	Повезани подаци скупа DailyMed
LSR	LS терминологија	Скупови података и терминологија која се користи у природним наукама
MeSH	Појмови и терминологије	Означавање дескриптора у хијерархијској структури за претрагу специфичности
PharmGKB	Генотип/фенотип	Подаци о генотипу/фенотипу, варијације гена, веза између ген-лек-болест
PubMed	Цитати	Цитати из MEDLINE и LS часописа за биомедицинске чланке након 1950-те године
Reactome	Путање	Основне путање и реакције у биологији човека
SABIO-RK	Биохемијске реакције	Биохемијске реакције, параметри и услови
SGD	Биохемијске реакцијes	Молекуларна биологија и генетика пивског квасца (Saccharomyces cerevisiae)
SIDER	Лекови	Реакција на лекове, учесталост контраиндикација и њихова класификација
Taxonomy	Таксономија	Организми у генетским базама података са једном нуклеотидном или протеинском секвенцом
WikiPathways	Мапе за путање	Отворене и јавно доступне колекције мапа за биолошке путање
WormBase	Геном	Геномика Caenorhabditis elegans и других сличних ваљкастих црва

Табела 3.5: Bio2RDF ендпоинти.

Скуп	**Година**	Величина и покривеност	**Ендпоинт**
Affymetrix	01/08/2014	86 942 371 триплета, 6 679 943 ентитета	http://cu.affymetrix.bio2rdf.org/sparql
BioModels	05/06/2014	2 380 009 триплета, 188 380 ентитета	http://cu.biomodels.bio2rdf.org/sparql
BioPortal	20/07/2014	19 920 395 триплета, 2 199 594 ентитета	http://cu.bioportal.bio2rdf.org/sparql
ChEMBL	X	409942525 триплета, 50 061 452 ентитета	http://cu.chembl.bio2rdf.org/sparql
ClinicalTrials	25/09/2014	98 835 804 триплета, 7 337 123 ентитета	http://cu.clinicaltrials.bio2rdf.org/sparql
CTD	09/06/2014	326 720 894 триплета, 19 768 641 ентитета	http://cu.ctd.bio2rdf.org/sparql cross-species
dbSNP	15/07/2014	8 801 487 триплета, 530 538 ентитета	http://cu.dbsnp.bio2rdf.org/sparql
DrugBank	25/07/2014	3 672 531 триплета, 316 950 ентитета	http://cu.drugbank.bio2rdf.org/sparql
GenAge	03/06/2014	73 048 триплета, 6 995 ентитета	http://cu.genage.bio2rdf.org/sparql
GenDR	03/06/2014	11 663 триплета, 1 129 ентитета	http://cu.gendr.bio2rdf.org/sparql
GOA	05/06/2014	97 520 151 триплета, 5 950 074 ентитета	http://cu.goa.bio2rdf.org/sparql
HGNC	04/07/2014	3 628 205 триплета, 372 136 ентитета	http://cu.hgnc.bio2rdf.org/sparql
HomoloGene	04/07/2014	7 189 769 триплета, 869 985 ентитета	http://cu.homologene.bio2rdf.org/sparql
InterPro	02/06/2014	2 323 345 триплета, 176 579 ентитета	http://cu.interpro.bio2rdf.org/sparql
iProClass	09/06/2014	3 306 107 223 триплета, 364 255 265 ентитета	http://cu.iproclass.bio2rdf.org/sparql
iRefIndex	22/06/2014	48 781 511 триплета, 3 110 993 ентитета	http://cu.irefindex.bio2rdf.org/sparql
KEGG	13/08/2014	50 197 150 триплета, 6 533 307 ентитета	http://cu.kegg.bio2rdf.org/sparql
MGI	05/06/2014	8 206 813 триплета, 924 257 ентитета	http://cu.mgi.bio2rdf.org/sparql
NCBI Gene	20/09/2014	2 010 283 833 триплета, 189 594 629 ентитета	http://cu.ncbigene.bio2rdf.org/sparql
NDC	02/08/2014	6 199 488 триплета, 488 146 ентитета	http://cu.ndc.bio2rdf.org/sparql
OMIM	19/09/2014	8 750 774 триплета, 1 013 389 ентитета	http://cu.omim.bio2rdf.org/sparql
Orphanet	02/06/2014	377 947 триплета, 28 871 ентитета	http://cu.orphanet.bio2rdf.org/sparql
PC	X	5 700 724 триплета, 1 024 572 ентитета	http://cu.pathwaycommons.bio2rdf.org/sparql
LinkedSPL	X	2 174 579 триплета, 59 776 ентитета	http://cu.linkedspl.bio2rdf.org/sparql
LSR	16/07/2014	55 914 триплета, 5 032 ентитета	http://cu.lsr.bio2rdf.org/sparql
MeSH	27/05/2014	7 323 864 триплета, 305 401 ентитета	http://cu.mesh.bio2rdf.org/sparql
PharmGKB	27/06/2014	278 049 209 триплета, 25 325 504 ентитета	http://cu.pharmgkb.bio2rdf.org/sparql
PubMed	27/06/2014	5 005 343 905 триплета, 412 593 720 ентитета	http://cu.pharmgkb.bio2rdf.org/sparql
Reactome	X	12 487 446 триплета, 2 461 010 ентитета	http://cu.reactome.bio2rdf.org/sparql
SABIO-RK	05/06/2014	2 716 421 триплета, 448 248 ентитета	http://cu.sabiork.bio2rdf.org/sparql
SGD	07/08/2014	12 494 945 триплета, 957 558 ентитета	http://cu.sgd.bio2rdf.org/sparql
SIDER	22/07/2014	17 627 864 триплета, 1 222 429 ентитета	http://cu.sider.bio2rdf.org/sparql
Taxonomy	27/05/2014	21 310 356 триплета, 1 147 211 ентитета	http://cu.taxonomy.bio2rdf.org/sparql
WikiPathways	X	514 397 триплета, 71 879 ентитета	http://cu.wikipathways.bio2rdf.org/sparql
WormBase	04/06/2014	22 682 002 триплета, 1 840 311 ентитета	http://cu.wormbase.bio2rdf.org/sparql

3.2.4 LinkedLifeData

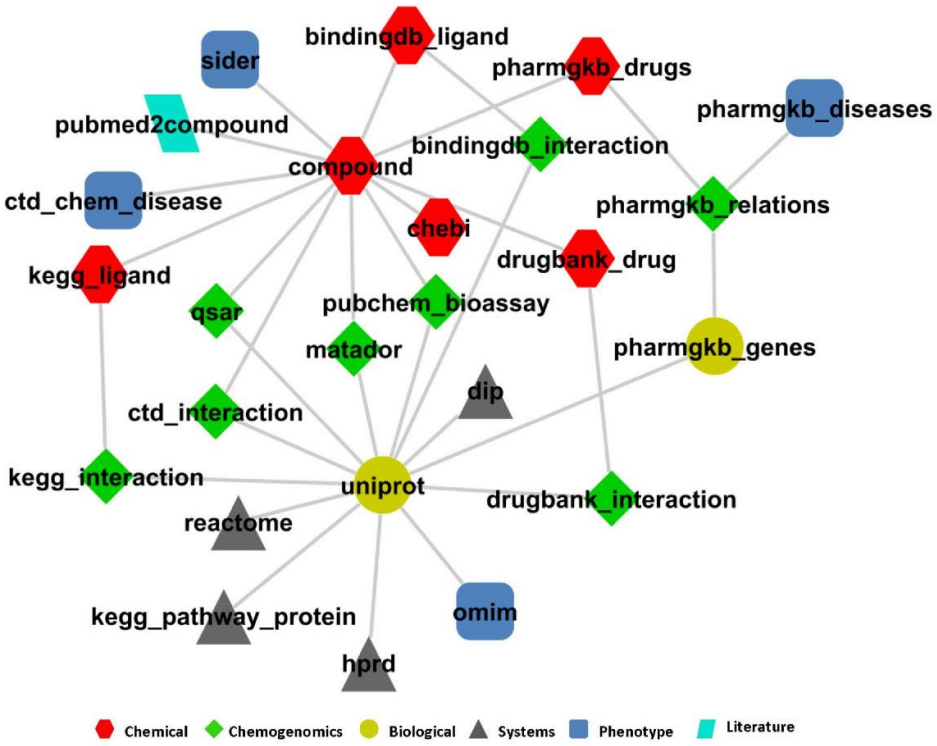
LinkedLifeData (LLD)[12] је семантички базирани репозиторијум који интегрише податке преузете из више од 25 различитих ресурса (DrugBank, UniProt, Reactome, SIDER, PubMed, итд).

LLD 2.0 у овом тренутку садржи 1 553 620 636 ентитета, и више од 10 милијарди повезаних триплета којима се приступа кроз један ендпоинт.

3.2.5 Chem2Bio2RDF

Иницијативе као што су Bio2RDF и LODD се баве повезивањем биолошких података и података о лековима користећи RDF, међутим могућности повезивања података из хемогеномике и хемијске биологије, који припадају тотално различитим доменима хемије и биологије, су и даље остале ограничене.

Chem2Bio2RDF пројекат је настао као одговор на поменута ограничења стварајући нови репозиторијум који агрегира податке са више других репозиторијума укључујући PubChem Bioassay [36], DrugBank [37], KEGG Ligand [38], CTD [39], BindingDB [40], PharmGKB [41], MATADOR [42], и многе QSAR (*Quantitative structure–activity relationship*) скупове који су доступни на Вебу [43], али у исто време умрежавајући Bio2RDF и LODD (Слика 3.4).



Слика 3.4: Скупови података интегрисани у Chem2Bio2RDF иницијативи.

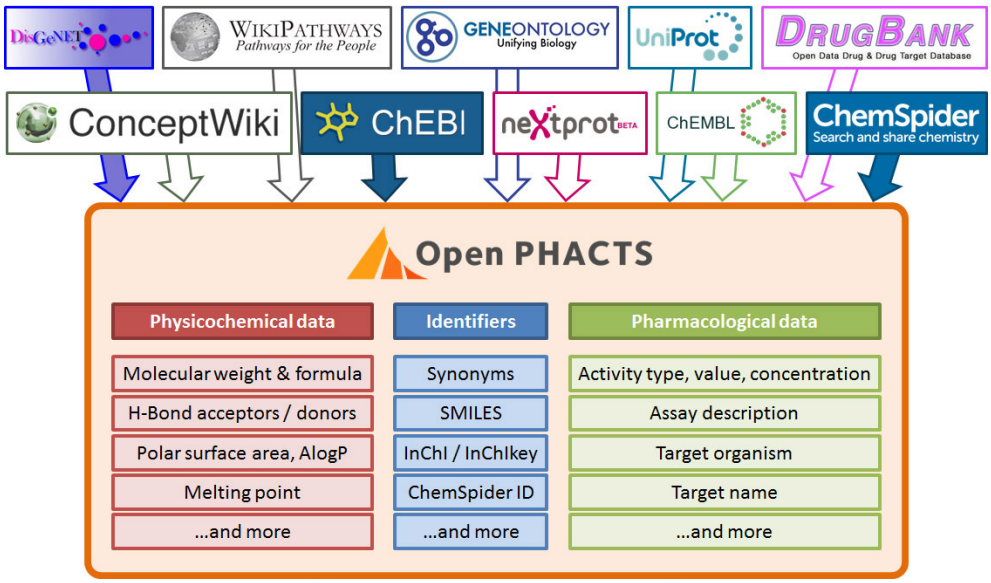
[12]http://linkedlifedata.com

3.2.6 Open PHACTS

Open PHACTS (Open Pharmacological Concept Triple Store) [44] је европска иницијатива основана од стране *Innovative Medicines Initiative* која почива на јавно-приватном партнерству између академских институција, издавача, предузећа, фамацеутских компанија и других организација које су се окупиле око заједничког циља који треба да омогући јефтиније и брже откривање лекова. За разлику од претходних иницијатива Open PHACTS је развијен за потребе фармацеутске индустрије. Са те тачке гледишта, иницијатива је развијена као одговор на све већи проблем претраге постојећих база података, што постаје све више уско грло како количина доступних података расте. Open PHACTS интегрише податке из скупова као што су ChEMBL, neXTProt и DrugBank да би се креирало више од 3 милијарде семантички повезаних триплета из области природних наука (Слика 3.5).

Open PHACTS иницијатива је тренутно подржана од стране 27 партнера:

- *Академске институције:* Maastricht University, University of Santiago de Compostela, University of Vienna, University of Manchester, University of Bonn, Swiss Institute of Bioinformatics, European Bioinformatics Institute, Vrije Universiteit of Amsterdam, Technical University of Denmark, University of Hamburg
- *Фармацеутске компаније:* Pfizer, Merck KGaA, Eli Lilly and Company, Novartis, GlaxoSmithKline, AstraZeneca
- *Друге компаније:* ChemSpider, Biovia, Eagle Genomics, Entagen
- *Издавачи:* Royal Society of Chemistry, Thomson Reuters



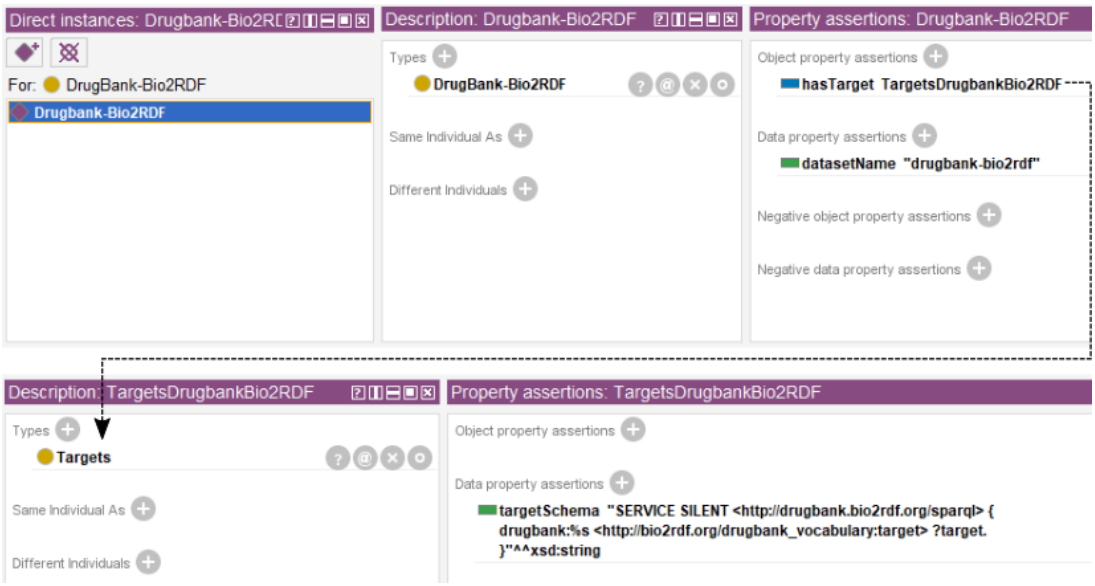
Слика 3.5: Скупови података интегрисани у Open PHACTS иницијативи.

Да би се уклониле баријере у откривању лекова у индустрији и академији, Open PHACTS конзорцијум је развио јавно доступну платформу под називом *Open PHACTS*

Discovery Platform [45]. Ова платформа омогућава интеграцију фармацеутских података из различитих извора пружајући алате и сервисе за претраживање овако интегрисаних података као подршку истраживању. Постоје наговештаји да ће се у наредном периоду радити на интеграцији Chem2Bio2RDF и Open PHACTS иницијатива.

3.3 Репозиторијуми, скупови и подупити

Како би се олакшао рад истраживачима, за сваки репозиторијум (скупу података) су сакупљене одређене информације (видети претходне табеле) и на основу њих је конструисана једноставна онтологија, *RepoIntegration.owl*, која садржи основне податке о свакој иницијативи (репозиторијуму). Посебна пажња у онтологији је посвећена скуповима који су интегрисани у оквиру репозиторијума, као и подацима које они складиште, како бисмо у сваком тренутку знали где се потенцијални одговори на питање могу пронаћи. За сваки од скупова чувају се три њима одговарајућа подупита који се односе на ћелијске линије, таргете и IC_{50} вредност. Ови упити представљају неку врсту образаца (енг. *patterns*), јер су одређени делови шаблона триплета непознате које добијају вредност током креирања и извршавања упита, док су предикати унапред дефинисани како бисмо добили прави одговор на питање корисника. Уколико одређени скупови не чувају ове податке, та ће поља остати непопуњена. На Слици 3.6 је дат пример везе скупа DrugBank (Bio2RDF) са својим подупитом за таргете. Према потреби се на једноставан начин могу додати други одговарајући подупити у случају да се интересовања прошире на нову тематику.



Слика 3.6: Скуп DrugBank и његов подупит за таргете.

У Подсекцији 5.2.1 је дат преглед постојећих аутоматских генератора за упите који нису развијени као алат који има употребну вредност са аспекта претраживања репозиторијума, већ за анализу перформанси извршавања упита. Овакви упити у већини случајева не одговарају потребама корисника, а доста често генератори креирају упите који нису валидни. Чак и када су генерисани упити валидни, готово их је немогуће

увезати све у један упит који се извршава над више репозиторијума. У овој докторској дисертацији се користе подупити (шаблони) који су делом сакупљени са портала различитих иницијатива, а делом ручно креирани и модификовани за наше потребе, пошто су валидини резултати један од главних циљева ове софтверске платформе. Неки примери коришћених шаблона су дати у Табели 3.6. Наглашени термини означавају непознате субјекте и објекте који се одређују у току покретања Платформе (*on-the-fly*) и мењају са одговарајућим инстанцама (URIs), док предикати остају фиксирани (*bounded*). Недостатак интегрисаних речника чини креирање упита још компликованијим, нарочито у ситуацији када URI-ји у репозиторијумима нису исти. Касније ће бити објашњено на који начин се подупити повезују у целину користећи тзв. "same as" релације у оквиру репозиторијума, без коришћења заједничке онтологије која их повезује. На овај начин се број међурезултата смањује, а самим тим су и коначни резултати унапред филтрирани.

Табела 3.6: Подупити скупова података унутар репозиторијума.

Скуп/репозиторијум	Подуити за претраживање таргета лекова
DrugBank/Bio2RDF	**?drugbank_id** <http://bio2rdf.org/drugbank_vocabulary:target> **?target**
DrugBank/ Chem2Bio2RDF	**?isValueOf** <http://chem2bio2rdf.org/drugbank/resource/DBID> **?drugbank_id** . **?isValueOf** <http://chem2bio2rdf.org/drugbank/resource/CID_GENE> **?target:**
Chembl/EMBL-EBI	**?activity** <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://rdf.ebi.ac.uk/terms/chembl#Activity> . **?activity** <http://rdf.ebi.ac.uk/terms/chembl#hasMolecule> **?chembl_id** . **?activity** <http://rdf.ebi.ac.uk/terms/chembl#hasAssay> **?assay** . **?assay** <http://rdf.ebi.ac.uk/terms/chembl#hasTarget> **?target** .

Онтологија је скалабилна тако да се лако могу додати нови скупови података оних институција које желе пре свега да своје податке истраживања учине јавно доступним, али и да буду интегрисани са постојећим иницијативама, отварајући тако могућност приступа резултатима других сличних институција.

Глава 4

Интеграција нових RDF скупова

Брзина којом се биомедицински подаци објављују и мењају, као и различитост формата у којима се они представљају, доводи до тога да се свим подацима јако тешко може приступити у одређеној интегрисаној форми. Ручна претрага велике количине података је доста напорна и компликована, а приступ сваком скупу посебно захтева доста времена и стрпљења. Ако говоримо о семантички базираним скуповима података, ту је ситуација и нешто неповољнија. Истраживачи морају најпре да истраже сваки скуп података појединачно, тј. да се упознају са његовом структуром, кључним ентитетима, предикатима и које ентитете они повезују. Веома често скупови података садрже стотине хиљада, а у већини случајева и неколико десетина милона RDF триплета. Након упознавања са структуром, неопходно је написати одговарајуће SPARQL упите, што представља додатну баласт за истраживаче. Затим, те упите треба извршити, уредити добијене резултате у смислено и корисно знање које треба да служи као подршка у доношењу одлука у току рада. У реалним аплкацијама резултати морају да се филтрирају и добро организују у кратком временском року, што је готово немогуће имајући у виду све дате околности и потребне кораке. Зато је интеграција скупова података од виталног значаја за истраживачку заједницу, јер може значајно да убрза и унапреди лабораторијске активности.

Као битан преуслов за интеграцију података на глобалном нивоу је могућност њиховог трансформисања из једног формата у други, у већини случајева из релационих база података (РБП) у скуп RDF триплета. Већина података је данас смештена у релационим базама података и зато постоји неколико студија које су се бавиле развојем техника и алата који треба да олакшају мапирање између РБП и локалне онтологије. Детаљни опис два основна приступа која се користе за овај задатак је дат у [46]. Један приступ је везан за креирање потпуно нове онтологије на основу структуре базе података. Овај приступ је примењен у нашем случају, где су сви подаци из тренутне базе података мигрирани у складиште RDF података у чијој се основи налази PIBAS (*Preclinical Investigation of BioActive Substances*) онтологија која осликава начин рада Лабораторије. У наредном тексту биће представљене класе новоразвијене онтологије и објашњено на које податке из базе података се односе. Други приступ се односи на мапирање базе података у онтологију помоћу мапе која служи као медијатор између њих. Као пример овог приступа наводимо OWL2RDB језик за мапирање који креира међуслој између РБП и OWL онтологије [47].

4.1 PIBAS онтологија

Фундаментални циљ науке је формални опис експеримената за ефикасну анализу, анотацију и дељење резултата, што је и сада случај. У Секцији 1.1 је описан начин извођења експеримената у Лабораторији, затим које се супстанце посматрају, која се процедура користи том приликом, као и на који начин се мери цитотоксичност супстанци. Први корак ка потенцијалној интеграцији података са глобаним иницијативама, како би резултати постали видљиви свима, је њихово презентовање помоћу технологија Семантичког Веба. Ове технологије имају највећи потенцијал за моделовање структуре и резултата експеримената, јер је помоћу њих могуће приказати сложене везе између објеката, а које се тешко представљају у релационим базама података. На пример, користећи предности OWL синтаксе класе *pibas:Rat* и *pibas:Fish* су представљене као дисјунктне. Сама конструкција онтологије која одговара пословању Лабораторије је била доста комплексан и итеративан процес имајући у виду сложеност саме структуре. Термини који фигуришу су хијерархијски повезани приказујући њихове сложене односе и везе.

Пре самог моделовања извршен је преглед свих одговарајућих над-онтологија (upper ontology) које садрже опште термине (као што су објекти, својства и везе) који су заједнички за више домена. Основна функција ових онтологија је подршка семантичкој интероперабилности између великог броја доменски специфичних онтологија. На овај начин је направљена заједничка полазна тачка за формулацију неких дефиниција. Над-онтологија која најбоље осликава рад у експерименталним наукама је EXPO онтологија [48] која дефинише више од 200 концепата за креирање семантички значајних ентитета за научне екперименте. EXPO онтологија[1] је развијена са циљем да формализује знање о дизајну научних експеримената, методологији и репрезентацији резултата. Креирање такве заједничке онтологије је било неопходно због тога што научници прате исте принципе приликом извођења експеримената. Конкретно, на основу класа *expo:ExperimentalResult* и *expo:ExperimentalGoal* креирани су одговарајући пандани, односно својства типа података: *pibas:result* и *pibas:theAimOfExperiment*. Такође, за концепте "Experiment" и "ExperimentalMethod" који су дефинисани у структури експеримената уочене су одговарајуће класе *expo:ScientificExperiment* и *expo:ExperimentalMethod*. Међутим, свака лабораторија има своје специфичности којима треба да буде посвећена посебна пажња. Одређени концепти EXPO онтологије су усвојени и модификовани као одговарајући, док постоје и концепти који у њој нису садржани и који су додати у хијерархију. Управо овакви кораци чине скупове података различитим и тешким за коришћење и даљу интеграцију.

Захваљујући истраживачима који свакодневно врше експерименте издвојени су битни термини, притом дефинишући њихове међусобне односе (класа-подкласа, класа-својство или класа-инстанца) који су неопходни за изградњу таксономије. Процес валидације је обављен од стране експерата, док је процес закључивања, у циљу провере конзистентности онтологије, обављен применом FACT алата у Protégé едитору. У наставку је детаљно представљена таксономија PIBAS онтологије.

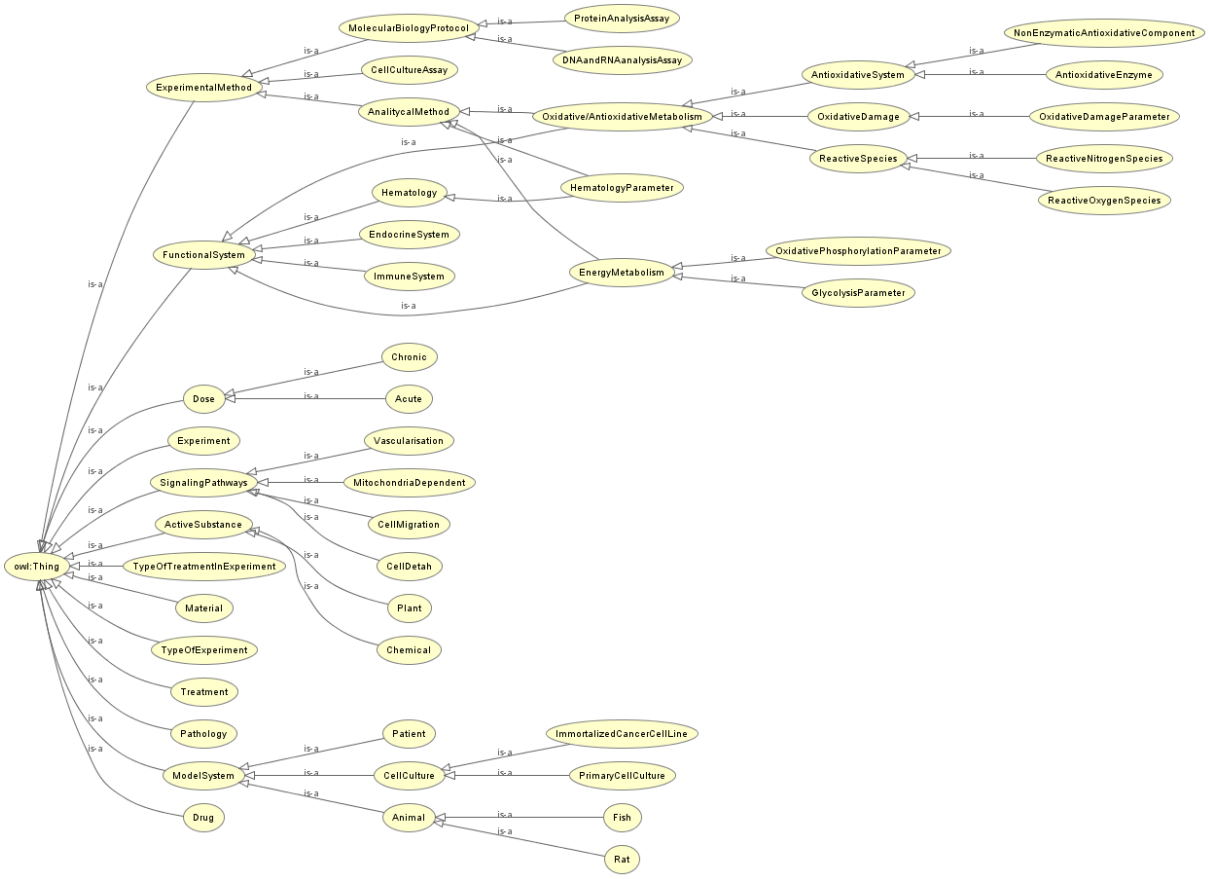
[1] http://expo.sourceforge.net/

4.1.1 Концепти PIBAS онтологије

У PIBAS онтологији на врху хијерархије се налази класа *owl:Thing* (скуп свих инстанци), према OWL стандарду, док су све остале класе њене подкласе. Ако се посматра први хијерархијски ниво међу класама, посебну важност има класа *pibas:Experiment* која је намењена за представљање концепта експеримента. 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. Класа *pibas:ModelSystem* означава концепт који представља биолошке системе (HCT-116, MDA-MB-231, итд). 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:Drug* се користи за дефиницију лека, али је заправо еквивалентна класи *pibas:ActiveSubstance*, јер је потребно да се изведе доста експеримената како би хемијска структура била квалификована као лек. Класа *pibas:Treatment* дефинише тип третмана (in vivo или in vitro). In vitro, за разлику од in vivo, се односи на студију или експеримент који се обавља у лабораторији у епруветама или лабораторијском посуђу. Док се in vitro односи на експерименте које се обављају на живом организму (животиња, човек). Класа *pibas:Material* дефинише алате, уређаје и хемијске структуре које се користе за спровођење експеримента. 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 кроз испитивање хируршки уклоњених органа, ткива (узорци биопсије), телесних течности, а у неким случајевима и целог тела (аутопсија). Класа *pibas:Dose* бележи концентрацију неког третмана који је коришћен у експерименту.

На другом хијерархијском нивоу су дефинисане подкласе (*owl:subClassOf*) одговарајућих класа из првог хијерархијског нивоа. Издвојићемо подкласе класа *pibas:ModelSystem* и *pibas:ExperimenatalMethod*. Подкласе, класе *pibas:ModelSystem* су *pibas:Animal* (концепт разноликости која се налази код животиња), *pibas:CellCulture* (представљање ћелијских линија) и *pibas:Patient* (представља пацијенте који учествују у експериментима). Међу подкласама класе *pibas:ExperimenatalMethod* издвајамо *pibas:MolecularBiologyProtocol* која представља протоколе (анализе) који ће се користити за тестирање биолошких система (*pibas:DNAandRNAanalysisAssay* и *pibas:ProteinAnalysisAssay*) и подкласу *pibas:CellCultureAssay* која дефинише есеје (ћелијске тестове) који се користе у експерименталним приступима. У оквиру PIBAS онтологије дефинисана су и одговарајућа објектна и својства типа података. Табеле 4.1 и 4.2 приказују њихове домене, односно кодомене. Хијерархијски преглед класа по нивоима се може видети на Слици 4.1.

Због сложене организације Лабораторије и начина чувања претходних података (релациона база података, MS Word и Excel извештаји), најподесније је било податке сваког експеримента сместити у посебне онтологије због њихове лакше манипулације. Назив који је коришћен за фајл сваког експеримента је *expID.owl*, где је ID ознака експеримента, по принципу примарног кључа у релационим базама података. С обзиром да су извештаји поседовали неке додатне информације у односу на структуру PIBAS онтологије као што су информације о истраживачима који су извели експеримент, за потребе које институције, у које време, ко је одговорно лице итд. извршено је увођење



Слика 4.1: Таксономија концепата PIBAS онтологије.

Табела 4.1: Објектна својства (Object Property) PIBAS онтологије са доменима и кодоменима.

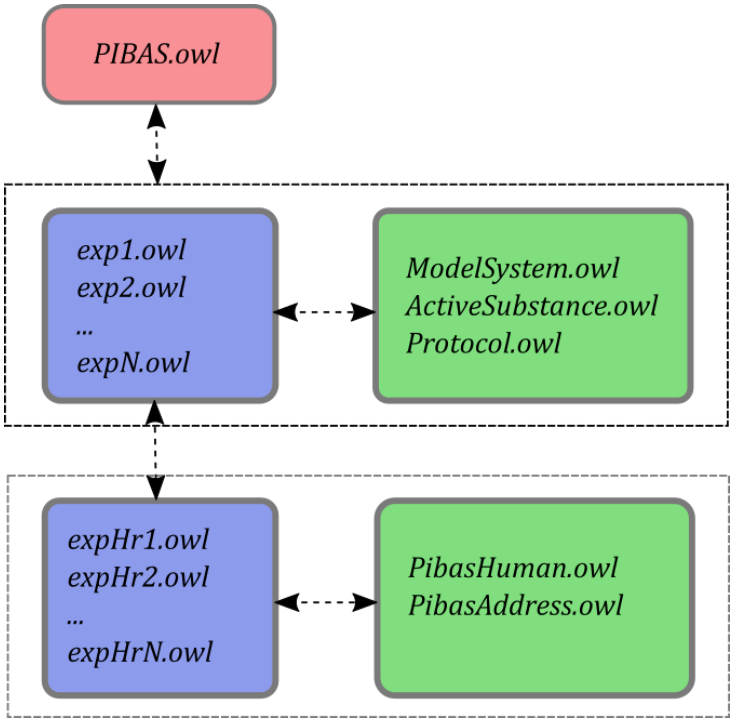
Назив	**Домен (rdfs:domain)**	**Кодомен (rdfs:range)**
pibas:experimentalMethod	pibas:Experiment	pibas:ExperimentalMethod
pibas:typeOfExperiment	pibas:Experiment	pibas:TypeOfExperiment
pibas:treatment	pibas:ModelSystem	pibas:Treatment
pibas:typeOfTreatmentInExperiment	pibas:ModelSystem	pibas:TypeOfTreatmentInExperiment
pibas:functionalSystem	pibas:TypeOfTreatment	pibas:FunctionalSystem
pibas:modelSystem	pibas:TypeOfTreatment	pibas:ModelSystem
pibas:signalingPathways	pibas:TypeOfExperiment	pibas:SignalingPathways
pibas:protocol	pibas:ExperimentalMethod	pibas:MolecularBiologyProtocol
pibas:material	pibas:ExperimentalMethod	pibas:Material
pibas:activeSubstance	pibas:TypeOfTreatmentInExperiment	pibas:ActiveSubstance
pibas:dose	pibas:TypeOfTreatmentInExperiment	pibas:Dose
pibas:drug	pibas:TypeOfTreatmentInExperiment	pibas:Drug
pibas:patology	pibas:TypeOfTreatmentInExperiment	pibas:Pathology

нових класа и својстава. Додатно је развијена онтологија *PibasHuman.owl* која садржи све информације које нису везане за сам експеримент, већ за учеснике и налогодавце. Такође, и за ову онтологију, су креирани посебни фајлови за сваки експеримент под називом *expHrID.owl*. Дакле, семантички базирана база података Лабораторије садржи PIBAS.owl онтологију (која представља структуру експеримента без инстанци), затим онтологије *ModelSystem*, *ActiveSubstance*, *Protocol*, *PibasHuman* и *PibasAddress* (са својим инстанцама), као и онтолошке фајлове (*expID.owl* и *expHrID.owl*) који садрже податке

Табела 4.2: Својства типа података (Datatype Property) PIBAS онтологије са доменима и кодоменима.

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

појединачних експеримента [49, 50]. Слика 4.2 представља декомпозицију базе података са две јасне целине за експеримент и истраживаче који га изводе.



Слика 4.2: CPCTAS база података.

Основни подаци везани за CPCTAS базу података се сада могу додати у RepoIntegration.owl онтологију, као и одговорајући подупити који се могу добити на основу знања о самој бази података. Како би се избегло стално упознавање истраживача са структуром нових скупова података, направљено је мапирање између супстанци наше базе података и неког од скупова, нпр. KEGG, DrugBank, итд. У наставку дајемо три примера SPARQL упита (Листинзи 4.1, 4.2, и 4.3) над креираном базом података који се могу извршити на ендпоинту: http://cpctas-lcmb.pmf.kg.ac.rs:2020/sparql.html.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX pibas: <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
SELECT ?Experiment ?ProtocolName ?CommentOfProtcol
{
    {
        ?Experiment rdf:type pibas:Experiment;
        pibas:ID 91;
        pibas:experimentalMethod ?ExperimentalMethod.
        ?ExperimentalMethod pibas:Protocol_name ?ProtocolName.
    }
    OPTIONAL
    {
        ?Experiment rdf:type pibas:Experiment;
        pibas:ID 91;
        pibas:experimentalMethod ?ExperimentalMethod.
        ?ExperimentalMethod rdfs:comment ?CommentOfProtcol.
    }
}
```

Листинг 4.1: Протоколи експеримента ID=91 и њихови коментари.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX pibas: <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
SELECT ?cellline ?IC50value
WHERE
{
    pibas:105 pibas:sameAs ?mapping_node.
    pibas:105 pibas:IC50value ?IC50value.
  pibas:105 pibas:hasCellLine ?cellline.
}
```

Листинг 4.2: Тестиране ћелијске линије и IC_{50} вредност за супстанцу pibas:105.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX pibas: <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
SELECT ?Experiment ?Manager ?Researcher ?ResponsibleResearcher ?User ?UserRepresentative
{
  ?Experiment rdf:type pibas:Experiment;
  pibas:ID 83;
  pibas:manager ?b;
  pibas:researcher ?c;
  pibas:responsibleResearcher ?d;
  pibas:user ?e;
  pibas:userRepresentative ?f.
  ?b foaf:name ?Manager.
  ?c foaf:name ?Researcher.
  ?d foaf:name ?ResponsibleResearcher.
  ?e foaf:name ?User.
  ?f foaf:name ?UserRepresentative.
}
```

Листинг 4.3: Особе које су одговорне за експеримент ID=83.

4.2 Додавање скупа података у репозиторијум

Ова секција је посвећена објављивању нових података на Вебу. Како бисмо податке потенцијално учинили јавно доступним, подаци би требало да буду повезани са подацима других скупова повезивањем ентитета. Према LOD cloud статистици[2] скоро сви скупови података имају више од хиљаду линкова ка другим ресурсима. Али, процес мапирања је временски захтеван, при чему сви скупови имају различите предикате у оквиру различитих репозиторијума. На пример, DrugBank предикати за таргете лекова у Bio2RDF-у и Chem2Bio2RDF-у су различити (http://bio2rdf.org/drugbank_vocabulary:-target; http://chem2bio2rdf.org/drugbank/resource/CID_GENE) што нам одмах говори и да се упити разликују. Користећи принцип јединственог идентификатора на којем се базирају релационе базе података, овде је предложено једноставно мапирање изабраних ентитета у скуповима података које се може урадити веома брзо. За процес мапирања су изабране хемијске структуре које су носиоци информација у репозиторијумима. Слично се може урадити са, на пример, ћелијским линијама, таргетима, или другим важним ентитетима. Ако бисмо рецимо узели област друштвеног умрежавања (из LOD облака), тада битни ентитети могу бити кориснички профили, групе, фан странице и слично. У следећем пасусу је приказано неколико детаља који се тичу повезивања нових скупова података са репозиторијумом.

Како би се подаци представили истраживачкој заједници, први предуслов који мора бити испуњен је представљање података помоћу технологија Семантичког Веба, пратећи принципе "Повезаних података"[3] (*Linked Data*). Овај услов је испуњен у претходном кораку. У циљу скалабилности платформе, процес мапирања је урађен тако да додавање нових скупова података не утиче на њену функционалност. Репрезентација једне супстанце из Лабораторије (*pibas:102*) треба да се мапира са репрезентацијом друге супстанце помоћу њеног идентификационог броја (cid) из оригиналног скупа (*pubchem:1235*), без обзира који URI јој је додељен у репозиторијуму. На овај начин се постиже флексибилност и за друге сличне лабораторије које желе да се интегришу. Приликом мапирања се никако не треба везивати само за ентитете из једног скупа података, јер супстанца може да постоји у једном репозиторијуму, али не и у другом. Због једноставности, у експериментима је коришћено мапирање које је ограничено на четири скупа: PubChem [36], DrugBank [37], ChEBI [51] и KEGG [38]. Ова листа се по потреби може проширити. Листинг 4.4 приказује пример једне такве мапе која повезује ентитете из Лабораторије са другим ентитетима. На исти начин се могу креирати мапе и за друге лабораторије. У експерментима су за потребе демонстрације методологије коришћене мапе за PIBAS [49] и CHEMBL [52].

```
<owl:NamedIndividual rdf:about="&pibas;102">
  <pibas:sameAs>pubchem:1235</pibas:sameAs>
  <pibas:sourceNumber>22</pibas:sourceNumber>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="&pibas;103">
  <pibas:sameAs>drugbank:DB00093</pibas:sameAs>
  <pibas:sourceNumber>2</pibas:sourceNumber>
</owl:NamedIndividual>
<owl:NamedIndividual rdf:about="&pibas;104">
```

[2] http://lod-cloud.net/state/

[3] http://www.w3.org/standards/semanticweb/data

```
    <pibas:sameAs>kegg_ligand:C10107</pibas:sameAs>
    <pibas:sourceNumber>6</pibas:sourceNumber>
</owl:NamedIndividual>
```

Листинг 4.4: Један део PIBAS мапе.

Глава 5

Претраживање повезаних података - преглед литературе

Након што је приказано на који начин се подаци представљају и како су међусобно повезани, следеће питање које се природно намеће је на који начин се може спровести претраживање репозиторијума, имајућу у виду да подаци могу бити дистрибуирани тј. смештени на једној или више локација.

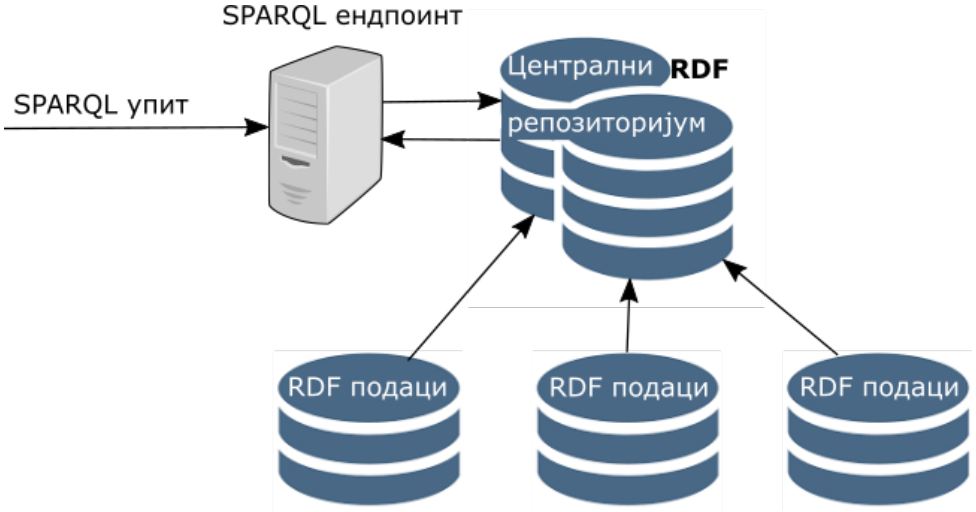
Начин на који су подаци смештени у великој мери одређује и начин на који се подаци могу претраживати. У првом делу поглавља биће описане инфрастуктуре за претраживање повезаних података са посебним нагласком на начине како се претраживање спроводи. Затим ће бити дат свеобухватни преглед литературе која је везана за начине претраживања репозиторијума и креирања SPARQL упита.

5.1 Инфраструктура за претраживање повезаних података

Данас постоји неколико семантички базираних репозиторијума (initiatives) за биолошке и хемијске скупове података који су детаљно описани у Поглављу 3: *Bio2RDF* [35], *LODD* [53], *Chem2Bio2RDF* [54], *EMBL-EBI* [55], *Open PHACTS* [44], *ChemSpider* [56] итд. Већина RDF инфраструктура смешта податке локално на основу одређених правила и третира их као један репозиторијум знања. То значи да су RDF модели локално реплицирани са удаљених ресурса и спојени у један модел, без обзира на дистрибуирану природу Семантичког веба. Међутим, у многим случајевима смо присиљени да приступамо удаљеним скуповима података из наше RDF инфраструктуре, без могућности да креирамо локалну копију скупа који садржи одређене податке. На пример, некада немамо дозволу да умножавамо податке, или је скуп јако велики да би се креирао само један модел који садржи све податке, подаци нису доступни у RDF формату и тако даље [57]. На другој страни, *Open PHACTS Discovery платформа* [45] увек креира локалне копије због бољих перформанси, остављајући податке у оригиналном формату и форми. Ова платформа пружа интегрисани приступ за 11 скупова података обухватајући податке из области хемије, биолошких путања, и протеина. Упити који на основу задатог контекста врше екстракцију одређених делова из скупова се заснивају на ком-

поненти *Identity Mapping Service* која повезује одређене појмове из различитих скупова података, налик мапирању које ми спроводимо у Платформи. Ипак, ниједан репозиторијум не може да обухвати на овај начин све скупове података, што говори у прилог тези да треба омогућити и удаљени дистрибуирани приступ подацима са различитих локација. На овај начин се обезбеђује да подаци буду увек „свежи" и омогућава се скалабилност самих апликација (лака интеграција нових података).

У зависности од локације података, инфраструктура за претраживање повезаних података може бити подељена на две категорије, када имамо приступ централном репозиторијуму, и случај када су репозиторијуми дистрибуирани. Концепт централизованог репозиторијума је сличан ономе који се користи за прикупљање података (енг. *data warehousing*) у традиционалним базама података, где се подаци најпре сакупљају и складиште у оквиру једног репозиторијума пре него што се приступи њиховој даљој обради и претраживању (Слика 5.1). Поред већ поменуте *Open PHACTS Discovery* платформе, пример још једног таквог репозиторијума је *Sindice* [58], који сакупља податке, индексира их и омогућава приступ подацима преко API-ја, али и приступ преко SPARQL ендпоинта. Централни репозиторијум такође има и *LinkedLifeData* иницијатива.

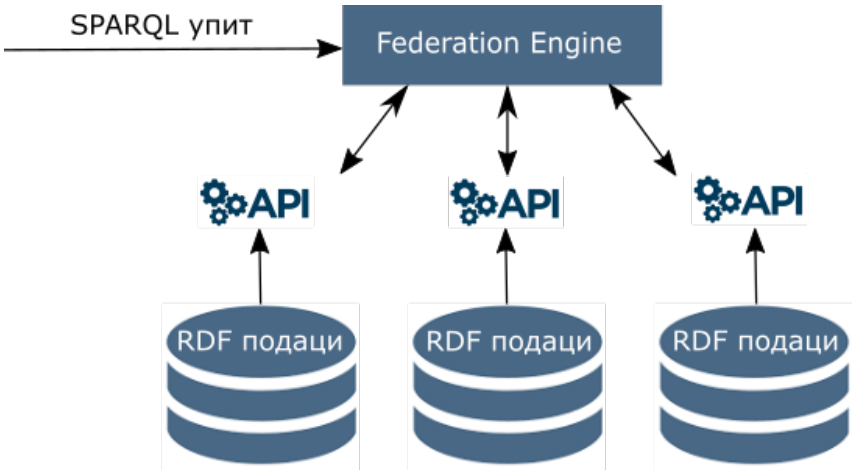


Слика 5.1: Претраживање централног репозиторијума.

Главна предност оваквог начина складиштења података је ефикасно време извршавања упита што проистиче из чињенице да је комуникација преко мреже елиминисана, као и избор ресурса. Међутим, због честих промена података, процес синхронизације података може да постане велики проблем [59]. На другој страни складиште за податке захтева пуно слободног простора да би сви подаци били на једном месту. Узмимо на пример компанију *Seven Bridges Genomics*[1] која складишти више од 1.5 петабајта података на својим серверима. Такође, поред простора интезивно се троше и други ресурси, јер истовремено треба обрадити велику количину података.

За разлику од централизованог складишта, претраживање повезаних података (Linked Data) у дистрибуираном окружењу не захтева сакупљање података на почетку. Овакави приступи се могу додатно груписати као *Link Traversal* и *Federation* приступи. Link traversal системи се базирају на откривању података пратећи HTTP URI-је. Дакле, без

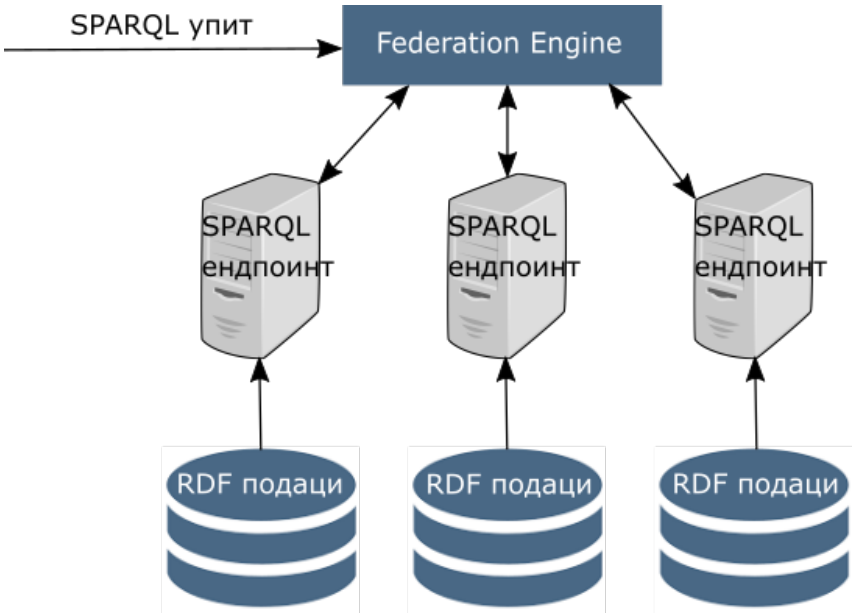
[1] https://www.sevenbridges.com/



Слика 5.2: Федерација над једним репозиторијумом коришћењем API позива.

било каквог почетног знања, релевантни скупови подака се откривају за време извршавања [60]. Добра страна оваквог приступа су увек најсвежији подаци. Међутим, у системима оваквог типа извршавање упита почиње једним шаблоном триплета што ствара велику зависност од почетног корака и потенцијалном увећању међурезултата. Још један недостатак овог система је ограничење у типу упита који се могу користити, као што је могућност коришћења дефинисаних предиката у шаблонима триплета ограничена.

На другој страни, системи за федерацију (енг. *federation*) користе посредника да би се кориснички упит трансформисао у више подупита и вратио резултате из интегрисаних извора података. С обзиром да подаци претходно не морају бити сакупљени у један репозиторијум, сви подаци су увек најновијег датума, али само извршавање упита је нешто дуже од оног када се ради са инфраструктуром која има централно складиште. Добре стране овог приступа су мало коришћење ресурса, али сама комуникација између посредника и скупа података је увећана, јер се избор ресурса врши у току извршавања упита. Додатно, постоје две врсте система за федерацију: *федерација над једним репозиторијумом* (Слика 5.2) и *федерација над више SPARQL ендпоинта* (Слика 5.3). Федерација над једним репозиторијумом се обично обавља коришћењем изворног API-ја његовог репозиторијума где се упит дели над подупите који се даље прослеђују API-ју. Пример једног оваквог приступа је *Sesame Sail Federation* [61]. Међутим, не подржавају сви скупови података API позиве, па је онда потребно посебно прикупљање података и смештање на репозиторијум. Федерација над више репозиторијума истовремено захтева SPARQL ендпоинт који се користи као мост између федерације и скупова података. Свеобухватни преглед ових приступа се може видети у раду [62]. Већина софтверских платформи за федерацију је компатибилно са оваквим приступом, јер су репозиторијуми за триплете углавном подржани SPARQL ендпоинтима (за детаље видети Секцију 3.1). За нас су од посебног интереса системи за федерацију над више SPARQL ендпоинта, јер Платформа користи ову инфраструктуру за приступ подацима. У тексту који следи биће поменути неки најважнији од њих.



Слика 5.3: Федерација помоћу SPARQL ендпоинт-а.

5.2 Претраживање семантичких репозиторијума помоћу федерације

За ефикасно процесирање упита у семантички оријентисаном окружењу развијени су софистицирани генератори упита и бенчмарк (енг. *benchmark*) системи за евалуацију њихових перформанси. Са наше тачке гледишта, недостаци упита из бенчмарк система проистичу из чињенице да се ослањају на предефинисане статичке упите који се извршавају над тачно одређеним скуповима података [63][64][65] тако да су тешко употребљиви за добијање одговора на питање, јер нису са том намером ни направљени. На другој страни, и аутоматски генератори упита се суочавају са многим проблемима и недостацима. Прво, процес постављања параметара за алгоритме и границе (енг. *thresholds*) може бити доста сложен без претходог знања о подацима. Све ово води ка сакупљању различитих статистика које се мењају током времена. Друго, у тако великом броју генерисаних упита, многи од њих не враћају резултате, док многи као резултат враћају непотребне податке. У исто време, процеси тражења "најбољих" упита, њихово извршавање и евалуација, захтевају доста времена. Треће, ови приступи не могу да претражују више репозиторијума (садрже више скупова података) пратећи њихову специфични начин интеграције. Веома често се дешава да интеграција репозиторијума није могућа, јер не постоји мапа (енг. *mapping schemes*) која их повезује. Додатна отежавајућа околност проистиче из чињенице да се све време ради са скуповима чија је структура у потпуности различита, при чему су и саме везе између скупова унутар репозиторијума различито дефинисане о чему је већ било речи у Секцији 3.

Аутоматски генератори упита су мање оријентисани ка тачно одређеном циљу и зато као резултат могу да генеришу много упита који су корисни само за евалуацију начина и брзине њиховог извршавања, али не и за крајње кориснике и њихове захтеве. Употребљиви и реални упити могу бити генерисани само ручно (*user-guided*) улажући при томе доста напора, јер садржај скупова података мора да буде унапред анализиран.

Ипак, проблем интеграције података из више скупова и репозиторијума још увек представља отворени изазов на који се истраживачка заједница фокусирала и који активно решава. Све претходно наведено јасно указује да ручно направљени упити захтевају много уложеног времена за упознавање структуре и садржаја скупова података, док аутоматски генерисани упити могу да произведу много упита који морају бити ручно проверени, изабрани и модификовани за даље коришћење. Наведени недостаци оба приступа остављају отворени простор за додатна побољшања и развој хибридних решења која користе предности како једних, тако и других приступа.

У наставку секције је дат преглед литературе за оба типа генератора упита, оне који упите генеришу аутоматски и оне које корисници сами креирају према својим потребама. Овом приликом је направљен и осврт на њихове главне разлике и недостатке у односу на нашу софтверску платформу. SpecINT софтверска платформа може да се третира као решење које се налази између ова два приступа, јер генерисани упити у оквиру ње враћају релевантне резултате, при чему не зависе од корисника и личног искуства. Прецизније, SpecINT није аутоматски генератор упита који упите генерише у потпуности самостално од почетка, већ користи уграђено експертско знање узимајући већ направљене подупуте и од њих креира коначни SPARQL упит. Такође, платформа се мање ослања на корисничку интервенцију, јер математички апарат који је имплементиран у платформи пружа задовољавајућу тачност упита и избор правих скупова података.

5.2.1 Аутоматски генератори упита

Постојећи аутоматски генератори упита су развијени за детаљну анализу и евалуацију различитих језгара (енг. *engines*) који извршавају Federated SPARQL упите. Ови системи за федерацију су базично развијени за оптимизацију времена извршења упита и смањење међурезултата, тако да њихови генератори упита не могу бити коришћени за добијање задовољавајућих резултата и корисничког искуства. Иако се неки генератори упита односе на дистрибуиране изворе података, они не могу да бирају скупове над којима ће да се извршавају (*on-the-fly*), нити могу да повежу два репозиторијума без глобалног мапирања. Могућност избора оних скупова података који имају највећу вероватноћу да садрже релевантне податке/триплете је прави бенефит и управо је на томе један од фокуса наше платформе. Примери оваквих генератора упита су дати у наставку.

FedX [63] је развијен за поређење опште сврхе система за SPARQL упите. Он је фокусиран на стратегију како смањити број трансмисија упита и на смањивање међурезултата, док његов недостатак лежи у чињеници да се базира на скупу предефинисаних статичких упита који обухватају тек неколико скупова података. *FedBench* [66] је једини бенчмарк (енг. *benchmark*) за Federated SPARQL упите који прати и мери перформансе инфраструктуре мерећи при томе време учитавања и време извршавања упита. Међутим, *FedBench* такође има статичке упите и скупове података.

DAW [67] садржи скуп статичких упита који се базирају на карактеристикама *BSBM (Berlin SPARQL Benchmark)* упита [31] из четири јавно доступна скупа података. Ипак, ови упити су генерисани коришћењем статистичких метода и као такви не могу да се искористе за специјализоване, наменске системе за федерацију. На другој страни, ови

упити су прилично једноставни по структури и комплексности (садрже максимално 4 триплета по упиту). Да би се решио овај проблем, неки системи генеришу случајне скупове упита за одређени скуп података. У студији која је спроведена од стране Umbrich-а и његовог тима [59], проширена је семантика упита која повезује *Linked Data упите (LidaQ)*. За креирање бенчмарк Federated SPARQL упита овај генератор упита креира скуп сличних упита коришћењем случајних шетњи по ширини или дубини, базирајући се на 3 основна облика (entity, star и path shapes). Скуп упита у оквиру *SPLODGE* [68] је базиран на карактеристикама скупа података које су добијене из статистике предиката. На основу случајног процеса за генерисање упита *SPLODGE* користи процену кардиналности и сасвим је реално да различити упити са истим каратеристикама дају различите величине резултата. *DARQ* [69] и *SPLENDID* [64] користе различите статистичке информације (користећи ручно сакуљене описе скупова или *VOID*) пре него сам садржај. Неки скупови података се константно увећавају, тако да апликације морају често да узимају најновије RDF скупове. Ипак, одржавање обимних и ажурираних податке у кешу је превише захтеван задатак. Ново побољшање долази са *ANAPSID*-ом [70] које се огледа у ажурирању каталога података и плану извршавања у току извршавања (*runtime*). За додатни преглед поменутих система видети [71].

За разлику од статистички базираних генератора, *FEASIBLE* [72] представља аутоматски приступ за генерисање бенчмаркова независно од историје упита у оквиру апликације тј. *log* фајлова упита. Генерисање упита се постиже избором прототипских упита чију величину дефинише корисник, а на основу улазног скупа упита. Слично, у раду [73] је предложен *SQCFramework*, SPARQL бенчмарк софтверски оквир који је генерише прилагођене SPARQL упите на основу историје претходних упита. Коришћењем различитих алгоритама за кластеризацију података, софтверски оквир може да генерише бенчмаркове различитих величина, са различитим значајним SPARQL особинама.

Поред раније наведених недостатака аутоматских генератора упита, ови генератори се извршавају над скуповима података који су дати унапред, и немају могућност укључивања нових скупова без претходног статистичког прорачуна или глобалног мапирања. Такође, они не могу да раде са више репозиторијума у исто време који имају различите предикате и везе. Једино решење које се експицитно бави овом тематиком тј. интегрисаним упитима над више RDF репозиторијума је описано у [57]. Stuckenschmidt и остали су теоријски описали како да се прошири *Sesame RDF* [61] репозиторијум да би се подржао рад *SeRQL* упита над више ових репозиторијума, при чему се препоручује коришћење специјалне структуре за индексирање како би се одредили релевантни скупови за упите. Ипак, овај приступ није имплементиран, већ је чисто теоријске природе.

5.2.2 Кориснички дефинисани упити

На другој страни, многе постојеће апликације пружају кориснички интерфејс за претраживање биоинформатичких скупова података, пре чему се корисницима дозвољава интуитивно креирање и извршавање Federated SPARQL упита. Ове апликације могу да креирају корисне упите што произилази из чињенице да корисници пролазе кроз одређене кораке у оквиру интерфејса, бирајуће одговарајуће скупове података (*endpoints*), предикате које апликација дозволи на основу ограничења, субјекте/објекте,

омогућавајући повезивање мањих делова у упите на основу експертског искуства. Примери оваквих апликације су: *GoWeb* [74], *SPARQLGraph* [75], *Smart* [76], *BioQueries* [77], *BioSearch* [78] итд. Ове апликације су дизајниране за визуелно креирање, ажурирање и извршавање SPARQL упита за област биологије. *PIBAS FedSPARQL* [79] је апликација која такође покреће Federated SPARQL упите за неколико биоинформатичких тема. У овој апликацији корисник пролази кроз систем и бира делове упита. Као напредну функционалност, *PIBAS FedSPARQL* пружа могућност детектовања сличних података користећи резултате предефинисаних упита као улаз. У оквиру евалуације наше софтверске платформе *PIBAS FedSPARQL* је коришћен за додатну проверу враћених резултата.

Међутим, све ове апликације се заснивају на корисничком искуству и афинитетима, док се недостаци рефлектују у немогућности да се лако додају нови скупови података и у подршци за мали број одређених ендпоинта.

5.2.3 SpecINT софтверска платформа

SpecINT софтверска платформа која је развијена као подршка у раду Лабораторије представља компромисно (енг. *trade-off*) решење између аутоматских и кориснички базираних генератора упита. Циљ који платформа треба да задовољи је омогућавање приступа потребним информацијама, иако корисницима синтакса SPARQL упита и организација репозиторијума нису блиски. SpecINT не захтева ангажовање корисника током фазе конструисања упита, за разлику од кориснички дефинисаних упита, јер се релевантни скупови података бирају на основу сопствених вектора графа.

SpecINT софтверска платформа је хибридно решење које подразумева улогу човека у креирању мањих упита (енг. *patterns, templates*) у фази препроцесирања подака, као битног корака за добијање задовољавајућих резултата. Касније се ови упити/обрасци користе приликом конструисања упита аутоматски, узимајући у обзир само најрелевантније скупове података (из различитих репозиторијума) који имају највећу вероватноћу да садрже триплете који нас интересују у том тренутку. За решавање овог проблема коришћени су сопствени вектори графа чије нам координате омогућавају праћење постојећих путања између скупова података и укључивање у упит оних најрелевантнијих за постављено питање. Све ове акције (избор скупова, повезивање термина, повезивање подупита итд.) се истовремено дешавају над више репозиторијума и у току извршавања (*on-the-fly*). Добијене путање у графовима који се посматрају доводе до доста добрих решења, при чему се репозиторијуми не морају посебно истраживати. За разлику од класичних *state-of-the-art* Federated SPARQL генератора упита који зависе од заједничких онтологија које повезују скупове подтака и њихове статистистичке анализе, наш приступ повезује исте те скупове унутар репозиторијума коришћењем сопствених вектора графа [80], а избор чворова (скупова) за упит се врши према њиховој значајности у графу [81] [82], без заједничких онтологија које повезују ресурсе.

Како би се због комплексности SpecINT софтверске платформе избегла додатна објашњења и искакање из оквира током представљања њене архитектуре, најпре је у наредном поглављу дат кратак преглед теоријских основа на којима се базирају одређени делови Платформе и које ће нам бити неопходне за разумевање предложене методологије за интеграцију и претраживање података. Ту ће пре свега бити дате основне

дефиниције из теорије графова, дефинисани појмови као што су спектар графа и њему одговарајући сопствени вектори, а затим приказан и математички доказ теореме на основу чијих резултата се одређује избор релевантних скупова података и креирање упита који их претражује. Након упознавања са свим технологијама Семантичког Веба, начином на који репозиторијуми приказују и повезују податке, тренутним решењима и математичком основом која се користи у одређеним корацима, целокупно стечено знање се сада може објединити и демонстрирати како је оно искоришћено на примеру Платформе која је овде представљена.

Глава 6

Математичке основе коришћене у развоју алгоритма за претраживање

Спектрална теорија графова је грана математичке комбинаторике, односно теорије графова, у којој се особине графа изучавају преко сопствених вредности и сопствених вектора матрица које су придружене графу. Данас, спектрална теорија графова има широку примену у различитим областима науке као што су хемија, физика, биологија, географија, економија, електротехника, итд. Осим поменутих, посебно се помиње и велика примена спектара графова у различитим областима информатике, медицине као и у друштвеним наукама. Како се непрестано откривају и нашироко примењују стари и нови резултати у овој теорији, то је и изучавање спектралних особина графа постало јако популарна и интересантна тема.

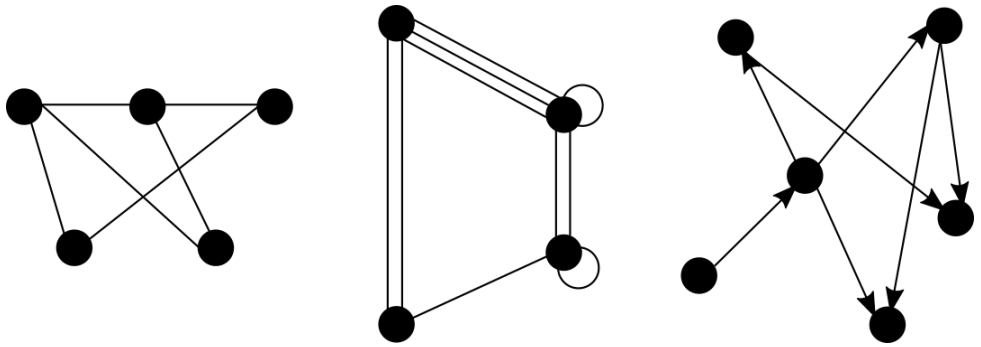
У последњих десетак година уочено је да спектри графова имају разне, веома важне примене у рачунарству. Спектри графови се појављују у Интернет технологијама, рачунарској обради слике, препознавању облика, обради велике количине података, статистичким базама података и многим другим областима. Постоји више хиљада таквих научних радова који говоре у прилог томе. Пре него што буде приказана сама архитектура SpecINT софтверске платформе, најпре ће бити речено нешто више о математичком појму графа на којем се она заснива.

6.1 Теорија графова

На почетку су дате неке основне дефиниције из теорије графова, а за више детаља и ширу слику о графовима и њиховим применама читаоца упућујемо на књиге [83, 84, 85, 86].

Дефиниција 6.1. Граф G је уређени пар (V, E), где је V коначан непразан скуп елемената који се називају *чворови* (енг. *vertex, node*), а E бинарна релација у V чији се елементи називају *ивице* (енг. *edge*) графа G тј. $E \subseteq \binom{V}{2}$.

Скупови V и E графа $G = (V, E)$ се најчешће означавају са $V(G)$ и $E(G)$, редом. За граф G кажемо да је реда $n = |V|$ и величине $m = |E|$.



Слика 6.1: Пример а) графа, б) мултиграфа и в) диграфа.

Дефиниција 6.2. Два чвора u и v графа G су суседна уколико постоји грана која их спаја ($\{u, v\} \in E$). Број чворова који су суседни са чвором v назива се степен чвора v и означава са $deg(v)$ или d_v, док се са $N(v)$ означава скуп свих суседа чвора v у графу G.

Дефиниција 6.3. Мултиграф је граф у чијем се скупу грана могу појављивати паралелне гране или петље (почетни и крајњи чвор се поклапају).

Дефиниција 6.4. Оријентисани граф или диграф је уређени пар (V, E), где је V коначан непразан скуп елемената који се називају *чворови*, а E коначан скуп уређених парова елемената из V које називамо *усмереним гранама*.

Код оријентисаних графова грана представља уређени пар чворова, односно, уколико је $e = (v_1, v_2) \in E$, тада кажемо да је грана e оријентисана од v_1 ка v_2. На слици 6.1 су представљени по један граф, мултиграф и диграф, редом. Код оријентисаних графова сваки крај гране се броји на различит начин. Другим речима, број грана које улазе у неки чвор је улазни степен, у ознаци $deg^+(v)$, а број грана које из чвора излазе је излазни степен, у ознаци $deg^-(v)$.

Граф који је коначан, неоријентисан, без петљи и вишеструких грана називамо простим графом и следеће дефиниције се односе на такве графове.

Дефиниција 6.5. Комплемент графа $G = (V, E)$ је граф $\overline{G} = (V, \binom{V}{2} \setminus E)$.

Дефиниција 6.6. Шетња дужине k у графу G је низ $v_0, e_1, v_1, e_2, v_2, \ldots, e_k, v_k$ чворова и грана тако да је $e_i = v_{i-1}v_i$ за $i = 1, 2, \ldots, k$. Чворови v_0 и v_k су крајњи чворови шетње W. Шетња је затворена уколико је $v_0 = v_k$. Стаза је шетња у којој се ниједна грана не понавља. Пут је шетња у којој се ниједан чвор не понавља. Циклус је затворена стаза у којој се ниједан чвор не понавља, изузев првог и последњег.

Чворови u и v графа G су повезани ако у G постоји пут чији су крајњи чворови управо u и v. Граф G је повезан уколико су свака два његова чвора повезана - у супротном је неповезан. Неповезани делови графа се зову компоненте повезаности графа.

Дефиниција 6.7. Чвор v графа G је артикулациони чвор ако се његовим уклањањем повећава број компоненти повезаности графа G.

Дефиниција 6.8. Грана e графа G је мост ако се њеним уклањањем повећава број компоненти повезаности графа G.

У наставку наводимо неколико неформалних дефиниција везаних за одређене типове графова.

Граф се назива *нетривијалним* ако садржи бар два чвора.

Граф реда n чији су сви чворови степена $n-1$ називамо *комйлетним їрафом*, у ознаци K_n, док граф реда n чији су сви чворови степена r називамо *реїуларним їрафом стейена* r.

Чворови степена 0 и 1 у графу називају се *изоловани* и *висећи* чворови, респективно.

Повезан граф са n чворова код кога су сви чворови степена 2 називамо *контуром* дужине n (енг. cycle) и означавамо са C_n.

Повезани графови T_n који не садрже контуре су *стабла* (енг. *tree*), где је n број чворова стабла.

Стабло са n чворова код кога је један чвор степена $n-1$, а сви остали чворови степена 1 називамо *звездом* (енг. *star*), у ознаци $S_{1,n-1}$. Чвор степена $n-1$ назива се *центар звезде*.

Повезан граф P_n са n чворова који не садржи контуре и у коме ниједан чвор није степена већег од 2 називамо *йутем* дужине n (енг. *path*).

6.2 Спектар матрице

Да бисмо увели појам спектра графа биће нам најпре потребно да дефинишемо одређене појмове који се односе на спектре матрица, јер ћемо графове посматрати кроз њихову матричну репрезентацију. Посматраћемо даље n-димензионални векторски простор $\mathbb{C}^n$, док ћемо векторе овог простора представљати матрицама колона типа $n\times 1$. На нивоу целе докторске дисертације са I ће бити означена дијагонална матрица где су на дијагонали све јединице, а са J матрица састављена само од јединица.

Ако је $\mathbf{x}\in\mathbb{C}^n$ и

$$\mathbf{x}=\begin{bmatrix} x_1\\ x_2\\ \vdots\\ x_n\end{bmatrix}=\begin{bmatrix} x_1 & x_2 & \cdots & x_n\end{bmatrix}^T,$$

тада се x_k назива k-том координатом вектора $\mathbf{x}$.

Дефиниција 6.9. Ако је A комплексна квадратна матрица реда n, тада сваки вектор $\mathbf{x}\in\mathbb{C}^n\setminus\{\mathbf{0}\}$ који задовољава услов

$$(\exists\lambda\in\mathbb{C})\ A\mathbf{x}=\lambda\mathbf{x} \tag{6.1}$$

називамо *сойственим вектором матрице* A (енг. *eigenvector*), а скалар λ *сойственом вредношћу матрице* A (A-сопствена вредност; енг. *eigenvalue*). За вектор $\mathbf{x}$ кажемо да припада сопственој вредности λ.

Једначина (6.1) се може записати у облику $(A-\lambda I)\mathbf{x}=\mathbf{0}$, одакле се може закључитити да су сопствене вредности матрице A нуле карактеристичног полинома $P_A(\lambda)=|\lambda I-A|$.

Дефиниција 6.10. Ако је r ред нуле λ карактеристичног полинома матрице A, тада кажемо да сопствена вредност λ матрице A има алгебарску вишеструкост r.

За матрицу A кажемо да је симетрична ако важи $A = A^\intercal$.

Дефиниција 6.11. Скуп сопствених вредности симетричне квадратне матрице A, заједно са њиховим алгебарским вишеструкостима, назива се *сīектром* те матрице (A-спектар). Ако су $\lambda_1 > \lambda_2 > \ldots > \lambda_k$ различите сопствене вредности матрице A, а $m(\lambda_1), m(\lambda_2), \ldots, m(\lambda_k)$, редом, њихове алгебарске вишеструкости, тада се спектар матрице A може записати у следећем облику

$$\sigma(A) = \begin{pmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_k \\ m(\lambda_1) & m(\lambda_2) & \cdots & m(\lambda_k) \end{pmatrix}$$

Релацију $A\mathbf{x} = \lambda\mathbf{x}$ можемо интерпретирати и на следећи начин: ако је $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, тада је $\lambda x_u = \sum_{v \sim u} x_v$, где се сумирање врши по свим суседима v чвора u.

Скуп сопствених вектора матрице A који одговарају сопственој вредности λ образује потпростор $\epsilon(\lambda)$ векторског простора $\mathbb{C}^n$. Димензија овог потпростора назива се *īеометријска вишеструкост* сопствене вредности λ.

За матрицу A кажемо да је реална ако су сви њени елементи реални бројеви. За реалну матрицу A кажемо да је ненегативна (позитивна), у ознаци $A \geq 0$ ($A > 0$) ако су сви елементи матрице ненегативни (позитивни). Ненегативна реална матрица $A \in R^{n \times n}$ је неразложива ако и само ако за сваки (i, j) постоји природан број s тако да је $(A^s)_{i,j} > 0$. У овој докторској дисертацији се ради само са реалним матрицама, па ћемо у наставку то и подразумевати.

Теорема 6.1 (Спектрална теорема за симетричне матрице)**.** Симетрична матрица $A \in R^{n \times n}$ има само реалне сопствене вредности, њихове одговарајуће алгебарске и геометријске вишеструкости су једнаке и постоји скуп њихових реалних сопствених вектора (величине n) који чини ортонормирану базу.

Теорема 6.2 (Perron-Frobenius)**.** Свака неразложива, ненегативна матрица A има позитивну сопствену вредност r, која је једнострука нула карактеристичног полинома. Модули свих осталих карактеристичних вредности нису већи од r. „Максималној" сопственој вредности r одговара сопствени вектор са позитивним координатама. Ако при томе матрица A има h карактеристичних вредности, по модулу једнаких r, ти бројеви су међусобно различити и задовољавају једначину $\lambda^h - r^h = 0$. Уопште, скуп карактеристичних вредности $\lambda_1 = r, \lambda_2, \ldots, \lambda_n$ матрице A, посматран као скуп тачака у комплексној равни, прелази сам у себе при ротацији равни за угао $\frac{2\pi}{h}$. За $h > 1$ могуће је пермутацијом врста и истом пермутацијом колона довести матрицу на следећи „циклички" облик:

$$\begin{bmatrix} O & A_{12} & O & \cdots & O \\ O & O & A_{23} & & O \\ \vdots & & & & \\ O & O & O & & A_{h-1,h} \\ A_{h1} & O & O & O & O \end{bmatrix},$$

где се дуж главне дијагонале налазе квадратне нула-матрице.

Варијанта претходне теореме за симетричне матрице гласи:

Теорема 6.3. Нека је $A \in R^{n\times n}$ неразложива симетрична матрица са сопственим вредностима $\lambda_1, \lambda_2, \ldots, \lambda_n$. Тада важи:

(i) $\lambda_1 > 0$,

(ii) алгебарска и геометријска вишеструкост највеће сопствене вредности λ_1 је 1,

(iii) $\lambda_1 > |\lambda_i|$ за $i = 2, \ldots, n$,

(iv) сопствени вектор који одговара сопственој вредности λ_1 има све строго позитивне координате,

(v) $\lambda_1 = -\lambda_n$ ако и само ако се A своди на форму

$$\begin{bmatrix} O & B \\ B^T & O \end{bmatrix}.$$

На основу Теореме 6.3 важи да је граф повезан ако и само ако му је највећа сопствена вредност проста и њој одговарајући сопствени вектор позитиван.

Матрица A је позитивно семидефинитна ако за сваки вектор $\mathbf{x} \in \mathbb{R} \setminus \{\mathbf{0}\}$ важи да је $\mathbf{x}^T A\mathbf{x} > 0$.

Теорема 6.4. Нека је $A \in R^{n\times n}$. Тада су следећа тврђења еквивалентна:

- A је позитивно семидефинитна
- све сопствене вредности матрице A су ненегативне
- постоји $m \in \mathbb{N}$ и матрица $B \in \mathbb{R}^{m\times n}$ тако да је $A = B^T B$

6.3 Спектар графа

У овом делу је представљена веза између теорије графова и линеарне алгебре. Графу (диграфу) G са n чворова се може на природан начин придружити матрица суседства која се дефинише на следећи начин.

Дефиниција 6.12. *Матрица суседства* графа, чији је скуп чворова $\{v_1, v_2, \ldots, v_n\}$, је квадратна матрица $A = (a_{ij})$ реда n чији су елементи дефинисани на следећи начин:

$$a_{ij} = \begin{cases} 1, & \text{ако } \ v_i v_j \in E \\ 0, & \text{ако } \ v_i v_j \notin E \end{cases}. \tag{6.2}$$

За прост граф матрица суседства је симетрична матрица, док се на главној дијагонали налазе нуле. На сличан начин се дефинишу матрице суседства (мулти)(ди)графа, где је a_{ij} вредност једнака броју грана, или стрелица, које излазе из чвора i и завршавају се у чвору j.

Дефиниција 6.13. *Карактеристични полином* $\Phi(G, \lambda)$ матрице суседства графа (дијаграфа) G назива се карактеристични полином графа G и означава се са $\Phi(G, \lambda) = |\lambda I - A|$.

Карактеристични полином графа (дијаграфа) G са n чворова можемо приказати и у следећем облику:

$$\Phi(G, \lambda) = \sum_{k=0}^{n} a_k \lambda^{n-k},$$

односно

$$\Phi(G, \lambda) = a_0\lambda^n + a_1\lambda^{n-1} + \ldots + a_n$$

при чему се a_k назива његовим k-тим коефицијентом.

Дефиниција 6.14. Нека су $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ нуле карактеристичног полинома $\Phi(G, \lambda)$ графа (дијаграфа) G, односно решења једначине $\Phi(G, \lambda) = 0$. Тада се $\lambda_1, \lambda_2, \ldots, \lambda_n$ називају сопственим (карактеристичним) вредностима графа (дијаграфа) G, а све заједно чине спектар графа (дијаграфа) G.

Пример 6.1. Спектар пута P_n се састоји из бројева облика $2\cos\frac{\pi}{n+1}\,k$ где је $k = 1, \ldots, n$.

Пример 6.2. Спектар контуре C_n се састоји из бројева облика $2\cos\frac{2k\pi}{n}$ где је $k = 0, 1, \ldots,$-$n-1$.

Пример 6.3. Нека је G комплетан граф K_n. Његова матрица суседства је $A = J - I$, а спектар чине бројеви $n-1$ и -1, вишеструкости 1 и $n-1$, редом.

Највећу сопствену вредност графа G означавамо са $\lambda_1(G)$ називамо је *индексом* или *спектралним радијусом* графа G (енг. *index, spectral radius*). Јединствени позитиван (свака координата је позитивна) сопствени вектор који одговара индексу повезаног графа G назива се *главни сопствени вектор* од G (енг. *principal eigenvector*).

Дефиниција 6.15. Граф $H = (V_1, E_1)$ је подграф (енг. subgraph) графа $G = (V, E)$, ако важи $V_1 \subseteq V$ и $E_1 \subseteq E \cap \binom{V_1}{2}$. Граф H је *индуковани подграф* графа G (енг. induced subgraph) ако скуп E_1 садржи све гране из E које повезују чворове из скупа V_1.

Теорема 6.5 (Теорема о укљештењу)**.** (видети [85], страна 18) Нека је G граф са n чворова и сопственим вредностима $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, и нека је H индуковани подграф од G са m чворова. Ако су $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_m$ сопствене вредности од H тада $\lambda_{n-m+i} \leq \mu_i \leq \lambda_i$ $(i = 1, 2, \ldots, m)$.

Пример 6.4. Индекс регуларног графа реда r је број r.

Нека је G неповезан граф са компонентама повезаности $G_1, G_2, \ldots, G_k$, које можемо посматрати и као самосталне графове. Нека су $A_1, A_2, \ldots A_k$ матрице суседства редом графова $G_1, G_2, \ldots G_k$. Тада матрица суседства графа G има облик

$$A(G) = \begin{pmatrix} A_1 & O & \cdots & O \\ O & A_2 & & O \\ \vdots & & \ddots & \\ O & O & & A_k \end{pmatrix}.$$

Одавде се може видети да за карактеристичне полиноме поменутих графова важи релација

$$\phi(G, \lambda) = \phi(G_1, \lambda) \cdot \phi(G_2, \lambda) \cdots \phi(G_k, \lambda). \tag{6.3}$$

Дакле, спектар графа G добија се обједињавањем спектара његових компоненти повезаности $G_1, G_2, \ldots G_k$.

Теорема 6.6. Број компоненти графа G једнак је максималном броју линеарно независних ненегативних сопствених вектора од G.

Теорема 6.7. Ако је G' граф добијен из графа G брисањем било којег чвора или гране, тада важи $\lambda_1(G') \leq \lambda_1(G)$. Неједнакост је стриктна када је G повезан.

Пре него што прикажемо на који начин је структура графа повезана са вредностима координата сопствених вектора увешћемо додатну нотацију. За произвољан вектор $z = (z_1, z_2, \ldots, z_n)^T \in R^n$ нека су скупови $\mathcal{P}(z)$, $\mathcal{N}(z)$ и $\mathcal{O}(z)$ дефинисани на следећи начин:

$$\mathcal{P}(z) = \{i : z_i > 0\},\ \mathcal{N}(z) = \{i : z_i < 0\},\ \mathcal{O}(z) = \{i : z_i = 0\}.$$

Претпоставимо да је z вектор чије се координате придружују скупу чвора $\{1, 2,$-$\ldots, n\}$ графа G. Казаћемо да је знак чвора i *йозитиван, неїативан или нула* (у односу на z) према томе да ли i припада скупу $\mathcal{P}(z)$, $\mathcal{N}(z)$ или $\mathcal{O}(z)$, редом.

6.3.1 Лапласова матрица

Дефиниција 6.16. Лапласова матрица $\mathcal{L}$ графа G дефинише се на следећи начин:

$$\mathcal{L} = D - A,$$

где је D дијагонална матрица чији су елементи на главној дијагонали степени чворова графа, а A матрица суседства графа.

Дефиниција 6.17. Лапласов карактеристични полином графа G задат је релацијом $\Psi(G, \mu) = |\mu I - \mathcal{L}(G)|$, где је $\mathcal{L}(G)$ Лапласова матрица.

Лапласов карактеристични полином можемо приказати и на следећи начин:

$$\Psi(G, \mu) = \sum_{k=0}^{n} (-1)^k c_k \mu^{n-k},$$

односно

$$\Psi(G, \mu) = c_0\mu^n - c_1\mu^{n-1} + \ldots + (-1)^{n-1} c_{n-1}\mu + (-1)^n c_n$$

при чему за коефицијенте c_k важи $c_k \geq 0$.

Слично, као у случају обичног карактеристичног полинома и $\Psi(G, \mu)$ је полином n-тог степена са променљивом μ. Једначина $\Psi(G, \mu) = 0$ има тачно n решења, од којих нека могу бити међусобно једнака и сва су реални бројеви. Означићемо их са $\mu_1, \mu_2, \ldots, \mu_n$.

Дефиниција 6.18. Нека су $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_n$ нуле Лапласовог карактеристичног полинома $\Psi(G, \lambda)$ графа G. Бројеви $\mu_1, \mu_2, \ldots, \mu_n$ се називају *Лапласове сопствене вредности* графа G (L-сопствене вредности), а сви заједно чине *Лапласов спектар* графа G (L-спектар).

Теорема 6.8. Нека је G тежински граф са ненегативним тежинама. Тада:

1. $L(G)$ има само реалне сопствене вредности,
2. $L(G)$ је позитивна семидефинитна,
3. Његова најмања сопствена вредност је $\mu_1 = 0$ и одговарајући сопствени вектор је $(1, 1, \ldots, 1)^T$. Мултиплицитет од μ_1 је једнак броју компоненти од G.

Теорема 6.9. Лапласове сопствене вредности графова G и $G' = G - v$ су укљештене, тј.

$$0 = \mu_1(G) = \mu_1(G') \leq \mu_2(G) \leq \mu_2(G') \leq \mu_3(G) \leq \ldots \leq \mu_n(G) \leq \mu_n(G').$$

Теорема 6.10. Ако $\overline{G}$ означава комплемент графа G тада

1. $\Psi(\overline{G}, \mu) = (-1)^{n-1} \frac{x}{n-x} \Psi(G, n - x)$
2. $\mu_1(\overline{G}) = 0$
3. $\mu_{i+1}(\overline{G}) = n - \mu_{n-i+1}(G)$, за $i = 1, 2, \ldots, n-1$

Друга најмања сопствена вредност μ_2 Лапласове матрице графа G се назива *алгебарска повезаност* графа G. Ова сопствена вредност је већа од 0 ако и само ако је G повезан граф. Ово је последица чињенице да број појављивања вредности 0 као сопствене вредности Лапласове матрице представља број повезаних компоненти графа. Величина ове сопствене вредности одређује колико је граф заиста повезан. Сопствени вектор придружен алгебарској повезаности се назива *Фидлеров вектор*.

Теорема 6.11. (видети [87]) Нека је G граф, и нека је G_1 добијен из графа G брисањем k чворова из G и свих њихових грана. Тада:

$$\mu_2(G_1) \geq \mu_2(G) - k.$$

Као што се може приметити, истом графу могу да одговарају различите матрице суседства, јер различите нумерације чворова доводе до различитих матрица. Насупрот томе постоје карактеристике графа које су инваријантне у односу на пренумерацију чворова. Таква важна инваријанта јесте спектар графа.

Оно чему желимо да посветимо посебну пажњу у овој докторској дисертацији су сопствене вредности и сопствени вектори графа који се користе у позадини софтверске платформе. Одређене сопствене вредности из спектара неких графова, као и њима одговарајући сопствени вектори носе доста информација о самој структури графа и управо то нас је мотивисало да их употребимо у алгоритму који треба да препозна којем репозиторијуму (граф) који скуп података (чвор) припада и да међу њима изабере оне најрелевантније за постављено питање. На основу тих вредности SpecINT доноси одлуке за креирање кориснички употребљивог SPARQL упита. Пре него што прикажемо на који начин се одређене сопствене вредности користе, поменућемо неке значајне сопствене вредности и сопствене векторе графа, као и њихове примене.

6.3.2 Значајне сопствене вредности графа

Неке сопствене вредности графа су важни параметри графа, па су самим тим интересантни и у применама. У овом делу текста биће нешто више речи о највећој A-сопственој вредности, другој најмањој L-сопственој вредност и другој највећој A-сопственој вредности. Више детаља о овим, али и другим значајним сопственим вредностима се могу наћи у раду [88].

Највећа A-сопствена вредност

Највећа сопствена вредност или индекс графа (познат и као спектрални радијус) је веома важан математички параметар који је приказан у прегледаном чланку [89]. На основу теореме 1.12 у [83], индекс графа је једнак одређеној врсти средње вредности степена чворова, такозваној динамичкој средњој вредности, која узима у обзир не само директне суседе чворова већ и суседе суседа итд. Највећа сопствена вредност λ_1 матрице суседства игра важну улогу у моделовању процеса ширења вируса у рачунарским мрежама. Што је мања највећа сопствена вредност, већа је робусност мреже у односу на ширење вируса. У раду Y. Wanga [90] је показано да је *праг епидемије* у ширењу вируса пропорционалан са $\frac{1}{\lambda_1}$. Други модел ширења вируса је развијен од стране P. Van Mieghema и осталих [91] са истим закључком који се односи на $\frac{1}{\lambda_1}$.

У радовима M. D. Königa et al. [92, 93] истраживање и развој мрежа се проучавају коришћењем индекса матрице суседства. У таквим мрежама пожељно је да се знање шири кроз мрежу што је више могуће. Према томе, тенденција је да се добије што већа вредност индекса, што је у супротности са ситуацијом код ширења вируса.

Алгебарска повезаност

Друга најмања Лапласова сопствена вредност графа G, у ознаци $\mu_2(G)$, се назива алгебарском повезаношћу графа и први пут је уведена у раду [87]. Значај ове сопствене вредности произилази из чињенице да се обе "графичке" мере повезаности графа *vertex connectivity* k и *edge connectivity* k', у многим ситуацијама не карактеришу као одговарајуће мере. На пример, обе су једнаке 1 за сва стабла, док је алгебарска повезаност највећа за звезду (такође једнака 1), а најмања за пут. Такође је познато (на основу теореме о укљештењу) да било који подграф повезаног графа настао брисањем произвољне гране има алгебарску повезаност која није већа од алгебарске повезаности полазног графа; тада је, као што је очекивано, присутна особина монотоности. Додатно, постоје многе друге неједнакости које повезују алгебарску повезаност са k и k'. Наиме, важе следеће неједнакости

$$2k'(G)(1-\cos\frac{\pi}{n}) \leq \mu_2(G) \leq k(G) \leq k'(G) \leq \delta(G)$$

где је G граф са n чворова и минималним степеном чвора δ (за више детаља погледати [94]). За још додатних информација о алгебарској повезаности и границама за њену вредност погледати [95, 96, 97].

Алегабарска повезаност се доста често користи за решавање *max-cut* и *min-cut* проблема одвајања, као и *bipartition width* (сви су NP-тешки). Рез (енг. *cut*) би-партиције

$S \cup \overline{S}$ скупа чворова тежинског графа $G = (V, E, w)$ се дефинише као $cut(S, \overline{S}) = \sum_{s \in S} \sum_{t \in \overline{S}} w(s, t)$. Прва два проблема су повезана са проналажењем реза максималне и минималне величине, док се трећи проблем представља као *min-cut* у којем је бипартиција парна (тако да се кардиналности скупова S и $\overline{S}$ разликују највише за 1). Кардиналности одговарајућих скупова S доводе до инваријанти графа (тј. мере одвајања), при чему треба напоменути да постоје многа ограничења над овим инваријантама у литератури која су повезана са алгебарском повезаношћу (7.51 у [85]). *Min-cut* проблем се може третирати сопственим вектором који одговара $\mu_2(G)$ (видети Подсекцију 6.3.3).

Друга највећа сопствена вредност

Друга највећа сопствена вредност r-регуларних графова је повезана са алгебарском повезаношћу ($\lambda_2 + \mu_2 = r$). Због тога је за регуларне графове проблем максимизације алгебарске повезаности еквивалентан минимизацији друге највеће сопствене вредности.

Добри експендери се могу лако конструисати из графова са по модулу малом другом највећом сопственом вредношћу. Ова класа графова обухвата и такозване *Ramanujan* графове. За више детаља о применама *Ramanujan* графова видети [89].

6.3.3 Сопствени вектори графа

Сопствени вектори графа такође садрже доста информација о структури графа. Ипак, неки указују да сопствени вектори нису инваријанте графа пошто зависе од ознака чворова. На другој страни, ово може бити предност, нарочито ако се говори о наменском означавању чворова. У овом делу говоримо о главном (енг. *principal*) и Фиедлеровом сопственом вектору, док ће се проблеми који су повезани са истовременим коришћењем неколико сопствених вектора бити описани у Подсекцији 7.4.1. Други значајни сопствени вектори и њихове примене су дате у прегледном чланку [88].

Главни сопствени вектор

Нормализовани позитивни сопствени вектор који припада највећој A-сопственој вредности (индексу) повезаног графа се назива *главним сопственим вектором*. Проблем рангирања појединаца и предмета уз помоћ сопствених вектора погодно изабраних матрица графа је стар у математичкој литератури. Основне референце везане за овај проблем се могу видети у докторској дисертацији [98]. Ове методе се често могу видети у области социологије већ дуже време.

Теорема 6.12. Нека је $N_k(i)$ број шетњи дужине k које почину у чвору i небипартитног графа G са чворовима $1, 2, \ldots, n$. Нека је $s_k(i) = N_k(i) \cdot (\sum_{j=1}^{n} N_k(j))^{-1}$. Тада, за $k \to \infty$, вектор $(s_k(1), s_k(2), \ldots, s_k(n))^T$ тежи ка сопственом вектору који одговара индексу графа G.

Према томе, рангирање чворова на основу координата главних сопствених вектора се своди на њихово рангирање према броју шетњи. Број шетњи $N_k(i)$ се може интепретирати и као "утицај" или "значај" чвора i и често се у литератури може пронаћи назив *централност чвора* i.

Алати за претрагу на Интернету се базирају на сопственим векторима матрице суседства и другим матрицама графа. Најпознатији системи за претрагу су *PageRank* који су развили Sergey Brin и Larry Page [81] (користи се у Гоогле-у) и *Hyperlinked Induced Topics Search* (HITS) који је развио Jon Kleinberg [99]. Чланак [100] садржи преглед обе технике. У овом контексту структура Интернета се може представити као дограф G у којем веб стране одговарају чворовима, а линкови између страна (*hyperlinks*) оријентисаним гранама графа.

HITS користи сопствене векторе који одговарају највећој сопственој вредности симетричних матрица AA^T и A^TA, где је A матрица суседства подграфа графа G који је индукован скупом веб страна које су добијене на основу кључних речи претраге и коришћењем неких хеуристика. Добијени сопствени вектори дефинишу одређени редослед изабраних веб страна.

PageRank алгоритам користи сличне идеје које у основи користе *случајне шетње* (енг. *random walks*). Уствари матрица суседства графа G је нормализована тако да се заправо користи матрица $P = D_+^{-1}A$. (На основу ових трансформација, да би се елиминисали нула-редови у P, свим чворовима без излазећих грана се додају гране које иду ка свим осталим чворовима. Додатно, да би се обезбедило да је матрица примитивна, вештачки је формиран најмање један непаран цикл уколико такав не постоји). Даље се формира конвексна комбинација $\bar{P}$ од матрице P и ранг матрице постаје један. Матрица $\bar{P}$ је транзициона матрица Марковљевог ланца и нормализовани сопствени вектор највеће сопствене вредности транспоноване матрице $\bar{P}$ се дефинише као стабилно стање ланца. Веб стране се рангирају према координатама сопственог вектора. Математичка позадина PageRank алгоритма је у више детаља описана у Подсекцији 7.4.2.

Фидлеров сопствени вектор

Сопствени вектор који одговара другој најмањој Лапласовој сопственој вредности повезаног графа се назива Фидлеров сопствени вектор. Овај сопствени вектор се користи као део хеуристике за решавање *min-cut* проблема где је потребно извршити поделу скупа чворова у делове који одговарају позитивним и негативним координатама овог вектора [101]. Показано је да се корисна подела скупа чворова графа може базирати на знаку координата сопствених вектора, мислећи се пре свега на "доње сопствене векторе" Лапласове матрице (али и "горње сопствене векторе" матрице суседства графа). Имајући у виду ове примедбе, у наставку ће бити описане две хеуристике за партиционисање графа:

(1) *Рекурзивна сйектрална бисекција:* користи се Фидлеров сопствени вектор да би се поделили чворови графа у два дела на основу знака координата, а затим се наставља са истом процедуром задовољавајући задати критеријум оптималности (видети [102, 103]).

(2) *Итеративна сйектрална бисекција:* почиње се као у делу под (1), с'тим што се надаље користе трећа најмања, четврта најмања, ... сопствена вредност за обраду партиција у међуфази, све док се не зауставимо због одређеног критеријума.

Ове идеје су доста коришћене у литератури на различите начине за осмишљавање нових хеуристика за спектрално партиционисање или кластеровање графа. На пример,

Shi и Malik [104] су показали како се знаци координата Фидлеровог сопственог вектора могу искористити за одвајање предње од задње структуре у сликама. Оригинална процедура из [101] је даље побољшана коришћењем матрице $D^{-1}L$ (у циљу максимизације нормализованог реза графа). Генерално, процес сегментације слика, где је циљ поделити слику на регионе на основу одређених критеријума, је од великог значаја у компјутерској визији (енг. *computer vision*) и препознавању шаблона (енг. *pattern recognition*). Веома често се за сегментацију слика користе управо сопствени вектори неких матрица графа (за више детаља видети [105]).

Познато да у регуларним графовима сопствена вредност λ_2 одговара алгебарској повезаности μ_2 са истим (Фидлеровим) сопственим вектором. Ипак, сопствени вектор z сопствене вредности λ_2 показује сличне особине и у нерегуларним графовима. На пример, у [106], аутори наводе пример случајног графа са 600 чворова, где знаци координата сопственог вектора од λ_2 доводе до бисекције високог квалитета, без јасног теоријског објашњења. У овој ситуацији, следећа Фидлерова теорема постоје релевантна (видети [84, 107]): наиме, подграфови индуковани чворовима који одговарају негативним или позитивним координатама од z, су повезани (видети [108]).

6.4 Прилози кластеровању одређених класа графова

Нека су K_n и K_m комплетни графови са n и m чворова редом, и нека је $G = K_n uv K_m$ граф добијен из графа $K_n \dot\cup K_m$ додавањем гране која повезује чвор u од K_n са чвором v од K_m. Означићемо овакав граф са $E_{n,m}$.

Нека је $C = \{(v_n, v_{n+1})\}$ један рез у графу $G = E_{n,m}$. Блокови реза C су индуковани чворовима графова K_n и K_m. Подесном нумерацијом, означимо чворове једног C-блока са $v_1, v_2, \ldots, v_n$, чворове другог C-блока са $v_{n+1}, v_{n+2}, \ldots, v_{n+m}$, при чему ћемо сматрати да су са v_n и v_{n+1} обележени артикулациони чворови. На почетку ћемо показати да важи следећа теорема:

Теорема 6.13. Нека је $w = (w_i)$ Фидлеров вектор графа $G = E_{n,m}$. Чворови који припадају скупу $N(w)$ су у једном C-блоку, док су чворови који припадају скупу $P(w)$ у другом блоку пресека C.

Доказ. Из Лапласове матрице графа G добијамо следећи систем једначина:

$$
\begin{array}{rrrrrrrrrrrrr}
(n-1)w_1 & & -w_2 & - & \ldots & & -w_{n-1} & -w_n & & & = \mu w_1 \\
-w_1 & +(n-1)w_2 & - & & & & -w_{n-1} & -w_n & & & = \mu w_2 \\
\vdots & & & & & & & & & & \\
-w_1 & & -w_2 & - & \ldots & +(n-1)w_{n-1} & & -w_n & & & = \mu w_{n-1} \\
-w_1 & & -w_2 & - & & & -w_{n-1} & +nw_n & -w_{n+1} & & = \mu w_n \\
\\
-w_n & +mw_{n+1} & & -w_{n+2} & - & \ldots & & & -w_{n+m} & & = \mu w_{n+1} \\
& -w_{n+1} & +(m-1)w_{n+2} & - & & & & & -w_{n+m} & & = \mu w_{n+2} \\
& \vdots & & & & & & & & & \\
& -w_{n+1} & & -w_{n+2} & - & & & & -w_{n+m} & & = \mu w_{n+m-1} \\
& -w_{n+1} & & -w_{n+2} & - & \ldots & +(m-1)w_{n+m} & & & & = \mu w_{n+m}
\end{array}
\tag{6.4}
$$

За $\mu = \mu_2$ и сопствени вектор $w = (w_i)$ који одговара сопственој вредности μ_2, из првих $n-1$ једначина добијамо да важи

$$w_1 = w_2 = \ldots = w_{n-1} \tag{6.5}$$

Користећи прву једнакост из (6.4) и резултат (6.5) добија се $(1-\mu)w_1 = w_n$. Брисањем једног артикулационог чвора из графа G добије се неповезани граф са две компоненте повезаности, при чему на основу Теореме 6.11 важи $1 - \mu > 0$. Одавде закључујемо да су сви чворови из првог C-блока истог знака. Аналогно добијамо да су сви чворови из другог C-блока истог знака.

Из услова ортогоналности вектора $w = (w_i)$ и $e = (1, 1, ..., 1)$ се може показати да важи $(n - \mu)w_1 = -(m - \mu)w_{n+2}$. Како на основу Теореме 6.11 важи да је $n - \mu > 0$ и $m - \mu > 0$, тада добијамо да су координате w_1 и w_{n+2} супротног знака, односно да чворови из различитих C-блокова припадају различитим скуповима, $N(w)$ и $P(w)$. ■

Процес слепљивања (енг. *coalescence*) два графа G_1 и G_2 започиње избором два произвољна чвора, v_1 у графу G_1 и чвора v_2 у графу G_2, над којима се врши манипулација. Тада, слепљивање графова, у ознаци $G_1 \cdot G_2$, се састоји од чворова и грана индивудуалних графова, изузев што се два изабрана чвора v_1 и v_2 претварају у један чвор v који је суседан са свим чворовима у G_1 који су суседни чвору v_1 и сваким чвором из G_2 суседним чвору v_2.

Теорема 6.14. (видети [85]) Нека је $G \cdot H$ слепљивање у којем је чвор v_1 из G_1 идентификован са чвором v_2 из G_2. Тада важи $P_{G\cdot H}(x) = P_G(x)P_{H-v}(x) + P_{G-u}(x)P_H(x) - xP_{G-u}(x)P_{H-v}(x)$.

Нека су K_n и K_m комплетни графови са n и m чворова редом, и нека је граф $G = K_n \cdot K_m$ њихово слепљивање. Означићемо овако добијени граф са $V_{n,m}$. Брисањем артикулационог чвора графа $G = K_n \cdot K_m$, добијамо неповезани граф са две компоненте. Чворови из једне компоненте графа заједно са чвором v_n индукују један блок у графу G. Подесном нумерацијом, означимо чворове једног блока са $v_1, v_2, \ldots, v_n$, чворове другог C-блока са $v_n, v_{n+1}, \ldots, v_{n+m}$, при чему ћемо сматрати да је са v_n означен артикулациони чвор.

Теорема 6.15. Нека је $w = (w_i)$ Фидлеров вектор графа $G = V_{n,m}$. Чворови који припадају скупу $N(w)$ су у једном блоку, док су чворови из $P(w)$ у другом блоку графа G, изузев артикулационог чвора v_n који припада скупу $O(w)$.

Доказ. Од раније је познато да важи $z_2(G) = z_n(\overline{G})$ (видети Теорему 6.10), где су $z_2(G)$ и $z_n(\overline{G})$ сопствени вектори који одговарају сопственим вредностима $\mu_2(G)$ і $\mu_n(\overline{G})$, редом.

Уместо да одредимо сопствени вектор који одговара другој најмањој сопственој вредности μ_2 графа G, одредићемо сопствени вектор који одговара сопственој вредности μ_n графа $\overline{G} = K_{n-1,m-1} \cup \{v_n\}$.

Пошто граф $\overline{G}$ има изловани чвор v_n, тада можемо израчунати сопствени вектор подграфа $H = K_{n-1,m-1}$ за μ_n, и након тога сопственом вектору додати координату са вредношћу 0 на n-то место. Даље, уместо одређивања сопственог вектора подграфа $H = K_{n-1,m-1}$ за μ_n, можемо одредити сопствени вектор графа $\overline{H} = K_{n-1} \cup K_{m-1}$ за μ_2.

За граф $\overline{H}$ важи да је $\mu_1 = \mu_2 = 0$. Сопствени вектори графова K_{n-1} и K_{m-1} који одговарају сопственој вредности 0 су $e(K_{n-1}) = (\underbrace{1, 1, ..., 1}_{\text{n-1}})$ и $e(K_{m-1}) = (\underbrace{1, 1, ..., 1}_{\text{m-1}})$. Из последњег закључујемо да вектори $y_1 = (\underbrace{1, 1, ..., 1}_{\text{n-1}}, \underbrace{0, 0, ..., 0}_{\text{m-1}})$ и $y_2 = (\underbrace{0, 0, ..., 0}_{\text{n-1}}, \underbrace{1, 1, ..., 1}_{\text{m-1}})$ чине базу потпростора $\varepsilon_{\overline{H}}(0)$.

Тада је вектор $x_2(\overline{H})$ је линеарна комбинација ова два вектора. Вектори $x_2(\overline{H})$ и $e(\overline{H})$ су ортогонални из чега следи да је $\alpha(n-1) + \beta(m-1) = 0$, одакле закључујемо да су α и β скалари са различитим знаковима.

$$
\begin{aligned}
& x_2(\overline{H}) = x_n(H) \\
\Rightarrow \quad & x_n(H) = (\underbrace{\alpha, \alpha, \ldots, \alpha}_{\text{n-1}}, \underbrace{\beta, \beta, \ldots, \beta}_{\text{m-1}}) \\
\Rightarrow \quad & x_n(\overline{G}) = (\underbrace{\alpha, \alpha, \ldots, \alpha}_{n-1}, 0, \underbrace{\beta, \beta, \ldots, \beta}_{\text{m-1}}), \text{ jer je } \overline{G} = HUK_1 \\
\Rightarrow \quad & x_2(G) = w = (\underbrace{\alpha, \alpha, \ldots, \alpha}_{\text{n-1}}, 0, \underbrace{\beta, \beta, \ldots, \beta}_{\text{m-1}})
\end{aligned}
$$

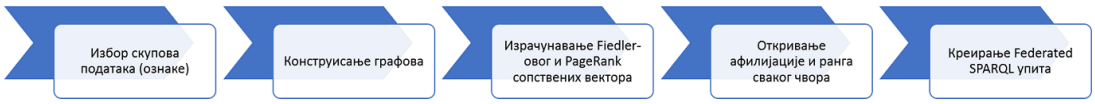
Пошто су α и β различитог знака, тада се може закључити да чворови из два блока графа G без v_n припадају различитим скуповима $N(w)$ и $P(w)$, док је $v_n \in O(w)$. ■

Глава 7

SpecINT архитектура

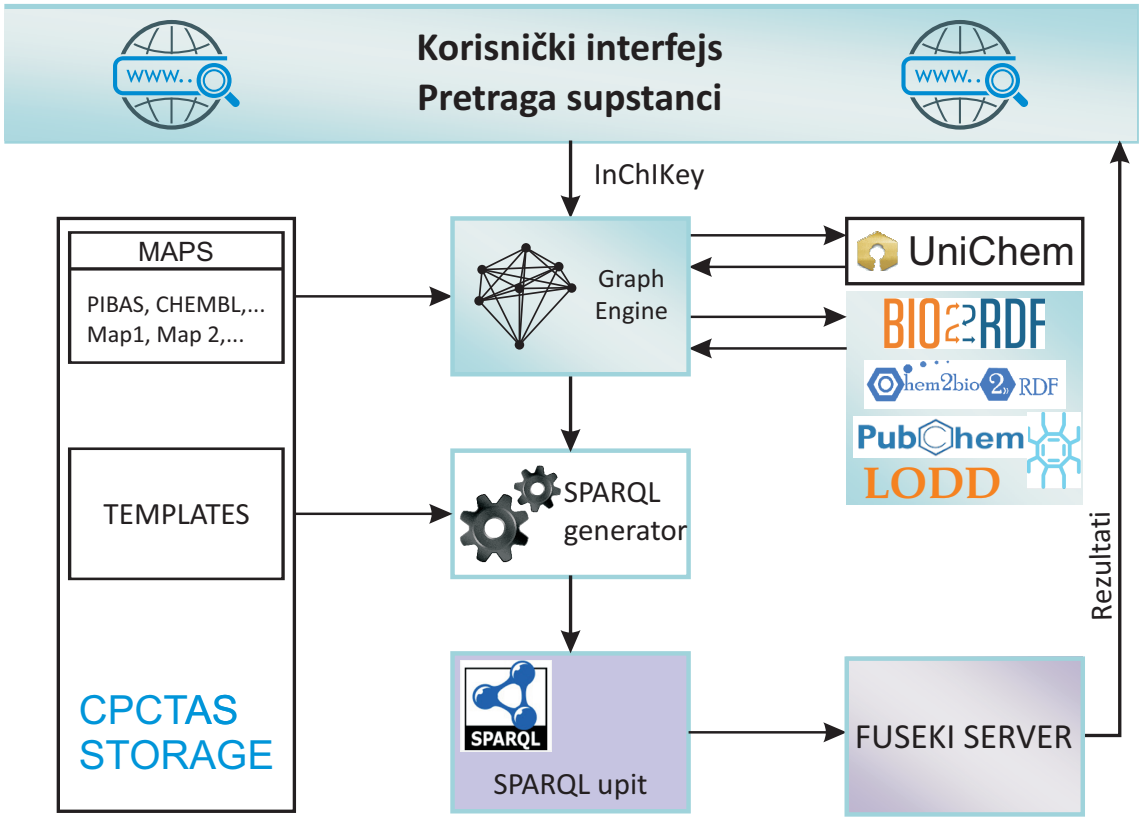
Константна експанзија нових скупова података доноси нове проблеме у анализи и претраживању неповезаних и хетерогених података, што је круцијално за њихово успешно и сврсисходном експлоатисање и добијање нових знања из њих. Захваљујући стандардима Семантичког Веба и „онлајн" претрази података преко ендпоинта, могуће је спровести претраживање ових података помоћу једног SPARQL упита. Процес интерграције података и добијање новог знања из њих је само по себи изазов, имајући у виду како су нови и постојећи скупови података повезани унутар репозиторијума. На почетку ове дисертације је описан проблем са којим се суочава научна заједница и због чега су јој информације на глобалном нивоу битне, па је самим тим и дата мотивација за развој једне овакве платформе. Кроз претходна поглавља је описан начин на који се подаци чувају и презентују на Вебу, као и проблеми које такво представљање доноси. Такође су приказана и тренутна решења која постоје, објашњено је зашто имплементације које се ослањају на статистику скупа не одговарају потребама истраживача и које су модификације неопходне како би се добила фунцкионалнија решења за претраживање једног репозиторијума. Међутим, и даље таква решења не могу да претражују више репозиторијума истовремено. Након описа *state-of-the-art* решења дате су математичке основе на којима се заснива Платформа, тачније процеси интеграције и претраживања репозиторијума. С обзиром да су сва неопходна знања за разумевање Платформе дата у претходним поглављима, у овом делу текста биће дат њен општији опис, а детаљна објашњења кроз наредне подсекције.

Кроз један пример биће објашњено, корак по корак, на који функционише Платформа, од тренутка када је прослеђен InChIKey хемијске структуре као улаз, па до тренутка генерисања Federated SPARQL упита и добијања резултата након његовог извршавања. На Слици 7.1 је приказан процес рада Платформе, где се могу уочити пет основних корака који се одвијају у њој: избор скупова који ће се разматрати, конструисање графова са гранама између њих, израчунавање одговарајућих сопствених вектора графа и на крају, креирање упита који треба да да одговор на постављено питање. Према овим корацима биће формиране наредне секције које ће сваки корак да објасне детаљно. Иначе, SpecINT платформа је смештена на серверу IBM производње који има 8 GB RAM меморије, и *Intel Xeon Processor E5620* израђен у *Nehalem-C* технологији са *12M Cache, 2.40 GHz, 5.86 GT/s Intel QPI*. Програмски језик који је коришћен за развој је *Python*, верзија 2.7. Док је за развој онологија коришћен *Protégé editor*.



Слика 7.1: SpecINT процес рада.

На Слици 7.2, где је приказана архитектура софтверске платформе, се може уочити део који представља кориснички интерфејс преко којег корисник прослеђује InChIKey одговарајуће хемијске структуре Платформи. Овај улаз се прослеђује модулу *Graph Engine* који на основу њега треба да конструише један граф и један дограф, користећи најпре *UniChem* као сервис, а затим и информације из репозиторијума (десни део слике). Улога неоријентисаног графа је да чува информације о афилиацији (припадности) скупова података у репозиторијумима, док оријентисани граф чува информације о начину на којем су супстанце међусобно повезане. На слици се може видети и део који садржи мапе - једноставан начин да се мала лабораторија прикључи неком репозиторијуму или иницијативи као што је то објашњено у Поглављу 4.



Слика 7.2: Архитектура SpecINT софтверске платформе.

Када су графови конструисани, они се даље користе као улаз у следећи модул - *SPARQL generator*, који на основу прослеђених информација треба да конструише валидан Federated SPARQL упит за питање које је постављено. У оквиру модула се доносе одлуке везане за афилијацију и избор најрелевантнијих скупова података у датом тренутку,

а затим се узимају одговарајући подупити (Секција 3.3). Пошто је одговарајући упит конструисан, он се извршава, а резултати за хемијску структуру која се претражује приказују кориснику.

Још једном да напоменемо да је овде дат груб опис функционисања софтверске платформе (енг. *workflow*), а детаљи везани за конструисање графова, избор најрелевантнијих скупова података и њихово повезивање и уклапање са подупитима су објашњени у наставку.

7.1 Јединствени идентификатори хемијских структура као улаз у Платформу

Претрага података о молекуларним структурама из база података и RDF скупова података се најбоље постиже коришћењем јединственог идентификатора. *SMILES* стандард представља поједностављену молекулску спецификацију улазних линијских података (енг. *simplified molecular input line entry specification*) која се користи за недвосмислено описивање структуре хемијских молекула користећи кратке ASCII стрингове. Неодостатак овакве спецификације је могућност приказа више различитих SMILE приказа (кодова) за једну исту супстанцу. У међувремену су развијени алгоритми који омогућавају да се исти SMILE генерише за молекул независно од редоследа атома у структури. Међутим, иако је SMILE за супстанцу јединствен, он зависи од начина на који је алгоритам изабрао редослед атома у току њиховог генерисања, па отуда назив канонички SMILES. Оно што канонички SMILES треба да обезбеде је индексирање и јединствености молекула у базама података. Други стандард који је добио значајну улогу као јединствени идентификатор и све се више користи да учини ресурсе и литературу доступном на Вебу [109] је *International Chemical Identifier (InChI)*, развијен од стране IUPAC-а (*International Union of Pure and Applied Chemistry*). С обзиром да се хемијске структуре на Вебу углавном представљају сликама и да InChI није погодан за брзу претрагу, дизајниран је InChIKey стандард - хеширана верзија InChI стандарда. InChIKey се састоји од 27 карактера и доста је погодан за брзу претрагу на Вебу. Репозиторијуми на Вебу (као и сви они представљени у Секцији 3.2) користе све три спецификације када чувају податке о хемијским структурама, а приметан је и тренд све већег коришћења InChIKey спецификације због једноставнијег записа и брже претраге. За улаз у софтверску платформу која је представљена у овој докторској дисертацији је изабран InChIKey, али се платформа може лако прилагодити и за друге стандарде. На Слици 7.3 се може видети како изгледа један део корисничког интерфејса.

7.2 Иницијализација скупова података

У овој секцији је описан процес избора иницијалних скупова података који потенцијало могу да садрже податке који су нам потребни. У каснијим корацима ови резултати ће бити додатно филтрирани. Генератор упита треба пажљиво да одреди који су скупови добри за упите, пошто погрешан избор или води ка скупој комуникацији са великим бројем међурезултата који се памте или систем не може да врати резултате.

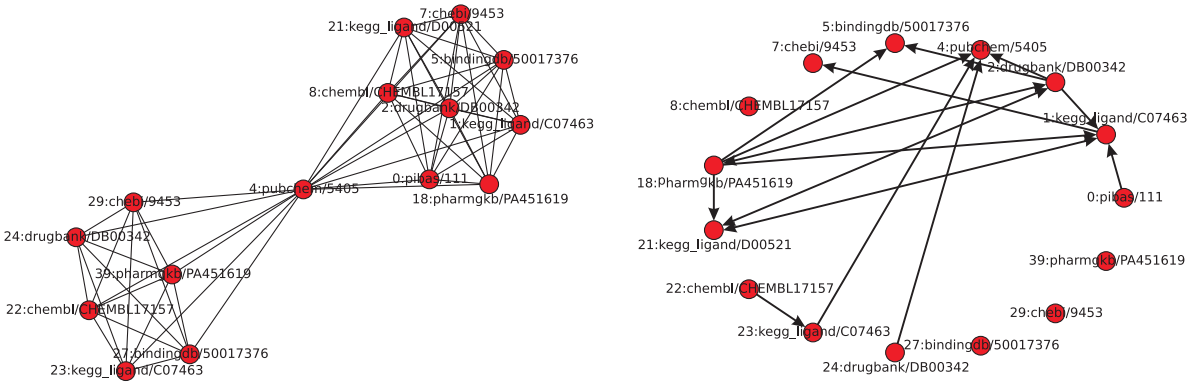
Слика 7.3: SpecINT интерфејс.

Најпрактичнији начин да се повежу два скупа података јесте да се користе најважнији појмови око којих се групишу остали појмови и подаци. На пример, све информације о лековима из скупа KEGG могу да се повежу са лековима из скупа DrugBank коришћењем *owl:sameAs* релације која је идентификујући линк који повезује два ентитета са истим InChIKey идентификатором. Да би се сакупиле све информације о одређеној хемијској структури, она најпре мора бити трансформисана у одговарајући InChIKey идентификатор. Тада се за добијање листе свих синонима тражене супстанце (без одговарајућих URI-ја) позива UniChem [110] API који је развијен у европском биоинформатичком институту (*European Bioinformatics Institute (EBI)*). UniChem је бесплатно доступан сервис који у себи садржи мапирања за мале молекуле и који се базира на прихваћеним и стабилним стандардима, InChIs и InChIKeys. Још прецизније, доступни синоними представљају ознаке исте супстанце која припада различитим скуповима података. На пример, за *InChIKey = GUGOEEXESWIERI-UHFFFAOYSAN*, неке од добијених синонима су: *CHEMBL17157*, *kegg_ligand C07463*, *drugbank DB00342*, *chebi9453*, *SCHEMBL5152* итд. Након тога се сви добијени синоними за једну супстанцу користе за креирање чворова графа, што ће детаљније бити описано у наредној секцији.

7.3 Конструисање графова

Узимајући у обзир да SPARQL упити проистичу из оријентисаних графова, прво је конструисан неоријентисан граф из добијених синонима (користе се као ознаке за чворове), а затим је, пратећи везе унутар сваког репозиторијума, из њега добијен диграф. Овај корак одређује начин на који ће репозиторијуми бити повезани и додатно филтрирани. Пратећи одређене путање у графу, редослед подупита је одређен тако да претраживање може да буде урађено без заједничке онтологије између ресурса. Ако су ове путање погрешне, упит неће бити у стању да повеже узастопне подупите и ау-

томатски упит неће бити валидан. Дакле, ова процедура укључује конструисање два графа (видети Слику 7.4) који имају одређену сврху: а) неоријентисаног графа $V_{n,m}$ (лева страна) и б) оријентисаног графа $D_{n,m}$ (десна страна). Сви кораци у добијању графова су приказани у Процедури 1.



Слика 7.4: Повезивање графова који одговарају Chem2Bio2RDF и Bio2RDF репозиторијумима.

А. *Конструисање неоријентисаноī ɪграфа.* Овај корак открива нашу скривену намеру да на неки начин сачувамо информацију о припадности чворова репозиторијуму и повезујући чвор између репозиторијума, пошто ће наредне промене у графу и уклањање грана могу да измене њихове афилиације. Афилиације репозиторијума (URIs) су јако важне у овој фази, јер иста супстанца у оквиру различитих репозиторијума има различите предикате и оријентацију грана. Ови URI-ји могу лако бити добијени из URI шаблона који припада релевантном репозиторијуму, али они су изостављени ради јасноће слике.

Сви синоними пронађени у претходном кораку су искоришћени за означавање чворова комплетних графова K_n и K_m (за сваки репозиторијум), пошто сви синоними одговарају истој супстанци, али из различитих скупова података. Графови у општем случају не морају да буду компплетни, о чему ће више речи у дискусији. На овај начин се повезују све репрезентације исте супстанце у оквиру једног репозиторијума. Пратећи идеју о чувању афилиација чворова, у наредном кораку се спроводи слепљивање два добијена графа у граф $V_{n,m} = K_n \cdot K_m$, са било којим чвором чија се ознака налази у скупу заједничких ознака чворова два графа. Изабрани чвор биће коришћен као ”мост” за прелазак са једног на други репозиторијум. Према резултатима Теореме 6.15 (видети Секцију 6.4), за граф $V_{n,m}$ се може одредити тзв. Фидлеров сопствени вектор s на основу којег се скуп чворова графа може поделити у два дисјунктна скупа са позитивним и негативним чворовима (знак одговарајуће координате у сопственом вектору), при чему координата чија је вредност 0 одговара $null$-чвору.

За лакше разумевање примера, поред ознака скупова, сви чворови такође садрже и бројеве који узимају вредности од 0 до $|V_{n,m}|$, где је $|V_{n,m}|$ број чворова графа. Ови бројеви представљају редни број координата у сопственим векторима графа. У нашем примеру, ознаци 0 на чвору одговара прва координата у вектору, ознака 1 другој координати и слично. Након елиминисања неповезаних чворова тј. оних чворова који не садрже информације о траженој хемијској структури, неки од бројева који се користе као део ознаке чворова се губе.

B. *Конструисање дијаграфа.* За брже и успешније добијање резултата неопходно је изабрати најподесније секвенце скупова података. Пре генерисања упита, Платформа мора да провери постојање и оријентацију грана између чворова. Међусобне везе између скупова података је могуће открити претрагом одређених кључних речи које се налазе као подстрингови назива везе између две супстанце. Са граном $(source,\ target)$ добијеном из триплета $(?source,\ ?property,\ ?target)$ граф $V_{n,m}$ се може трансформисати у диграф $D_{n,m}$ што одговара природи SPARQL упита. Овај корак укључује брисање свих непостојећих грана, али не и изолованих чворова пошто је Фидлеров вектор претходно одређен са свим чворовима графа, а дужине свих вектора са којима радимо треба да имају исту дужину. Међутим, ови чворови могу и да се обришу (Слика 7.4), јер свакако не утичу на одређивање значајности осталих чворова. Са Фидлеровим вектором смо сачували одређене информације о графу пре самог процеса трансформисања графа. Када је диграф $D_{n,m}$ (без изолованих чворова) неповезан, целокупна процедура се понавља. За сваку супстанцу се добију различити графови $V_{n,m}$ и $D_{n,m}$, што умногоме зависи од начина формирања репозиторијума, али и од доступности ендпоинта у датом тренутку.

Добро позната је чињеница да је важност сваког чвора графа пропорционална суми важности сваког чвора који показује ка њему. Прост рачун каже се ово може третирати као проблем сопствене вредности и сопственог вектора (за више детаља погледати [81]), што ће и бити показано касније. Сада, за оријентисани граф $D_{n,m}$ можемо да одредимо ненегативни сопствени вектор r (ранг), чије координате одређују релативни значај сваког од чворова. Када добијемо сопствени вектор, најзначајнији чвор је онај са највећом вредношћу координате у сопственом вектору, следећи најзначајнији са другом највећом вредношћу и тако даље. Сада се може пратити највероватнија путања која ће прећи преко репозиторијума, водећи рачуна о томе којем репозиторијуму који чвор припада, управо како то раде машине за претрагу.

Процедура 1: КОНСТРУИСАЊЕ ГРАФА И ДИГРАФА.

Улаз: InChIKey, репозиторијуми R_1 и R_2
Резултат: Диграф $D_{n,m}$, сопствени вектори s и r

1 $Intersection$:={$R_1 \cap R_2$} ; `// заједнички скупови података`
2 $UniChem$:= {UniChem синоними за InChIKey};
3 $firstGraph$:= {синоними из R_1 за сваку $UniChem$ ознаку (label)};
4 $secondGraph$:= {синоними из R_2 за сваку $UniChem$ ознаку};
5 Конструисање комплетних графова K_n са ознакама из $firstGraph$ и K_m са ознакама из $secondGraph$;
6 Направити „слепљивање" $V_{n,m}$ са било којом ознаком из скупа $Intersection$ и израчунати сопствени вектор s;
7 Трансформација графа $V_{n,m}$ у диграф $D_{n,m}$;
8 Брисање непостојећих грана и фаворизација одређених чворова;
9 **if** *$D_{n,m}$ је неповезан* **then**
10 иди на корак 6 и употреби неискоришћену ознаку из $Intersection$;
11 **end**
12 Израчунати сопствени вектор r од $D_{n,m}$;

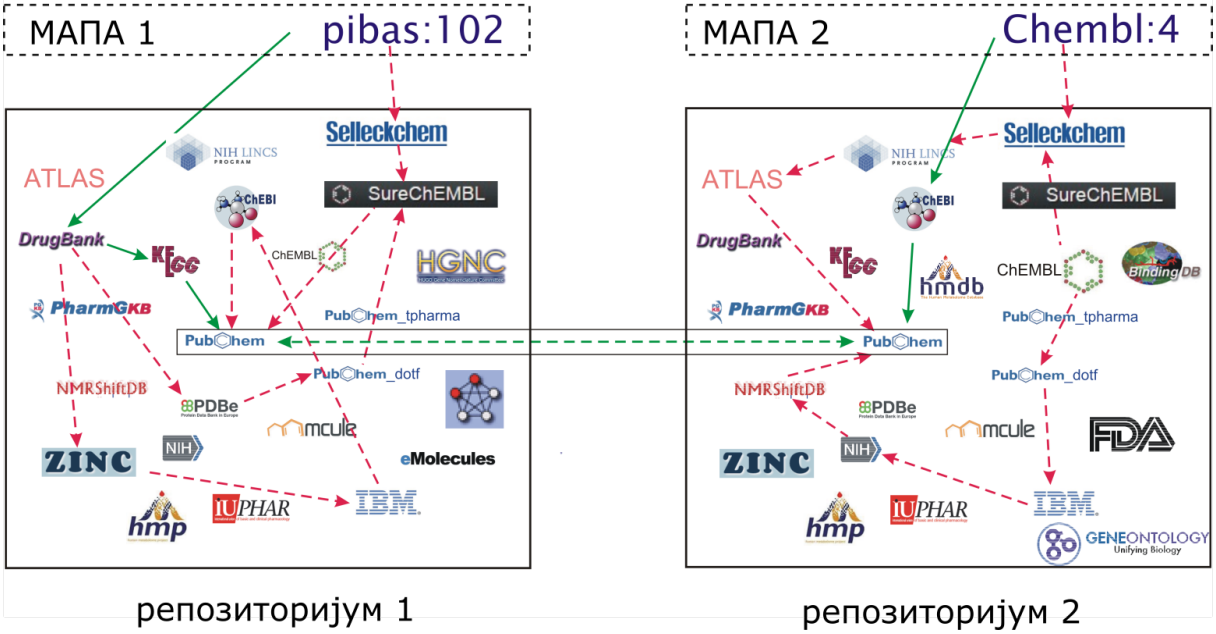
Међутим, процедура не може да гарантује да су сви одговарајући скупови података обухваћени. У случају када је нови скуп података интегрисан (мала вредност ранга) или се преферирају специфични одговори (они који одговарају постављеном питању), неопходно је осмислити начин како да управо ти скупови буду узети у разматрање. На пример, ако су таргети лекова у фокусу, одређени скупови података који носе такве информације морају бити фаворизовани за добијање бољих резултата. За потребе избора најрелевантнијих скупова података за постављено питање, испробане су три хеуристике које могу да утичу на вредности координата вектор r. Ове хеуристике се базирају на одређеним особинама графа и садржају онтологије која је развијена за ову намену и која садржи информације о сваком скуп података посебно. Тестиране хеуристике и сви резултати биће презентовани у секцији са резултатима.

7.4 Одређивање афилиација чворова

Ова секција је намењена алгоритиму за конструисање SPARQL упита на основу координата два вектора s и r и већ креираних подупита које ћемо у наставку звати обрасцима. Овде ћемо објаснити на који начин алгоритам прати припадност чвора репозиторијуму и по којем принципу се врши рећање најрелевантнијих скупова података у низ да би се осигурали резултати са репозиторијума. Такође, у овој фази се одређују и префикси који одговарају скуповима података и њима одговарајући обрасци. Нулти чвор се третира као неутрални чвор (мост) за RDF скуп података и може да се користи и као субјекат и као објекат у шаблону графа. Коначно, када је идејна нит представљена, чворови са највећим вредностима ранга биће искоришћени за тражење најбоље путање до централног чвора са обе стране (два репозиторијума), позитивне и негативне.

Како се једна супстанца са одређеном InChIKey вредношћу може наћи у оквиру више репозиторијума, поставља се питање како се може доћи до информација које су везане за њу тј. којем репозиторијуму припада, који се префикси користе, на који начин је та супстанца повезана са истим или сличним супстанцама у оквиру репозиторијума и слично. Одговор на нека питања се може добити помоћу метода кластеризације као типа ненадгледаног учења. Добијени граф који се састоји од више репозиторијума је по природи ствари густ тамо где су чворови репозиторијума, због веза које постоје у њему, док су везе које ми креирамо између репозиторијума ретке. Ово нас доводи до појма кластеризације која може са великом прецизношћу да детектује кластере и тачно одреди који чвор припада којем кластеру (репозиторијуму), иако може да се ради о чворовима са истим ознакама који се јављају у више репозиторијума (видети Слику 7.5). На почетку ће бити објашњен појам кластеризације података и набројани њени различити приступи, а затим приказани кораци алгоритма за *спектралну кластеризацију* који је од посебног интереса за нас. Посебан део биће посвећен оригиналним резултатима који дају математичко објашњење у кластеризацији одређене врсте графова.

Кластер анализа или, краће, *кластеровање* је задатак додељивања скупа објеката у групе (кластере) тако да су објекти у истом кластеру слични један другом, а мање слични у односу на објекте у другим кластерима. Кластеровање је главни задатак проучавања истраживачког *data mining*-а, технике за статистичку анализу података која се користи у многим областима, укључујући машинско учење, препознавање облика, анализу



Слика 7.5: Заједнички скупови подтака два репозиторијума.

слика, претраживање информација (енг. *Information Retrieval*) и биоинжењеринг. Сама по себи кластер анализа није један специфичан алгоритам, већ општи задатак који би требало решити помоћу различитих алгоритама који се значајно разликују у начину на који конструишу кластере и колико ефикасно их проналазе. У популарном смислу, кластери представљају групе са малим растојањима између чланова кластера, густе области у простору података, интервале или специфичну статистичку расподелу. Кластеровање се зато може и формулисати као више-критеријумски проблем оптимизације. Прикладан алгоритам за кластеровање и својства параметара (укључујући вредности као што је функција растојања, праг густине или број очекиваних кластера) зависе од скупа података са којим се ради и намене резултата. Кластер анализа као таква није аутоматски задатак, већ итеративни процес откривања знања или интерактивна више-критеријумска оптимизација која укључује испробавање и неуспехе и често је неопходно мењање параметара док се не постигну жељени резултати.

7.4.1 Типови кластеризације

Појам „кластер” се разликује од алгоритма до алгоритма и представља један од многих услова који одређује избор одређеног алгоритма за специфични проблем. Прва асоцијација на реч кластер је група објеката. Ипак, кластери добијени различитим алгоритмима се значајно разликују по њиховим карактеристикама, тако да је разумевање ових „кластер модела” кључ у разумевању разлика између различитих алгоритама. У наставку је дат кратак преглед различитих типова алгоритама за кластеризацију података са посебним освртом на моделима који се базирају на графовима. Детаљан преглед свих алгоритама за кластеризацију података се може наћи у књизи [111].

Кластеровање базирано на повезаности

Кластеровање базирано на повезаности, такође познато и као хијерархијско кластеровање, је базирано на основној идеји да су објекти више повезани са ближим објектима него са даљим. Као такви, ови алгоритми повезују објекте да би формирали кластере према њиховој удаљености. Кластер у великој мери може да буде описан преко минималног растојања које је потребно да повеже делове кластера. Са различитим растојањима различити кластери ће се формирати, што може бити приказано помоћу стабла које објашњава одакле заједничко име, *хијерархијско кластеровање*, долази. Ови алгоритми не обезбеђују једну поделу скупа података већ обимну хијерархију кластера који се спајају једни са другима на одређеним удаљеностима. У стаблу y-оса означава растојање на којем се кластери спајају, док су дуж x-осе смештени објекти тако да се кластери не мешају.

Кластеровање базирано на повезаности је читава фамилија метода која се разликује по начину на који се одређује растојање. Осим уобичајеног избора функција растојања, кориснику је такође потребно да одлучи о критеријуму повезивања (пошто кластер садржи више објеката, постоји више кандидата за израчунавање растојања) који ће се користити. Популарни избори су познати као једноструко повезивање кластера (минимално растојање међу објектима), комплетно повезивање кластера (максимално растојање међу објектима) или UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*), познат и као просечно повезивање кластера.

Иако су ови алгоритми једноставни за разумевање, резултати нису увек лаки за коришћење. Све док се не направи јединствена подела скупа података, корисник ће и даље морати да бира одговарајуће кластере из хијерархије. У општем случају сложеност алгоритма је $\mathcal{O}(n^3)$ што га чини спорим за велики скуп података. За неке специјалне случајеве, као што су *SLINK* за једноструко повезивање и *CLINK* за комплетно повезивање кластера, сложеност алгоритма је $\mathcal{O}(n^2)$).

Кластеровање базирано на централности

У кластеровању базираном на централности, кластери су представљени помоћу централног вектора који не мора бити члан скупа података. Ако је k фиксиран број кластера, k-means алгоритам за кластеризацију података даје формалну дефиницију оптимизационог проблема: одредити k центара кластера и додели објекте најближим центрима тако да су квадрати растојања између кластера минимизовани. Оптимизациони проблем је сам по себи NP-тежак и стандардни приступ у његовом решавању је трагање за приближним (енг. *approximated*) решењима. Добро познати апроксимативни метод је *LLoyd*-ов алгоритам, који се често назива k-means алгоритам (иако је други алгоритам увео овај назив). Ипак он само налази локални минимум и обично се покреће више пута са различитом случајном иницијализацијом.

Варијације k-means алгоритма обухватају такве оптимизације као што су избор најбољег корака, затим ограничења да центроиди кластера морају да буду чланови скупа података (k-medoids), избор медијана кластера (k-medians clustering), даље, бирање иницијалних центара мање случајно (k-means++) или дозвољавајући фази кластер додељивање (*fuzzy clustering*).

Већина алгоритама типа k-means захтева да број кластера k буде дат унапред што је једна од највећих мана ових алгоритама. Даље, алгоритми преферирају кластере сличне величине, док се објекти увек додељују најближем центроиду. Ово често води ка лошој одлуци везаној за границе између кластера (што није изненађујуће, јер алгоритам оптимизује центре кластера, али не и кластере ивица). k-means има бројне интересантне теоријске особине. На једној страни, он дели простор података у структуре које су познате као Воронојеви дијаграми, док је на другој страни, концептуално сличан класификацији најближег суседа и као такав је популаран у машинском учењу.

Кластеровање базирано на расподели

Метода за кластеровање која је највише повезана са статистиком се базира на моделима расподеле. Кластери се лако могу дефинисати као објекти који припадају највероватније истој расподели. Добра особина оваквог приступа је начин на који су вештачки скупови података генерисани: узорковањем случајно изабраних објеката из расподеле. Док је теоријска основа ових метода добра, оне имају један кључни проблем познат као "преучавање" модела (енг. *overfitting*), осим ако се у комплексност модела не укључе ограничења. Комплекснији метод ће увек моћи да објасни податке боље, што избор одговарајућег модела чини компликованим.

Метода из ове класе алгоритама која се највише користи је позната под називом Gaussian Mixture models која се базира на *Expectation-Maximization (EM)* алгоритму. Скуп података се обично моделује са фиксним бројем Гаусових расподела (како би се избегло "преучавање" модела) које се иницијализују случајним избором и чији се параметри итеративно оптимизују да би се скуп података боље симулирао. Ови модели ће конвергирати ка локалном минимуму, тако да ће вишеструко понављање произвести различите резултате.

Кластеровање базирано на расподели је семантички „снажан" модел, јер не обезбеђује само кластере, већ креира комплексне моделе за кластере који могу да открију везе и зависности између атрибута. Ипак, коришћење ових метода ствара додатни терет кориснику у тренутку када треба изабрати одговарајући модел података за оптимизацију, што некада није могуће (Гаусова расподела подразумева јаку претпоставку о подацима).

Кластеровање базирано на густини

У кластеровању базираном на густини, кластери су дефинисани као области веће густине од осталог дела. Објекти у ретким областима (за које се захтева да раздвајају кластере) се обично третирају као шум или граничне тачке.

Најпопуларнији метод за кластеровање који је базиран на густини је *DBSCAN*. За разлику од многих новијих метода, он има јасно дефинисан принцип кластеризације података који се зове „густина-досег" (енг. *density-reachability*). Слично као код кластеровања базираног на повезаности, метод се заснива на повезивању тачака унутар одређене границе растојања. Међутим, он повезује само оне тачке које задовољавају одређени критеријум густине, оригинално дефинисан као минимални број других објеката унутар радијуса. Кластер садржи све густо повезане објекте (што може да кре-

ира кластер произвољног облика), и додатно, све објекте који су унутар досега ових објеката. Друга интересантна особина *DBSCAN*-а је сасвим мала сложеност (захтева линеаран број упита по опсегу над базом података) и он ће у суштини дати исте резултате при сваком покретању, па зато нема потребе да се алгоритам покреће више пута. *OPTICS* је генерализација *DBSCAN* методе која елиминише потребу за избором одговарајуће вредности за параметар ε, и који на излазу даје хијерархијски резултат (дендограм) који је у спрези са кластеровањем базираним на повезивању. Кључни недостатак *DBSCAN* и *OPTICS* алгоритама је тај што се очекује одређена врста брисања густине како би се детектовале границе кластера. На скупу података са преклапајућим Гаусовим расподелама (чест случај коришћења у вештачки генерисаним подацима), границе кластера које су добијене овим алгоритмима често изгледају произвољно, зато што густина кластера константно опада. За овакве скупове података, *EM* алгоритам скоро увек даје боље резултате у односу на њих. Варијација *DBSCAN*-а, *EnDBSCAN*, ефикасно детектује овакву врсту структура.

Спектрална кластеризација

У последњих неколико година спектрално кластеровање (енг. *spectral clustering*) је постало један од најпопуларнијих алгоритама за кластеризацију података. Алгоритам је веома једноставан за имплементацију, може се ефикасно решити стандардним софтвером за линеарну алгебру, и веома често превазилази традиционалне алгоритме за кластеровање као што је k-means алгоритам. На први поглед спектрално кластеровање је крајње мистериозно, јер није очигледно зашто добро ради и шта се заиста дешава са подацима након трансформације.

Нека је дат скуп тачака $x_1, x_2, \ldots, x_n$ и неки је сличност између парова тачака x_i и x_j одређена вредношћу $s_{ij} \geq 0$. Интуитивни циљ кластеровања је подела скупа података на неколико група тако да су тачке у истој групи сличне међусобно, а тачке из различитих група нису. Ако немамо пуно информација изузев сличности између података, добар начин представљања података је у форми *тежинског графа* $G = (V, E)$ или преко њему одговарајуће матрице сличности $S = (s_{ij})_{i,j=1,\ldots,n}$. Сваки чвор v_i у графу представља податак x_i. За два чвора казаћемо да су повезана ако је сличност s_{ij} између одговарајућих тачака x_i и x_j позитивна или већа од одговарајућег прага, при чему је s_{ij} тежина гране. Проблем кластеровања се сада може преформулисати коришћењем тежинског графа: желимо да одредимо поделу (рез) графа тако да гране између различитих група имају веома малу тежину (значи да се тачке у различитим кластерима разликују међусобно), а гране унутар групе имају велику тежину (тачке унутар истог кластера су сличне међусобно).

Приликом конструисања тежинског графа циљ је измоделирати локалне везе између суседа које најбоље осликавају односе између њих. Постоји неколико популарних начина за трансформацију скупа података $x_1, x_2, \ldots, x_n$ у тежински граф:

1. *ε-околина графа:* Тачке се повезују ако су међусобна растојања мања од одређеног прага ε.

2. *k-најближих суседа графа:* Циљ је повезати чвор v_i са чвором v_j, ако је v_j међу k-најближих суседа чвора v_i.

3. *потпуно повезан граф:* Повезују се све тачке са позитивном међусобном сличношћу.

Технике спектралног кластеровања користе спектар матрице сличности података да би се добило смањење димензије за кластеровање. Једна таква техника је *алгоритам нормализованог одсецања* или *Shi-Malik алгоритам*, који су развили Jianbo Shi и Jitendra Malik, која се најчешће користи за сегментацију слика. Скуп тачака се дели на два скупа (S_1, S_2) помоћу сопственог вектора v, који одговара другој најмањој сопственој вредности нормализоване Лапласове матрице матрице L

$$L = I - D^{-1/2} S D^{-1/2},$$

где је D дијагонална матрица $d_{ii} = \sum_j s_{ij}$.

Подела графа се може извршити на различите начине, као што је на пример узимање средње вредности m свих координате из v, и смештањем свих чворова чије су ордговарајуће координате из v веће од m у скуп S_1, а остале у скуп S_2. Алгоритам се може искористити за хијерархијско кластеровање вишеструким понављањем исте процедуре на подскупове.

Алгоритми спектралног кластеровања

У наставку су дате најчешће коришћене верзије алгоритма спектралног кластеровања. Претпоставићемо да подаци садрже n „тачака" $x_1, x_2, \ldots, x_n$ које могу бити произвољни објекти. Сличност s_{ij} међу њима је исказана неком функцијом сличности која је симетрична и ненегативна, а одговарајућа матрица сличности је означена са $S = (s_{ij})_{i,j=1\ldots n}$.

Алгоритам 1 Ненормализовано спектрално кластеровање.

Улаз: Матрица сличности $S \in R^{n \times n}$, број кластера k.
Резултат: Кластери $A_1, \ldots, A_k$ са $A_i = \{j | y_j \in C_i\}$.

1 Одредити матрицу сличности S тежинског графа G.
2 Одредити Лапласову матрицу $L = D - S$.
3 Одредити првих k сопствених вектора $u_1, u_2, \ldots, u_k$ од L.
4 Нека је U матрица која садржи векторе $u_1, u_2, \ldots, u_k$ као колоне.
5 За $i = 1, \ldots, n$ нека је $y_i \in R^k$ вектор који одговара i-том реду матрице U.
6 Кластеризовати тачке $(y_i)_{i=1,\ldots,n}$ из R^k са k-means алгоритмом у кластере $C_1, \ldots, C_k$.

За разлику од Алгоритма 1, Shi и Malik су у свом раду [104] за израчунавање првих k сопствених вектора, уместо Лапласове матрице L, користили генерализовани проблем сопствених вредности $Lu = \lambda Du$. Нешто другачији је алгоритам који је коришћен у раду [112]:

Алгоритам 2 Спектрално кластеровање према Ng, Jordan и Weiss.

Улаз: Матрица сличности $S \in R^{n \times n}$, број кластера k.
Резултат: Кластери $A_1, \ldots, A_k$ са $A_i = \{j | y_j \in C_i\}$.

1 Одредити матрицу сличности S тежинског графа G.

2 Израчунати нормализовану Лапласову матрицу L_{sym}

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} S D^{-1/2}.$$

3 Одредити првих k сопствених вектора $u_1, u_2, \ldots, u_k$ од L_{sym}.

4 Нека је U матрица која садржи векторе $u_1, u_2, \ldots, u_k$ као колоне.

5 Формирати матрицу $T \in R^{n \times k}$ од матрице U нормализацијом редова на вредност 1, при чему се добија $t_{ij} = u_{ij} / (\sum_k u_{ik}^2)^{1/2}$.

6 За $i = 1, \ldots, n$ нека је $y_i \in R^k$ вектор који одговара i-том реду матрице T.

7 Кластеризовати тачке $(y_i)_{i=1,\ldots,n}$ из R^k са k-means алгоритмом у кластере $C_1, \ldots, C_k$.

У сва три алгоритма, главна идеја је промена репрезентације апстрактних података x_i у тачке y_i.

За имплементацију спектралног кластеровања у пракси потребно је израчунати првих k сопствених вектора потенцијално велике Лапласове матрице. Срећом, ако се користи k-најближих суседа графа или ε-околина графа, све Лапласове матрице су ретке (енг. *sparse*). Међу најпопуларнијим методама за израчунавање првих сопствених вектора ретких матрица су *метод степеновања* (енг. *power method*), *потпростор методи Krylov-a*, QR алгоритам, и *Lanczos метод*. Детаљи везани за ове методе се могу наћи у [113]. Брзина конвергенције ових алгоритама зависи од величине спектралног јаза (енг. *spectral gap*) $\gamma_k = \|\lambda_k - \lambda_{k+1}\|$. Што је већи спектрални јаз, алгоритам брже израчунава првих k сопствених вектора.

Број кластера

Један од основних проблема који се јављају у свим алгоритмима кластеровања је избор броја кластера k. Данас постоје различите методе које се користе како би се овај проблем решио. На пример, у кластеровању које се базира на моделима постоји добро оправдан критеријум за избор броја кластера из података. Алат који је посебно дефинисан за спектрално кластеровање је хеуристика која се базира на спектралном јазу (енг. *eigengap*) [80, 88], који се може искористити за сва три претходно поменута Лапласова графа. Овде је циљ изабрати број k тако да су све сопствене вредности $\lambda_1, \ldots, \lambda_k$ мале, а λ_{k+1} релативно велика. Постоји неколико оправдања за овакву процедуру. Једно од њих се базира на пертурбационој теорији, где у идеалном случају од k комплетно неповезаних кластера, сопствена вредност 0 има мултиплицитет k, и ту постоји јаз у односу на $(k+1)$. сопствену вредност $\lambda_{k+1} > 0$.

7.4.2 Одређивање најутицајнијих чворова

Поред тога што је потребно одредити афилиацију сваког чвора тј. његову припадност репозиторијуму, потребно је додатно одредити и који су то најутицајнији чворови у диграфу како бисмо управо њих искористили за повезивање два репозиторијума и смањивање међурезултата, али и фаворизовање одређених скупова података.

Светски познати претраживачи Веба (енг. *search engines*) као што су Google, Bing, Yahoo и Yandex користе различите методологије како би рангирали резултате претраге. Основна идеја која чини основу ових претраживача је рангирање страница на Интернету на основу структуре линкова графа чији су чворови странице, док се показује да је важност неке странице пропорционална одговарајућој координати Пероновог сопственог вектора.

PageRank је техника за анализу линкова која сваком чвору графа додељује нумеричку вредност између 0 и 1, и која зависи од структуре линкова. За задати упит, машине за претрагу израчунавају сложени скор за сваку Веб страницу која комбинује стотине својстава (енг. *features*) са PageRank скором, као што су то, на пример, косинусна сличност (енг. *cosine similarity*) или близина термина (енг. *term proximity*). Овај сложени скор који се добија помоћу различитих метода се користи да прикаже рангиране резултате претраге. Показано је да се PageRank скор графа може одредити израчунавањем одређеног сопственог вектора графа што ће и бити показано у наставку. У наредном тексту биће објашњен начин на који се долази до рангова за сваки веб-сајт (чвор), а затим ће иста идеја бити употребљена у SpecINT софтверској платформи, где ће се по узору на рангирање Веб сајтова одредити најутицајнији скупови података.

PageRank вектор

Замислимо на почетку просечног Интернет корисника који почиње претрагу на одређеној Веб страници и случајно се креће Вебом. У сваком временском тренутку, он наставља кретање са своје тренутне стране A ка случајно изабраној страни према којој постоји хиперлинк на страни A. Нека од чвора A постоје три хиперлинка и то према чворовима B, C и D на које корисник (сурфер) може да пређе. Корисник у следећем тренутку наставља претрагу прелазећи на један од ова три чвора са једнаком вероватноћом која износи $\frac{1}{3}$. Како сурфер наставља случајну шетњу од чвора до чвора, он одређене чворове посећује више него друге; инутитивно, то су чворови који имају много линкова из других често посећених чворова који указују на њих (улазни степен има велику вредност). Идеја иза Page-Rank алгоритма је да странице које су чешће посећене у шетњи буду више важне.

Прво питање на које треба дати одговор је шта ако тренутна локација на којој се налази сурфер, чвор A, нема излазне линкове или је корисник унео нову адресу у URL пољу претраживача? Да би се решио овај проблем уводи се додатна операција за случајног корисника: операција *телепортовања*. У операцији телепортовања сурфер скаче на било који други чвор графа. Нова дестинација телепорт операције се бира униформно од свих могућих Веб страна. Другим речима, ако је N укупан број чворова у Веб графу, ова операција пребацује сурфера на било који други чвор са вероватноћом $\frac{1}{N}$. Сурфер се такође телепортује на тренутну позицију са вероватноћом $\frac{1}{N}$.

У циљу додељивања PageRank скора сваком чвору графа, операција телепортације се користи на два начина:

- Када у чвору не постоје излазни линкови, сурфер позива телепорт операцију.
- У произвољном чвору који има излазне линкове, сурфер позива телепорт са вероватноћом $0 < \alpha < 1$, а наставља стандардну случајну шетњу (следи из униформно изабраног излазног линка) са вероватноћом $1 - \alpha$, где је α унапред задати параметар, најчешће $\alpha = 0.1$.

У наредном тексту биће приказано коришћење теорије Маковљевих ланаца да бисмо објаснили да када сурфер прати овај комбиновани процес (случајна шетња и телепорт) он посећује сваки чвор v Веб графа у фиксном делу времена $\pi(v)$ које зависи од *структуре īграфа* и *вредности* α. Вредност $\pi(v)$ се назива PageRank чвора v и у наставку ћемо показати како се ова вредност израчунава.

Марковљеви ланци

Марковљев ланац означава дискретни Марковљев случајни процес: процес који се појављује у серији временских корака у којима је направљен случајни избор. Марковљев ланац садржи N корака. Свака Веб страна одговара стању у Марковљевом ланцу који ћемо формулисати касније.

Марковљев ланац се описује $N \times N$ транзиционом матрицом вероватноћа P у којој је сваки елемент у интервалу [0, 1] при чему је сума вредности у сваком реду матрице је једнака броју 1. Марковљев ланац може се наћи у једном од N стања у датом тренутку, тада, елемент P_{ij} представља вероватноћу да је стање у следећем тренутку j, под условом да је тренутно стање i. Сваки елемент P_{ij} се назива транзициона вероватноћа и зависи само од тренутног стања i тј. поред датог тренутног стања, будуће стање система не зависи од прошлих. Другим речима, то значи, да опис садашњости у потпуности садржи информацију која може утицати на будуће стање процеса. Ово својство је познато као Марковљево својство. Тада, за Марковљево својство важи:

$$P_{ij} \in [0, 1], \text{ за } i, j = 1, \ldots, N$$

и

$$\sum_{j=1}^{N} P_{ij} = 1, \text{за } i = 1, \ldots, N. \tag{7.1}$$

Матрица са ненегативним елементима која задовољава једнакост (7.1) се назива стохастичком матрицом. Кључно својство стохастичке матрице је да главни леви сопствени вектор одговара највећој сопственој вредности која је једнака јединици.

Функција расподеле Марковљевог ланца у односу на његова стања се може третирати као вектор вероватноћа: вектор чије су координате из интервала $[0, 1]$, при чему је њихова сума једнака јединици. N-димензиони вектор вероватноћа, чија свака координата одговара једном од N стања Марковљевог ланца, се може третирати као функција расподеле за сва његова стања. За прост Марковљев ланац од 3 чвора, вектор вероватноћа имаће 3 координате чија је сума једнака 1.

Кретање случајног сурфера на Веб графу се може посматрати као Марковљев ланац, са једним стањем за сваку страницу, при чему свака транзициона вероватноћа представља вероватноћу преласка са једне странице на другу (операција телепортовања такође утиче на ове вредности). Означимо са A матрицу суседства Веб графа реда N, где су странице чворови графа, а хиперлинкови гране између њих.

Из матрице A се може добити транзициона матрица вероватноћа P за Марковљев ланац на следећи начин:

1. Ако ред у матрици A не садржи ниједну јединицу, тада сваки елемент треба заменити са вредношћу $\frac{1}{N}$ и тако за сваки ред.
2. Поделити сваку вредност 1 у A са укупним бројем јединица у том реду.
3. Помножити резултујућу матрицу са $1 - \alpha$.
4. Додати вредност $\frac{\alpha}{N}$ сваком елементу резултујуће матрице да би се добила матрица P.

Функција расподеле позиције сурфера у произвољном тренутку може бити описана помоћу вектора вероватноће $\mathbf{x}$. У тренутку $t = 0$ сурфер може да започне претрагу од стања које у вектору $\mathbf{x}$ има 1, док су све остале вредности једнаке нули. На основу дефиниције, расподела у тренутку $t = 1$ је дата вектором вероватноће $\mathbf{x}P$, у тренутку $t = 2$ са $(\mathbf{x}P)P = \mathbf{x}P^2$ итд. Другим речима, расподела стања клијента у било којем тренутку је одређена почетном расподелом и транзиционом матрицом вероватноћа P.

Ако се Марковљев ланац извршава више пута, свако стање се посећује различитом фреквенцијом која зависи од структуре ланца. Ово је случај и на Вебу где, на пример, корисник одређене спортске странице посећује доста чешће него рецимо оне које пишу о моди. Сада је циљ дефинисати под којим условима се ове фреквенције неће мењати, већ ће да конвергирају ка стабилном стању (енг. steady-state). PageRank вредност сваког чвора v биће управо вредност овог стабилног стања за фреквенцију посете.

Дефиниција 7.1. За ланац Маркова се каже да је *ергодичан* ако је могуће прећи из било ког стања у било које стање (не обавезно у једном кораку).

Да би Марковљев ланац био ергодичан, потребно да је буду испуњена два услова: *иредуцибилност* и *апериодичност*. Први услов обезбеђује постојање низа транзиција са не-нула вероватноћом из произвољног стања у било које друго стање, док други услов обезбеђује да стања нису подељена у скупове тако да се све транзиције стања понашају циклично прелазећи из једног у други скуп.

Теорема 7.1. За произвољан ергодичан Марковљев ланац постоји јединствени вектор вероватноћа за стабилно стање π који је главни леви сопствени вектор матрице P, такав да ако је $\eta(i, t)$ број посета стања i у t корака, тада важи

$$\lim_{t \to \infty} \frac{\eta(i, t)}{t} = \pi(i)$$

где је $\pi(i)$ вероватноћа стабилног стања за стање i.

Из Теореме 7.1 следи да случајна шетња са телепортовањем резултује јединственом расподелом вероватноћа стабилног стања за сва стања индукованог Марковљевог ланца. Вероватноћа стабилног стања за одређено стање је заправо PageRank вредност одговарајуће Веб странице.

Израчунавање PageRank вредности

Сада би требало дати одговор на који начин се могу израчунати PageRank вредности за сваки чвор графа. На почетку, да се подсетимо да је леви сопствени вектор транзиционе матрице P заправо N-димензиони вектор π такав да

$$\pi P = \lambda \pi. \tag{7.2}$$

Као што већ речено, N вредности главног сопственог вектора π су вероватноће стабилног стања случајне шетње са телепортацијом, па су самим тим ово PageRank вредности за одговарајуће Веб странице. Претходна једначина 7.2 се може интерпретирати на следећи начин: ако је π функција расподеле преласка сурфера са странице на страницу, он остаје унутар стабилног стања расподеле π. Ако је π стабилно стање расподеле, тада важи $\pi P = 1\pi$, што значи да је 1 сопствена вредност од P. Тада, ако израчунамо главни, леви сопствени вектор матице P који одговара сопстевној вредности 1, онда смо заправо израчунали PageRank вредности.

Леви сопствени вектор матрице се може израчунати методом *степене итерације* (енг. power method). Ако је x почетна расподела стања, тада је расподела у тренутку t дата са xP^t. Када се t повећава, очекивано је да расподела xP^t буде слична расподели xP^{t+1}, јер се очекује да Марковљев ланац достигне стабилно стање за велико t, при чему је ова вредност је независна од почетне расподеле x према Теореми 7.1.

7.5 Алгоритам за генерисање Federated SPARQL упита

Пошто смо видели на који начин се сопствени вектори графа могу искористити за одређивање афилиације и ранга чворова графа, у овој секцији биће описан начин на који се ови вектори могу употребити за креирање Federated SPARQL упита. У наставку је дат алгоритам који на основу посматране хемијске структуре (супстанца или комплекс) и постављеног питања конструише Federated SPARQL упите који враћају тражене резултате из два посматрана репозиторијума.

Пре него што представимо алгоритам на којем се базира софтверска платформа, потребно је да најпре скренути пажњу на одређене детаље. Прво, треба имати у виду да се за сваку супстанцу добијају различити графови, јер оријентација грана не мора увек да буде иста. На пример, гране не иду увек од скупа DrugBank ка скупу KEGG, тако да се не може рећи да је грана увек усмерена од супстанце из DrugBank-а ка супстанци из KEGG-а или обрнуто, што зависи од тога како је репозиторијум формиран. Такође, супстанце не морају да буду присутне у сваком скупу репозиторијума, шта више, дешава се да ендпоинт одређеног скупа у датом тренутку није активан. Друго, овде је представљен алгоритам који ради са два репозиторијума и довољно је, због начина повезивања графова, употребити само Фидлеров сопствени вектор уместо целокупног

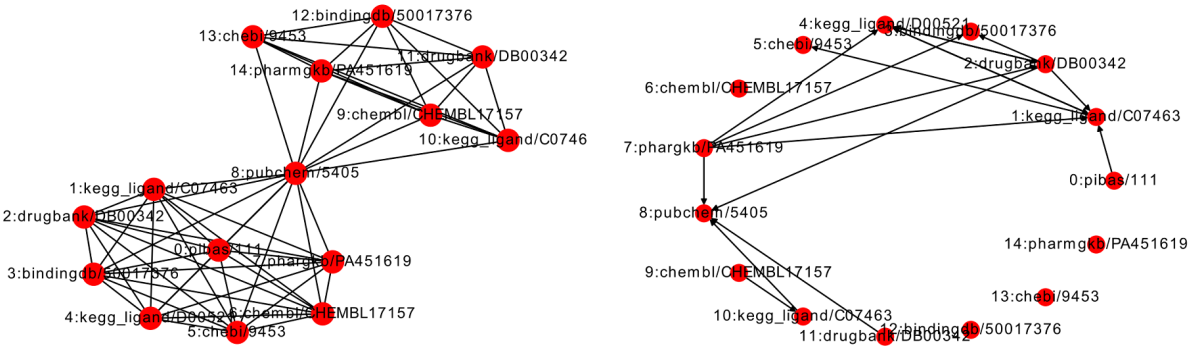
алгоритма за спектрално кластеровање. На основу резултата Теореме 6.15 довољно је израчунати само један сопствени вектор и имати исте информације које бисмо добили алгоритмом за спектрално кластеровање. Према теореми важи да артикулационом чвору у Фидлеровом сопственом вектору одговара координата чија је вредност 0, док чворовима са једне његове стране одговарају само позитивне, а са друге стране само негативне координате вектора. Алгоритам спектралног кластеровања је неопходан када се ради са три или више репозиторијума, при чему је тада потребно израчунати k сопствених вектора и након тога применити k-means алгоритам.

Нопходни кораци које је потребно проћи да би се креирали валидни SPARQL упити су приказани у оквиру Алгортима 3. Улаз у алгоритам су сопствени вектори s и r графова $V_{n,m}$ и $D_{n,m}$, редом, као и називи два репозиторијума (већ изабрани приликом конструисања графова). Дакле, пре позива алгоритма се за посматрану супстанцу најпре одреде графови на основу Процедуре 1, а затим се израчунају тражени сопствени вектори који се користе као улаз у алгоритам. Користећи информације које у себи носе сопствени вектори, алгоритам треба да обезбеди повезивање одговорајућих подупита и да као излаз да Federated SPARQL упит који може да врати резултате са различитих репозиторијума. Начин на који алгоритам бира како ће да повеже подупите биће објашњен у наставку, корак по корак, узимајући у обзир један од могућих случајева. Такође, након тога ћемо објаснити и саму процедуру за комплексније случајеве, на пример, када два репозиторијума нису повезана само артикулационим чвором већ са једном или више грана, или када се генерише упит који се извршава над више репозиторијума и друго.

Алгоритам 3 Генератор за Federated SPARQL упите.

Улаз: Фидлеров сопствени вектор $s = \{s_0, s_1, ..., s_{n+m-2}\}$,
rank сопствени вектор $r = \{r_0, r_1, ..., r_{n+m-2}\}$,
репозиторијуми R_1 и R_2
Резултат: Federated SPARQL упит

1 $query = \varnothing$
2 $null\_vertex \leftarrow n - 1$
3 $subject \leftarrow label(i)$, i - најбоље рангирани позитивни чвор
4 **repeat**
5 $neighbors \leftarrow$ позитивни суседи за $subject$
$object \leftarrow$ label(најбоље рангирани сусед)
$add\_subquery(subject, object, R_1, pattern)$
$subject \leftarrow object$
6 **until** $object = null\_vertex$;
7 $add\_subquery(subject, null\_vertex, R_1 \, or \, R_2, pattern)$
$subject \leftarrow label(i)$, i - најбоље рангирани негативни чвор
8 **repeat**
9 $neighbors \leftarrow$ негативни суседи за $subject$
$object \leftarrow$ label(најбоље рангирани сусед)
$add\_subquery(subject, object, R_2, pattern)$
$subject \leftarrow object$
10 **until** $object = null\_vertex$;
11 return $query$



Слика 7.6: Повезивање графова који одговарају Chem2Bio2RDF и Bio2RDF репозиторијумима.

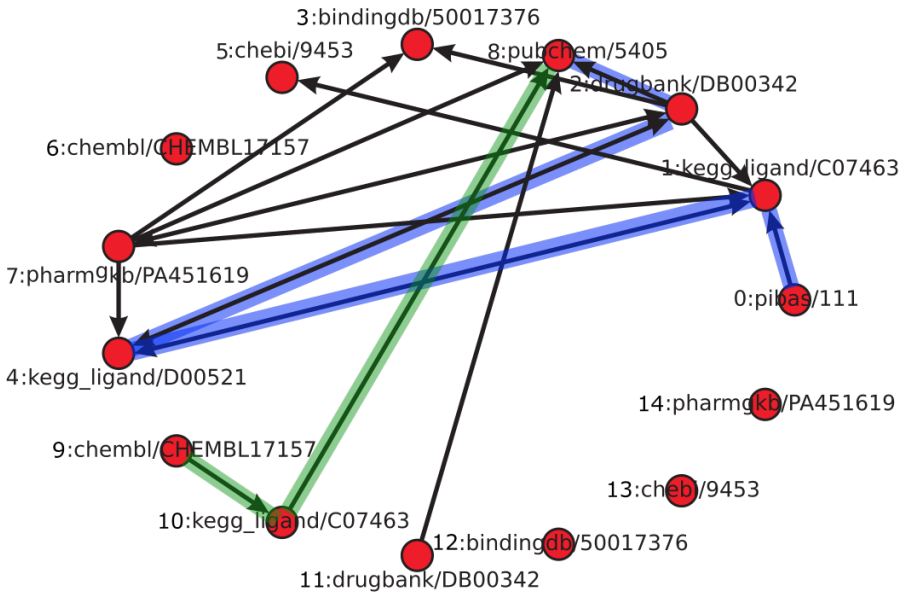
Претпоставимо да истраживачи желе да сазнају нешто више о супстанци (антихистамин) која је позната под називом Терфенадин, а чији је *InChIKey = GUGOEEXESWIERI-UHFFFAOYSA-N*. Пратећи кораке Процедуре 1 конструисани су графови који одговарају овој супстанци. На Слици 7.6, су приказани добијени графови у датом тренутку узимајући у обзир два репозиторијума Bio2RDF и Chem2Bio2RDF. Због једноставности и без губитка општости претпоставимо да прва повезана компонента садржи чворове са позицијама $0, 1, \ldots, n-2$, за повезујући чвор је резервисана позиција $n-1$, док друга повезана компонента садржи чворове са позицијама $n, n+1, \ldots, n+m-2$. За добијене графове се затим одређују два сопствена вектора, s и r. Први сопствени вектор s је Фидлеров сопствени вектор графа $V_{n,m}$ који дели оба графа, $V_{n,m}$ и $D_{n,m}$, на две повезане компоненте са различитим знаковима координата и једним повезујућим $null$-чвором. Овај сопствени вектор носи информације о припадности чворова репозиторијумима, и након трансформација које вршимо ови знакови носе информацију о њиховом пореклу. $null$-чвор представља артикулациони чвор (мост) и њему одговарајућа координата у вектору s има вредност 0. Координате вектора s графа $V_{n,m}$ су

$$s = [0.231, 0.231, 0.231, 0.231, 0.231, 0.231, 0.231, 0.231, 0, -0.309, \\ -0.309, -0.309, -0.309, -0.309, -0.309].$$

Други сопствени вектор r представља најзначаније чворове у оквиру оба репозиторијума. Координате у r заправо сугеришу највероватнију путању до артикулационог чвора унутар знаковних зона, обезбеђујући на тај начин повезивање репозиторијума (видети десну страну Слике 7.6). Његове координате су

$$r = \{0 : 0.1477, 1 : 0.0882, 2 : 0.0223, 3 : 0.0143, 4 : 0.0393, 5 : 0.0344, 6 : 0.0094, 7 : 0.0125, \\ 8 : 0.0490, 9 : 0.1477, 10 : 0.0722, 11 : 0.0094, 12 : 0.0094, 13 : 0.0094, 14 : 0.0094\}.$$

Први корак у генерисању упита је детектовање две путање од којих једна иде од изабраног позитивног чвора до $null$-чвора, а друга од изабраног негативног такође до $null$-чвора. Кренувши од једног чвора, детектују се сви његови суседи за које постоје



Слика 7.7: Спојене путање између два репозиторијумима.

гране које иду ка њима, а затим се бира чвор са највећом PageRank вредношћу. Ако постоји више чворова са истим рангом, онда се путања рачва посебно за сваки од чворова. Тада се као резултат потенцијално може добити више упита, али су то ретки случајеви, јер не могу све путање да се заврше у $null$-чвору, а тај корак је неопходан због повезивања репозиторијума. Процедура повезивања се наставља даље све до избора $null$-чвора. У оваквој путањи, објекат триплета је добијен избором најбоље рангираног чвора из суседства субјекта. Кроз неколико корака приказаћемо како то заиста изгледа.

Нека је један од почетних чворова чвор са позицијом 0 (0:pibas) на Слици 7.6. Његовој нултој позицији одговара нулта координата вектора s и њена вредност је позитивна (0.231). Једини чвор са којим има везу и ка којем је стрелица усмерена је чвор са ознаком 1:kegg_ligand/C07463. Даље, чвор 1 је повезан са чворовима 4:kegg_ligand:D00521, 7:Pharmkgb/PA451619, 5:chebi/9453, 0:pibas/111 и 2:drugbank/DB00342. Међу постојећим гранама се прво „изабере" оне које иду од чвора 1. У том кораку, чворови 2:drugbank/DB00342, 0:pibas/111 и 7:Pharmkgb/PA451619 се елиминишу из ужег избора, јер не постоји грана ка њима. Од преостала два чвора, 4:kegg_ligand:D00521 и 5:chebi/9453, оба имају позитивне координате, при чему PageRank чвора 4 има вредност 0.0393, док је за чвор 5 она 0.0344. Овом приликом узимамо најбоље рангиран чвор, а то је чвор 4:kegg_ligand/D00521 који наставља наш низ. Да би се избегао повратак у исти чвор, приликом следећег избора он се не разматра. Процедура се наставља све док се не достигне чвор 8:pubchem/5405 (артикулациони чвор). Исто се ради и са друге стране, када је иницијални чвор неки од чворова којем одговара негативна координата. За више детаља погледати Слику 7.7. Коначно, путања преко позитивних чворова (плава боја) која одговара Bio2RDF репозиторијуму је: $0 \rightarrow 1 \rightarrow 4 \rightarrow 2 \rightarrow 8$ ($pibas/111 \rightarrow kegg\_ligand/C07463 \rightarrow kegg\_ligand/DB00521 \rightarrow drugbank/DB00342 \rightarrow pubchem/5405$). Друга путања која одговара негативним чворовима (зелена боја) припада Chem2Bio2RDF репозиторијуми и она је: $9 \rightarrow 10 \rightarrow 8$ ($chembl/CHEMBL17157 \rightarrow kegg\_ligand/C07463 \rightarrow pubchem/5405$).

Пошто су путање креиране, а самим тим и целокупна путања која прелази преко оба репозиторијума, информације о једној супстанци постају доступне са више репозиторијума. Сада се за сваки чвор у путањи користи одређени образац за повезивање подупита у оригинални упит који нам одговара (видети примере у Табели 3.6). Гране између чворова графа $D_{n,m}$ се користе за ланчано повезивање подупита на такав начин да објекат једног шаблона графа постаје субјекат следећег. На пример, субјекат ***drugbank:DB00342*** је добијен као објекат триплета $(kegg\_ligand : D00521, http : //bio2rdf.org/kegg\_vocabulary : x - drugbank, ?drugbank)$, чији предикат представља грану у графу $D_{n,m}$. На овај начин смо успели да повежемо различите репрезентације исте супстанце у више различитих репозиторијума, при чему је конструисан је SPARQL упит (видети Листинг 7.1) који враћа таргете за унету супстанцу.

Сам алгоритам у оваквом облику има неколико недстатака. Прво, поставља се питање на који начин у путању можемо да потенцијално укључимо нове скупове података, као што је PIBAS, који немају пуно грана и самим тим имају ниску PageRank вредност. Друго, овакав алгоритам ће увек враћати исту путању, без обзира што се постављена питања разликују. На пример, неки чворови не садрже информације о ћелијским линијама, а изабрани су у путањи, док други садрже, а ипак нису изабрани. Све ово нас доводи до ризика да ће одређени подаци да остану ван домета корисника. И треће, да ли постоји начин да будемо сигурни да ће се путање спојити у артикулационом чвору. Одговоре на сва ова питања даћемо у следећем поглављу где ћемо описати три коришћене хеуристике и резултате добијене њиховом применом. Ту ћемо показати да одређени чворови могу бити фаворизовани и да је могуће за различита питања урадити додатну селекцију (филтрирање) скупова података које ћемо узети у обзир.

```
PREFIX  drugbank: <http://bio2rdf.org/drugbank:>
PREFIX  pibas: <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS#>
PREFIX  drugbank1: <http://chem2bio2rdf.org/drugbank/resource/drugbank_drug/>
PREFIX  kegg_ligand: <http://bio2rdf.org/kegg:>
PREFIX  chembl_molecule: <http://rdf.ebi.ac.uk/resource/chembl/molecule/>
PREFIX  cco:  <http://rdf.ebi.ac.uk/terms/chembl#>
PREFIX  chembl_mapp: <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/chembl#>

SELECT DISTINCT  ?target
FROM <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS/pibasmapping.owl>
FROM <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS/chemblmapping.owl>
WHERE
  { {   { pibas:111 pibas:sameAs kegg_ligand:C07463 .
          pibas:111 pibas:hasTarget ?target.
        }
      UNION
        { SERVICE SILENT <http://kegg.bio2rdf.org/sparql>
            { kegg_ligand:C07463 <http://bio2rdf.org/kegg_vocabulary:gene> ?target;
                                 <http://bio2rdf.org/kegg_vocabulary:same-as> ?kegg_ligand.
            }
        }
      UNION
        { SERVICE SILENT <http://kegg.bio2rdf.org/sparql>
            { kegg_ligand:D00521 <http://bio2rdf.org/kegg_vocabulary:gene> ?target;
                                 <http://bio2rdf.org/kegg_vocabulary:x-drugbank>
                                     ?drugbank.
            }
        }
      UNION
        { SERVICE SILENT <http://drugbank.bio2rdf.org/sparql>
            { drugbank:DB00342 <http://bio2rdf.org/drugbank_vocabulary:target> ?target;
                               <http://bio2rdf.org/drugbank_vocabulary:x-pubchemcompound>
                                   ?pubchem.
            }
        }
      UNION
        { SERVICE SILENT <http://147.91.203.161:8890/sparql>
            { ?value <http://chem2bio2rdf.org/pubchem/resource/CID> pubchem:5405.
              ?value <http://chem2bio2rdf.org/pubchem/resource/CID_GENE> ?target.
            }
        }
      UNION
        { SERVICE SILENT <http://147.91.203.161:8890/sparql>
            { ?isValueOf <http://chem2bio2rdf.org/drugbank/resource/DBID>
                drugbank1:DB00342.
              drugbank1:DB00342 <http://chem2bio2rdf.org/drugbank/resource/CID> ?pubchem.
              ?isValueOf <http://chem2bio2rdf.org/drugbank/resource/CID_GENE> ?target.
            }
        }
      UNION
        { SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/chembl/sparql/>
            { ?activity <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> cco:Activity.
              ?activity cco:hasMolecule chembl_molecule:CHEMBL17157.
              chembl_molecule:CHEMBL17157 cco:moleculeXref ?drugbank1.
              ?activity cco:hasAssay ?assay.
              ?assay cco:hasTarget ?target.
            }
        }
    }
  }
```

Листинг 7.1: Финални SPARQL упит.

Глава 8

Евалуација

У овом поглављу су приказани резултати евалуације SpecINT софтверске платформе и том приликом су демонстриране њене способности да у оквиру репозиторијума изабере најрелевантније скупове података, узимајући притом у обзир све њихове специфичности и податке које чувају. Такође, у оквиру овог поглавља биће проверена и исправност конструисаних упита, тј. да ли за почетак генерисани упити враћају било какве резултате, а онда и да ли су ти резултати кориснику употребљиви. Чињеницу коју треба још једном нагласити је да се методологија представљена у овој докторској дисертацији не заснива на заједничкој онтологији која повезује репозиторијуме, већ на тзв. "same as" везама између скупова података. На основу ових предиката (грана), који представљају везе унутар и између репозиторијума, платформа конструише одговарајуће графове који постају улаз у следећи модул платформе, а из њега се даље, као излаз, добијају одговарајући упити. Везе оваквог типа се могу искористити за било које појмове (термине) који су од значаја у датом тренутку. У нашем случају, то су идентификатори хемијских структура, јер се око њих групишу све остале информације. Још једном да се прецизира, евалуација софтверске платформе је урађена са два аспекта како би се: 1) измериле перформансе платформе у виду избора релевантних скупова података за постављено питање и 2) проверила исправност конструисаних SPARQL упита. Дакле, евалуација целокупне платформе се врши у контексту провере да ли је генерисани упит испунио очекивања клијената, то јест да ли као резултат на постављено питање упит враћа релевантне резултате из различитих скупова података (и репозиторијума) и да ли се у добијеним резултатима могу уочити трендови у свету за откривање лекова који ефикасно делују на различите ћелија канцера. У следећем тексту биће описане поставке експеримената (енг. *experimental setup*) и приказана евалуација резултата.

8.1 Поставке експеримената

SpecINT је креиран са циљем да пружи информације о: особинама супстанце (хемијске структуре), физичким и хемијским; начину на који супстанце интерагују са различитим протеинским таргетима; цитотоксичности супстанце на различитим ћелијским линијама и тако даље. Да би се у потпуности испитале све могућности које платформа пружа, покренута је претрага за 50 супстанци/комплекса који су случајно

изабрани из скупова података са којима платформа ради. За проверу методологије, бирање су само супстанце које се налазе у оба репозиторијума. У Табели 8.1 је приказана листа коришћених InChIKeys са њима одговарајућим молекулским формулама.

Штавише, за експерименте су коришћене и супстанце које су оригинално синтетисане за потребе истраживања у Лабораторији. Лабораторија поседује одређени број различитих здравих и канцером заражених ћелијских линија, и на почетку сваког истраживања је од велике важности знати да ли је посматрана супстанца већ негде анализирана и какву је цитотксичност показала у одређеном експерименту. На овај начин истраживачи могу добити додатне информације о сличним супстанцама и њиховим својствима, али исто тако и могућност за поређење резултата који су добијени у Лабораторији са резултатима других институција.

8.2 Репозиторијуми

Како би се унапредиле истраживачке активности на глобалном нивоу, у последњој деценији је покренуто неколико великих хемијско-информатичких и био-информатичких инцијатива које интегришу одређене скупове података. Свака иницијатива има своје специфичности, било да се оне огледају у начину чувања и интегрисања података или тематике која је у њиховом фокусу. Чињеницу коју треба посебно нагласити је да је свака од ових иницијатива направила огроман напредак у презентовању података истраживачкој заједници и дала немерљив допринос у популаризацији и развоју Повезаних података. Неке од најпопуларнијих инцијатива које се базирају на технологијама Семантичког Веба су већ раније описане у Подсекцији 3.2. У нашим експериментима фокус ће бити на два репозиторијума, *Chem2Bio2RDF* и *Bio2RDF*, пре свега због њихове доступности, а затим и специфичне структуре која је погодна за проверу предложене методологије. *Chem2Bio2RDF* је један од најпопуларнијих репозиторијума који складишти више од 80 милиона триплета. *Chem2Bio2RDF* покрива око 25 различитих скупова података који се односе на хемију/биологију агрегирајући податке о генима, хемијским комплексима, болестима, релевантним научним радовима и др, док *Bio2RDF* интегрише јавне биоинформатичке базе података и претвара их у 11 милијарди триплета у више од 35 скупова података. За потребе демонстрације интеграције нових скупова података, у приказаним експериментима су обухваћене *ChEMBL* база података коју одржава *EBI* и *PIBAS* база података која представља интелектуалну својину Лабораторије.

На овакав избор репозиторијума су утицали технички изазови са којима смо се сретали током фазе тестирања. *LODD* ендпоинти доста често нису доступни (проблем са серверима), док *Open PHACTS* припада класи API базираних репозиторијума, што не спада у домен ове дисертације. *Open PHACTS* SPARQL ендпоинти и даље не функционишу у пуном капацитету, јер се помоћу API позива добијају свеобухватнији резултати. Због тога смо *Open PHACTS* искористили за поређење добијених резултата две платформе. Додатни мотив за избор *Bio2RDF* и *Chem2Bio2RDF* репозиторијума је одлична комуникација коју смо остварили са истраживачима који су део ових иницијатива, а који су увек били расположени да продискутују и пруже подршку за било који проблем који се јавио током рада.

Табела 8.1: Тестиране супстанце са њиховим основним подацима.

Ид	**Назив**	**Формула**	**InChIKey**
1.	Alfuzosin	C19-H27-N5-O4	WNMJYKCGWZFFKR-UHFFFAOYSA-N
2.	Trimethadione	C6-H9-N-O3	IRYJRGCIQBGHIV-UHFFFAOYSA-N
3.	Nalidixic Acid	C12-H12N2O3	MHWLWQUZZRMNGJ-UHFFFAOYSA-N
4.	Clobazam	C16-H13-Cl-N2-O2	CXOXHMZGEKVPMT-UHFFFAOYSA-N
5.	Methixene	C20-H23-N-S	MJFJKKXQDNNUJF-UHFFFAOYSA-N
6.	Terfenadine	C32-H41-N-O2	GUGOEEXESWIERI-UHFFFAOYSA-N
7.	Ofloxacin	C18-H20-F-N3-O4	GSDSWSVVBLHKDQ-UHFFFAOYSA-N
8.	Cladribine	C10-H12-Cl-N5-O3	PTOAARAWEBMLNO-KVQBGUIXSA-N
9.	Nitisinone	C14-H10-F3-N-O5	OUBCNLGXQFSTLU-UHFFFAOYSA-N
10.	Methylergometrine	C20-H25-N3-O2	UNBRKDKAWYKMIV-QWQRMKEZSA-N
11.	Chlorzoxazone	C7-H4-Cl-N-O2	TZFWDZFKRBELIQ-UHFFFAOYSA-N
12.	Aminoglutethimide	C13-H16-N2-O2	ROBVIMPUHSLWNV-UHFFFAOYSA-N
13.	alpha-2-Piperidyl-2,8-bis(trifluoromethyl) quinoline-4-methanol	C17-H16-F6-N2-O	XEEQGYMUWCZPDN-UHFFFAOYSA-N
14.	Sulfadiazine	C10-H10-N4-O2-S	SEEPANYCNGTZFQ-UHFFFAOYSA-N
15.	2-amino-6-(1,2-dihydroxypropyl)-5,6,7,8-tetrahydro-1H-pteridin-4-one	C9H15N5O3	FNKQXYHWGSIFBK-UHFFFAOYSA-N
16.	Sapropterin	C9-H15-N5-O3	FNKQXYHWGSIFBK-RPDRRWSUSA-N
17.	Pd(II)complex (palladium(II) complex with 3-[(2-hydroxybenzylidene)amino]-2-thioxoimidazolidin -4-one)	X	BJKBPOHVNYFYGV-WYRZPCFZSA-L
18.	Vinorelbine	C45-H54-N4-O8	GBABOYUKABKIAF-BXZSYHTRSA-N
19.	Anidulafungin	C58-H73-N7-O17	JHVAMHSQVVQIOT-MFAJLEFUSA-N
20.	Clozapine	C18-H19-Cl-N4	QZUDBNBUXVUHMW-UHFFFAOYSA-N
21.	Grepafloxacin	C19-H22-F-N3-O3	AIJTTZAVMXIJGM-UHFFFAOYSA-N
22.	Doxylamine	C17-H22-N2-O	HCFDWZZGGLSKEP-UHFFFAOYSA-N
23.	Norgestrel	C21-H28-O2	WWYNJERNGUHSAO-XUDSTZEESA-N
24.	Norepinephrine	C8-H11-N-O3	SFLSHLFXELFNJZ-QMMMGPOBSA-N
25.	Cidofovir	C8-H14-N3-O6-P	VWFCHDSQECPREK-LURJTMIESA-N
26.	Mirtazapine	C17-H19-N3	RONZAEMNMFQXRA-UHFFFAOYSA-N
27.	Meprobamate	C9-H18-N2-O4	NPPQSCRMBWNHMW-UHFFFAOYSA-N
28.	Thiethylperazine	C22-H29-N3-S2	XCTYLCDETUVOIP-UHFFFAOYSA-N
29.	5-Fluorouracil	C4H3FN2O2	GHASVSINZRGABV-UHFFFAOYSA-N
30.	Timolol	C13-H24-N4-O3-S	BLJRIMJGRPQVNF-JTQLQIEISA-N
31.	Treprostinil	C23-H34-O5	PAJMKGZZBBTTOY-ZFORQUDYSA-N
32.	Trihexyphenidyl	C20-H31-N-O	HWHLPVGTWGOCJO-UHFFFAOYSA-N
33.	Palonosetron	C19-H24-N2-O	CPZBLNMUGSZIPR-DOTOQJQBSA-N
34.	Dydrogesterone	C21H28O2	JGMOKGBVKVMRFX-HQZYFCCVSA-N
35.	Mexiletine	C11-H17-N-O	VLPIATFUUWWMKC-UHFFFAOYSA-N
36.	Dexrazoxane	C11-H16-N4-O4	BMKDZUISNHGIBY-ZETCQYMHSA-N
37.	Amlodipine	C20-H25-Cl-N2-O5	HTIQEAQVCYTUBX-UHFFFAOYSA-N
38.	Tacrine	C13-H14-N2	YLJREFDVOIBQDA-UHFFFAOYSA-N
39.	Oxyphencyclimine	C20-H28-N2-O3	DUDKAZCAISNGQN-UHFFFAOYSA-N
40.	Triamterene	C12-H11-N7	FNYLWPVRPXGIIP-UHFFFAOYSA-N
41.	Valstar	C34-H36-F3-N-O13	ZOCKGBMQLCSHFP-ZQUOIQDWSA-N
42.	Valrubicin	C34-H36-F3-N-O13	ZOCKGBMQLCSHFP-KQRAQHLDSA-N
43.	Procyclidine	C19-H29-N-O	WYDUSKDSKCASEF-UHFFFAOYSA-N
44.	Phenylephrine	C9-H13-N-O2	SONNWYBIRXJNDC-VIFPVBQESA-N
45.	Carbimazole	C7-H10-N2-O2-S	CFOYWRHIYXMDOT-UHFFFAOYSA-N
46.	Palonosetron	C19-H24-N2-O	CPZBLNMUGSZIPR-NVXWUHKLSA-N
47.	Lanoxin	C41-H64-O14	LTMHDMANZUZIPE-YUICGFAKSA-N
48.	Digoxin	C41-H64-O14	LTMHDMANZUZIPE-PUGKRICDSA-N
49.	Sulpiride	C15-H23-N3-O4-S	BGRJTUBHPOOWDU-UHFFFAOYSA-N
50.	Profenamine	C19-H24-N2-S	CDOZDBSBBXSXLB-UHFFFAOYSA-N

8.3 Валидација резултата (*Ground-truth*)

Задатак провере софтверске платформе је поверен истраживачима који су запослени у Лабораторији за ћелијску и молекуларну биологију. Тим доменских експерата чинили су хемичари и билози (докторанти и професори) који имају вишегодишње искуство у процедурама за синтетисање супстанци, у испитивању цитотксичности супстанци на одређеним ћелијама канцера, а у прилог томе говори више десетина објављених научних радова из ове области у светски реномираним часописима.

За стицање опште слике о свим скуповима података, истраживачи су се најпре упознали са структуром посматраних репозиторијума, притом су се упознали са подацима који се у њима чувају, а затим уочили важне термине и везе између њих. Сваки од коришћених линкова је ручно прегледан како би у потпуности било проверено да ли су гране (оне које су коришћене у графовима) између супстанци (чворова) реалне, тј. да ли постоји триплет који садржи одговарајуће ентитете за супстанце. Проверене су такође и позиције ентитета у триплету (субјекат или предикат) које су важне за финалне резултате, пошто оријентација грана одређује правац путање, а самим тим и редослед подупита у генерисаном упиту. У доста случајева постоје гране које имају обе орјентације, као на пример у случају супстанце Terfenadine у Bio2RDF репозиторијуму, где постоји грана од *kegg_ligand/D00521* до *drugbank/DB00342* и обратно.

За унапред дефинисане задатке (таргети, ћелијске линије и IC_{50} вредност) ручно су избројани релевантни скупови података који су обухваћени упитом. Такође, урађена је провера исправности и употребљивости упита, односно провера да ли добијени резултати за одговарајућу супстанцу одговарају постављеном питању. Како би се у потпуности елиминисале људске грешке, додатно је урађена провера добијених резултата коришћењем *PIBAS FedSPARQL* апликације [79] која је такође развијена у оквиру Лабораторије. Ова платформа омогућава извршавање предефинисаних (ручно креираних) Federated SPARQL упита над иницијалним и кориснички селектованим базама података. Као и у случају SpecINT платформе, ова платформа пружа подршку за извођење Federated SPARQL упита над више иницијалних извора података (*CPCTAS, Bio2RDF, Chem2Bio2RDF, EMBL-EBI*) уз могућност додавања кориснички изабраног скупа података.

8.4 Коришћене хеуристике

Иако је показано да знакови координата Фидлеровог сопственог вектора (видети Секцију 6.4) недвосмислено деле чворове графа са Слике 7.6 на два кластера (што у општем случају не мора да важи) и тако одређује репозиторијум којем чвор припада, и даље се не може са сигуношћу тврдити да ће нас PageRank вредност чвора одвести до жељених резулата. Другим речима, главни задатак који се намеће у овом тренутку је проналажење начина да се повежу две путање у артикулационом чвору (енг. *cut-vertex*), преко позитивних и негативних чворова, али да овај избор укључи што је могуће више релевантних скупова података за постављени упит, са оба репозиторијума. Као што је већ објашњено, артикулациони чвор служи као спона за прелазак са једног на други репозиторијум. Оно што је и био циљ развоја овакве софтверске платформе је избегавање претраге свих доступних скупова података као и свих њихових путања.

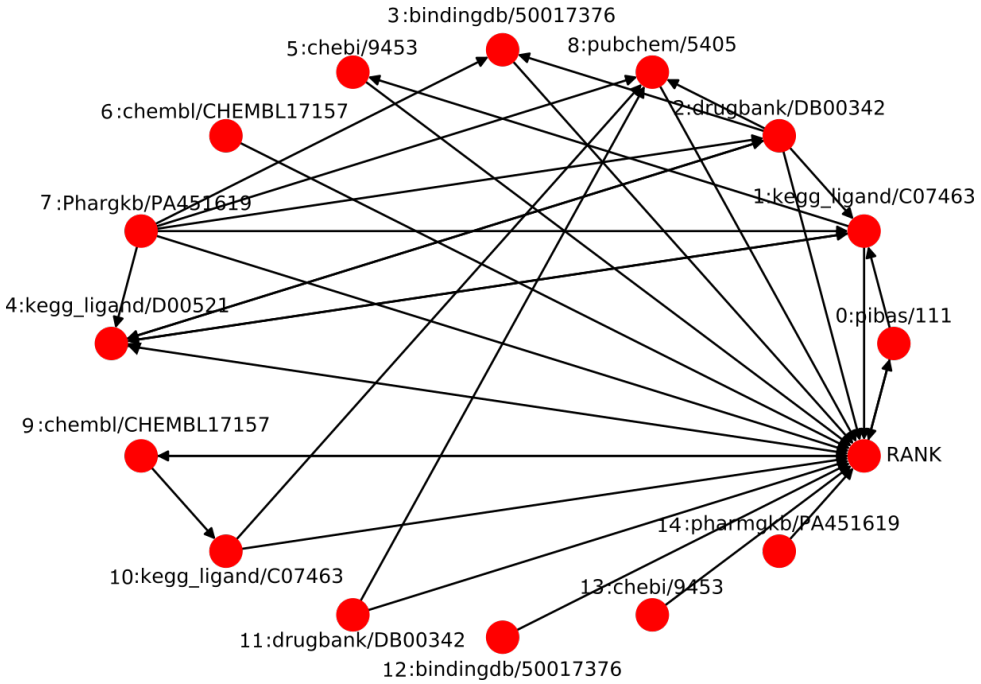
Исцрпљујућа претрага (енг. *brute-force search*) од истраживача захтева пуно времена, а ресурси постају превише оптерећени. Традиционални (енг. *state-of-the-art*) алгоритми за проналажење минимално разапињућег стабла за повезани тежински граф, као што су *Prim*-ов и *Kruskal*-ов алгоритам, се не могу применити у случају када тежине грана нису дефинисане и када одређене чворове треба фаворизовати. Из истих разлога, алгоритми за тражење најкраће путање између два чвора такође немају употребну вредност. Једна од могућих идеја је пребацивање фокуса на тежине грана и укључивање оних чворова који представљају крајеве грана са највећом тежином. Од ове идеје се након детаљне провере одустало, јер се постојеће гране разликују од хемијске структуре до хемијске структуре, а тежине се тешко могу интуитивно дефинисати. Такав проблем би се могао превазићи статистичком анализом скупова података и њихових триплета, али то је приступ који има бројне недостатке (стално рачунање нових вредности) и који је од самог почетка избегнут.

Прво питање које се намеће је на који начин изабрати путање, али тако да оне обухвате што је могуће више релеватних скупова података, а да губитак информација буде што мањи. За потребе тражења компромисног (*trade-off*) решења су предложене и тестиране три хеуристике базиране на степену чворова диграфа $D_{n,m}$ које се лако могу интегрисати у Алгоритам 3. У кораку када треба изабрати следећег суседа приликом тражења путање, нови чвор се може изабрати коришћењем једне од три предложене хеуристике, а које се базирају на:

- **Степену чвора:** бира се чвор(ове) са највећим излазним степеном $deg^{-}(v)$.
- **PageRank вредности чвора:** бира се чвор(ове) са највећом PageRank вредношћу. Ова хеуристика је због једноставности приказана у основном Алгоритму 3.
- **PageRank вредности фаворизованог чвора:** бира се чвор(ове) са највећом PageRank вредношћу која је кориснички навођена.

Концепт прве две хеуристике је јасан, зато их нећемо посебно описивати, док за трећу хеуристику треба дати додатно објашњење. Хеуристика *Фаворизовани PageRank* је уведена због навођеног (форсираног) избора чворова, који је потребан предуслов за фаворизовање одређених скупова података у зависности од задатог питања. За потребе ове хеуристике у граф се додаје још један фиктивни чвор са високим рангом. Нека је овај чвор означен са речју *RANK* (Слика 8.1). Његов високи ранг се може вештачки створити тако што ће сви чворови из $D_{n,m}$ имати грану ка њему. На овај начин он тако постаје битан чвор графа, али у исто време и утицајан, јер својом снагом може повећати важност других чворова додавањем гране ка њима. Претходно додате гране (ка *RANK* чвору) и гране од високо рангираног ка ниско рангираном чвору могу да повећају важност „слабијих" чвора без губљења информација о постојећој структури графа, јер су оне сачуване у првом вектору s. Ако је потребно да у упит укључимо слабо рангирани чвор, у ознаци *weak* (на пример, *PIBAS* или *CHEMBL* чворове), тада се у $D_{n,m}$ додаје нова грана од *RANK* ка *weak* чвору. Додатно, изабраним чворовима се могу додати петље (*self-loops*) које благо утичу на повећање њиховог ранга, за разлику од фиктивног чвора који то ради доста наглашеније. Овај приступ није коришћен у предложеном алгоритму, већ се само наводи као алтернатива коју треба додатно испитати. Доменско знање које се користи у овим ситуацијама је смештено у посебној онтологији која је развијена

за ову намену *RepoIntegration.owl*, а која је описана у Секцији 3.3. За сваки скуп података (репозиторијум) се чува његова афилијација, тематика којом се бави, кључни термини и друго. На основу постављеног питања се врши претрага онтологије, која служи као база доменског знања и из које се добијају информација о релевантним скуповима података. Те информације се затим користе за додавање одређених грана у графу које фаворизују изабране чворове (видети Табелу 8.2 која приказује ранг вредности за сваки чвор након фаворизације *PIBAS* и *Chembl* чворова).



Слика 8.1: Фаворизација PIBAS, Chembl и Kegg_ligand чворова.

Табела 8.2: PageRank вредности чворова пре и након фаворизације.

PageRank		**Фаворизовани PageRank**	
Скуп	**Ранг**	**Скуп**	**Ранг**
8:pubchem/5405	0.1453	0:pibas/111	0.1482
1:kegg_ligand/C07463	0.1416	9:chembl/CHEMBL17157	0.1482
4:kegg_ligand/D00521	0.1207	1:kegg_ligand/C07463	0.0886
2:drugbank/DB00342	0.0922	10:kegg_ligand/C07463	0.0724
10:kegg_ligand/C07463	0.0647	8:pubchem/5405	0.0492
3:bindingdb/50017376	0.0605	4:kegg_ligand/D00521	0.0395
5:chebi/9453	0.0951	5:chebi/9453	0.0345
0:pibas/111	0.0350	2:drugbank/DB00342	0.0219
6:chembl/CHEMBL17157	0.0350	3:bindingdb/50017376	0.0144
7:phargkb/PA451619	0.0350	6:chembl/CHEMBL17157	0.0094
9:chembl/CHEMBL17157	0.0350	7:phargkb/PA451619	0.0094
11:drugbank/DB00342	0.0350	11:drugbank/DB00342	0.0094
12:bindingdb/50017376	0.0350	12:bindingdb/50017376	0.0094
13:chebi/9453	0.0350	13:chebi/9453	0.0094
14:pharmgkb/PA451619	0.0350	14:pharmgkb/PA451619	0.0094
		RANK чвор	**0.3267**

Пример фаворизације и избора додатних чворова је дат у Табели 8.3. Поред чворова са ознаком 0 и 22, прво је фаворизован чвор *7:chebi*, чији се ранг истиче у први план, док се у другом случају то дешава са чвором *21:kegg_ligand*.

Табела 8.3: Фаворизација CHEBI и Kegg_ligand чвора.

CHEBI		**Kegg_ligand**	
Скуп	**Ранг**	**Скуп**	**Ранг**
5:chebi/9453	0.1302	**4:kegg_ligand/D00521**	0.1356
0:pibas/111	0.1101	**0:pibas/111**	0.0978
9:chembl/CHEMBL17157	0.1101	**9:chembl/CHEMBL17157**	0.0978
1:kegg_ligand/C07463	0.0707	1:kegg_ligand/C07463	0.0990
10:kegg_ligand/C07463	0.0562	10:kegg_ligand/C07463	0.0510
8:pubchem/5405	0.0420	8:pubchem/5405	0.0447
4:kegg_ligand/D00521	0.0342	2:drugbank/DB00342	0.0491
2:drugbank/DB00342	0.0204	5:chebi/9453	0.0374
3:bindingdb/50017376	0.0142	3:bindingdb/50017376	0.0191
7:phargkb/PA451619	0.0094	6:chembl/CHEMBL17157	0.0094
6:chembl/CHEMBL17157	0.0094	7:phargkb/PA451619	0.0094
11:drugbank/DB00342	0.0094	11:drugbank/DB00342	0.0094
12:bindingdb/50017376	0.0094	12:bindingdb/50017376	0.0094
13:chebi/9453	0.0094	13:chebi/9453	0.0094
14:pharmgkb/PA451619	0.0094	14:pharmgkb/PA451619	0.0094
RANK чвор	**0.3557**	**RANK чвор**	**0.3122**

8.4.1 Избор иницијалих чворова

Оно што треба посебно нагласити је избор почетних чворова од којих креће формирање путања. У овој докторској дисертацији је посебан акценат стављен на нове институције које желе да своје податке учине јавно доступним. Отуда су у већини приказаних примера почетни чворови управо такви скупови, а за потребе демонстрације методологије су искоришћени *PIBAS* и *Chembl* скупови. У случају да таквих скупова нема, алгоритам посебно треба да води рачуна да изабрани чворови никако не буду они који су изоловани. Ови чворови се могу појавити у $D_{n,m}$ након трансформација, тј. након операције у којој се из $V_{n,m}$ брише доста непостојећих грана.

У ситуацији када нема нових скупова података, у експериментима се показало да је за иницијални чвор најбоље изабрати чвор са просечним, а не највећим рангом, јер ће управо чворови са највећим рангом обезбедити међусобно повезивање скупова преко репозиторијума. Показало се да такви чворови, у улози артикулационих чворова, садрже велики број грана који иде од и ка њима и тако повећавају шансу за повезивање репозиторијума.

8.5 Резултати

У овој секцији су приказани добијени резултати за 50 изабраних супстанци које су дате у Табели 8.1. За сваку супстанцу алгоритам је покренут пет пута, за сваку хеуристику посебно, при чему су тражене одговарајуће путање које ће повезати два репозиторијума, *Chem2Bio2RDF* и *Bio2RDF*.

У Табели 8.4 је приказан заокружени просечан број релевантних скупова података који су обухваћени упитом за сваку хеуристику, као и праве вредности (енг. *ground-truth*) до којих се дошло ручном провером. Приказани резултати у табели се односе само на питање о таргетима на које утичу лекови, пошто су резултати за ћелијске линије и IC_{50} вредност слични и нема их потребе посебно приказивати. Треба имати у виду да су за потребе евалуације у обзир узети само скупови података који садрже релевантне податаке (везане за постављено питање), иако крајњи упит том приликом

може обухватити и друге скупове података.

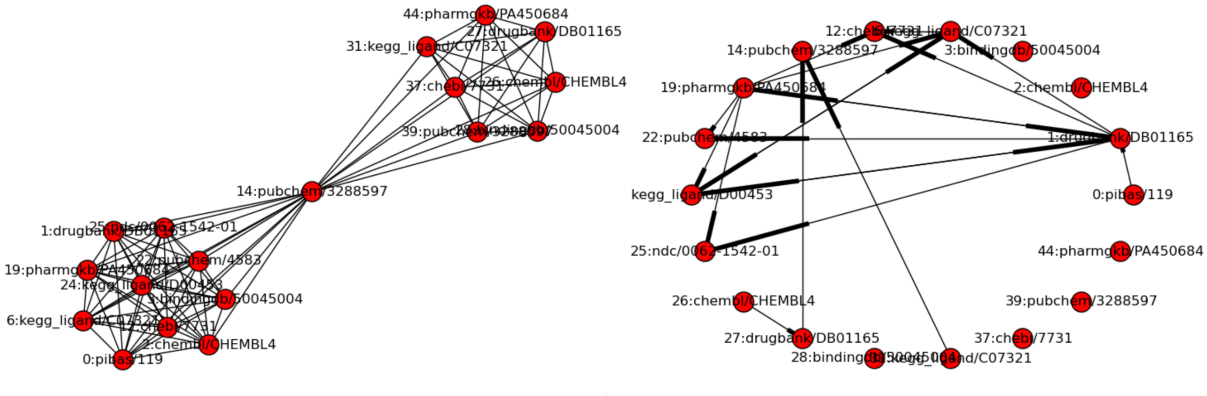
Табела 8.4: Број изабраних релевантних скупова података за таргете, за сваку хеуристику посебно.

InChIKey	Степен	PageRank	Фав. PageRank	Тачна вредност
WNMJYKCGWZFFKR-UHFFFAOYSA-N	4	4	6	6
IRYJRGCIQBGHIV-UHFFFAOYSA-N	0	4	6	7
MHWLWQUZZRMNGJ-UHFFFAOYSA-N	0	0	4	7
CXOXHMZGEKVPMT-UHFFFAOYSA-N	0	4	6	7
MJFJKKXQDNNUJF-UHFFFAOYSA-N	4	4	6	6
GUGOEEXESWIERI-UHFFFAOYSA-N	0	0	5	6
GSDSWSVVBLHKDQ-UHFFFAOYSA-N	0	0	0	8
PTOAARAWEBMLNO-KVQBGUIXSA-N	4	4	5	6

Из резултата се може приметити да је најефикаснија хеуристика управо хеуристика *Фаворизовани PageRank*. Укључивање доменског знања побољшава коначне резултате алгоритма, што нам говори да се алгоритам може и даље унапређивати новим доменским знањем које је посебно значајно када су непознати ресурси интегрисани у читав систем. Ретки су случајеви у којима *Degree* хеуристика успева да повеже путање преко репозиторијума. Чак и када је то случај, упит обухвата мали број релевантних ресурса. Такође, овај приступ није подесан из практичних разлога, јер доводи до великог гранања и генерисања већег броја упита. Гранање упита доводи до успорености платформе, јер велики број упита подразумева њихово извршавање и додатну евалуацију. За разлику од *Degree* хеуристике, *PageRank* хеуристика даје доста боље резултате по питању избора релевантних скупова, укључујући и већи проценат успешности у повезивању путања. Бољи резултати се објашњавају чињеницом да су најбоље рангирани чворови повезани са многим чворовима, па је самим тим и повезивање путања лакше. Основни недостатак овог приступа је немогућност укључивања нових скупова података који имају мали број веза, па самим тим и мали ранг. На пример, супстанца из Лабораторије је повезана само са једном супстанцом из скупа *KEGG* и као таква је неприметна за алгоритам за разлику од оних које потичу из великих иницијатива. Такође, овај приступ нема могућност обухватања најрелевантнијих скупова података за свако постављено питање, а о чему је већ било речи. Након додатне провере добијених резултата показано је да се коришћењем SpecINT софтверске платформе (*Фаворизовани PageRank* хеуристика) добија прецизност од 86% у покривању најрелевантнијих скупова података за таргете лекова. Прецизност од 71% и 75% је постигнута за ћелијске линије и IC_{50} вредност, редом. Важно је напоменути да резултати могу да варирају од тренутка до тренутка што зависи од доступности одговарајућих ендпоинта. Другим речима, неке гране у графу могу бити обрисане или додате тако да редослед повезивања подупита може бити нешто другачији.

Поред избора релевантних скупова, тестирана је и исправност тако генерисаних упита, јер велики број изабраних скупова не гарантује да ће се путање повезати у артикулационом чвору. Иако коришћењем графа добијамо механизам за обилазак и претраживање репозиторијума, некада се дешава да не постоји грана која повезује последњи чвор у путањи са артикулационим чвором. Чак и ако таква грана постоји, то не значи да је њена оријентација одговарајућа. У око 13% таквих случајева упити не могу бити направљени. Тада је неопходно спровести додатне модификације алгорит-

ма како би у обзир били узети и овакви случајеви, што додатно усложњава алгоритам. На пример, за хемијску структуру којој одговара InChIKey = GSDSWSVVBLHKDQ-UHFFFAOYSA-N путање је немогуће повезати у чвору *14:pubchem* (Слика 8.2). Самим тим ни резултати не могу бити враћени на овај начин, већ се претраживање мора спровести ручно, тј. упит по упит за сваки скуп посебно. За избор артикулационог чвора (чворова) који се користи за повезивање два графа узети су они скупови око којих су изграђена оба репозиторијума. Избор таквих скупова, који су поменути у кораку 1 ($Intersection = \{R_1 \cap R_2\}$) Процедуре 1, је ограничен на скупове *PubChem*, *Kegg*, *Chebi* и *DrugBank*. Ово свакако није ограничавајући фактор платформе, а овакво ограничење се нужно уводи због чињенице да се иницијативе увек концентришу око 3-4 скупа података који су у сфери интересовања, а онда све друге скупове интегришу са њима.



Слика 8.2: Граф и дограф за InChIKey = GSDSWSVVBLHKDQ-UHFFFAOYSA-N.

Као додатак случајевима коришћења из Секције 1.2, за сваку хемијску структуру (Ид) је примењена *Фаворизовани PageRank хеуристика.* Табела 8.5 приказује број враћених резултата за таргете, ћелијске линије и IC_{50} вредности за сваки ресурс посебно. У ширем контексту, ова табела даје кратак преглед локација на којима су супстанце тестиране и где се додатни резултати, уколико су потребни, могу пронаћи.

Табела 8.5: Број добијених резултата за изабране супстанце. За сваку супстанцу је приказан број пронађених таргета (target), ћелијских линија (CL) и IC_{50} вредности, за сваки од репозиторијума, укључујући и нове скупове података CHEMBL и CPCTAS.

	Bio2RDF			**Chem2Bio2RDF**			**CHEMBL**			**CPCTAS**		
Ид	*Target*	*CL*	IC_{50}	*Target*	*CL*	IC_{50}	*Target*	*CL*	IC_{50}	*Target*	*CL*	IC_{50}
1.	4	0	0	0	0	0	32	2	6	1	1	1
2.	1	0	0	0	0	0	129	0	0	1	1	1
3.	1	0	0	0	0	0	161	0	0	1	1	1
4.	1	0	0	0	0	0	7	0	0	1	1	1
5.	5	0	0	0	0	0	2	0	0	1	2	4
6.	7	0	0	2	0	0	210	10	29	1	1	1
7.	3	0	0	0	0	0	260	2	2	1	1	1
8.	10	0	0	2	0	0	170	24	27	1	1	1

8.6 Поређење са другим платформама

У овој секцији је приказно поређење SpecINT софтверске платформе са другом великом платформом, *Open PHACTS (Open Pharmacological Concept Triple Store)* [44], коју користе фармацеутске компаније. *Open PHACTS* је европска иницијатива, основана од стране *Innovative Medicines Initiative*, чији је циљ да омогући јефтиније и брже откривање нових лекова. *Open PHACTS* интегрише податке из скупова података као што су *ChEMBL*, *neXTProt*, *ChemSpider* и *DrugBank* да би се креирало више од 3 милијарде семантички повезаних триплета из области природних наука. Иако је *SpecINT* софтверска платформа *proof-of-concept* пројекат, дизајнирана да ради са више репозиторијума без заједничке онтологије између њих, она има способност да врати најновије резултате, али и да интегрише нове скупове података. Ресурси који се користе у евалуационом процесу за *SpecINT* су приказани у Секцији 3.2 (укључујући и њихове верзије). Две платформе су упоређене са аспекта враћених резултата за унету супстанцу, при чему је главни циљ овог поређења да се покаже да Federated SPARQL упити добијени помоћу платформе могу да пруже комплементарне резултате у односу на резултате који су добијени помоћу *Open PHACTS Discovery Platform* [45]. За потребе поређења посматрана су као и раније три задатака: 1) таргети, 2) ћелијске линије, и 3) одговарајуће IC_{50} вредности. Евалуација је обухватила 24 упита, по 8 за сваки задатак, при чему су упити за *SpecINT* генерисани на начин који је раније описан у Секцији 7.5. *Open PHACTS* платформи се може приступити на два начина, преко API позива или преузимањем комплетног RDF фајла са свим подацима над којим се затим извршавају класични SPARQL упити. У духу циљева софтверске платформе које је овде описана, изабран је рад са rdf фајловима. Пример SPARQL упита који приказује све таргете за CXOXHMZGEKVPMT-UHFFFAOYSA-N хемијску структуру је дат у Листингу 8.1. У следећем тексту је приказана и објашњена разлика у резултатима који су добијени помоћу ове две платформе.

За сваку InChIKey вредност која је коришћена као улаз у *SpecINT*, пронађен је одговарајући URI супстанце који је коришћен у *Open PHACTS* платформи. Експерти из Лабораторије су проверили враћене резултате са обе платформе и спровели њихово поређење. Један део тестираних супстанци и враћених резултата за таргете су приказани у Табели 8.6. Може се приметити да је разлика у броју враћених резултата релативно мала. У враћеним резултатима постоји велико преклапање, јер обе платформе обухватају *ChEMBL* скуп података. За питање таргета *SpecINT* пружа комплементарне резулатате који су добијени из *DrugBank* скупа (и са *Bio2RDF* и са *Chem2Bio2RDF* репозиторијума), као и *PIBAS* онтологије. *Open PHACTS API* није вратио ове додатне резултате иако је *DrugBank* као скуп укључен у платформу, па отуда потиче та мала разлика у резултатима. Преклапања две платформе су доста велика и у случају ћелијских линија и IC_{50} вредности, пошто обе платформе обухватају *EBI-RDF ChEMBL* и *DrugBank* скупове података. Међутим, за хемијску структуру којој одговара InChIKey=GSDSWSVVBLHKDQ-UHFFFAOYSA-N, *Open PHACTS* пружа више резултата за тестиране ћелијске линије него *SpecINT*. Нови скупови у *Open PHACTS* иницијативи се стално додају, а стари обнављају због поседовања великих рачунарских ресурса, што се за *SpecINT* не може увек гарантовати. Додатно, сви добијени резултати су јавно доступни на figshare репозиторијуму[1].

[1] https://figshare.com/articles/Evaluation/6352496

```
PREFIX  rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX  owl:  <http://www.w3.org/2002/07/owl#>
PREFIX  xsd:  <http://www.w3.org/2001/XMLSchema#>
PREFIX  rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX  chembl_molecule: <http://rdf.ebi.ac.uk/resource/chembl/
   molecule/>
PREFIX  cco:  <http://rdf.ebi.ac.uk/terms/chembl#>

SELECT (count(distinct ?target_op) as ?count_op) (count(distinct ?
   target_specint) as ?count_specint) (count(distinct ?matching) as
    ?count_matching)
FROM <http://cpctas-lcmb.pmf.kg.ac.rs/2012/3/PIBAS/CXOXHMZGEKVPMT-
   UHFFFAOYSA-N.rdf>
WHERE
  { ?target_op ?p ?o.
  FILTER(CONTAINS(str(?target_op), "http://rdf.ebi.ac.uk/resource/
     chembl/target/")).

    SERVICE SILENT <https://www.ebi.ac.uk/rdf/services/sparql/>
          { ?activity rdf:type cco:Activity .
            ?activity cco:hasMolecule chembl_molecule:CHEMBL70418.
            ?activity cco:hasAssay ?assay .
            ?assay cco:hasTarget ?target_specint.
          }

    BIND(IF(?target_specint=?target_op,?target_specint,"") AS ?
       matching)
  }
```

Листинг 8.1: Претрага таргета за лек *Clobazam* (познат још под именима *Onfi* и *Sympazan*).

На основу добијених резултата, закључујемо да оба приступа пружају добру полазну основу за претраживање података од интереса. Иако су разлике између репозиторијума мале, обе платфоме могу да понуде комплементарне резултате истраживачкој заједници што је и био циљ овог поређења. Другим речима, квантитатива разлика у враћеним резултатима није толико битна, колико да се покаже да обе платформе могу да дају допринос и поспеше даља истраживања и да је сарадња између институција на глобалном нивоу неопходна. Због тога ће у неком тренутку сви резултати нових експеримента који се добију у Лабораторији бити у потпуности доступни истраживачкој заједници кроз *PIBAS* онтологију. Такође, и *Open PHACTS* и *SpecINT*, пружају могућност за интеграцију нових скупова података тако је простор за даљу сарадњу на вишем нивоу већ отворен. Ако се пореде додатни алати које пружају обе платформе, *Open PHACTS Discovery Platform* их поседује знатно више као што је и очекивано. Због величине иницијативе, броја укључених институција (27) и истраживача, поред интеграције података, *Open PHACTS* је доста пажње посветио графичком интерфејсу и визуелизацији резултата, као и алатима који олакшавају претрагу и доношење даљих одлука.

Табела 8.6: Број добијених резултата. Један део тестираних кључева за претрагу таргета лекова.

Ид	InChIKey	OpenPhact	SpecINT
1.	WNMJYKCGWZFFKR-UHFFFAOYSA-N	32	40
2.	IRYJRGCIQBGHIV-UHFFFAOYSA-N	129	132
3.	MHWLWQUZZRMNGJ-UHFFFAOYSA-N	161	163
4.	CXOXHMZGEKVPMT-UHFFFAOYSA-N	7	10
5.	MJFJKKXQDNNUJF-UHFFFAOYSA-N	2	13
6.	GUGOEEXESWIERI-UHFFFAOYSA-N	210	220
7.	GSDSWSVVBLHKDQ-UHFFFAOYSA-N	260	267
8.	PTOAARAWEBMLNO-KVQBGUIXSA-N	170	183

Глава 9

Дискусија

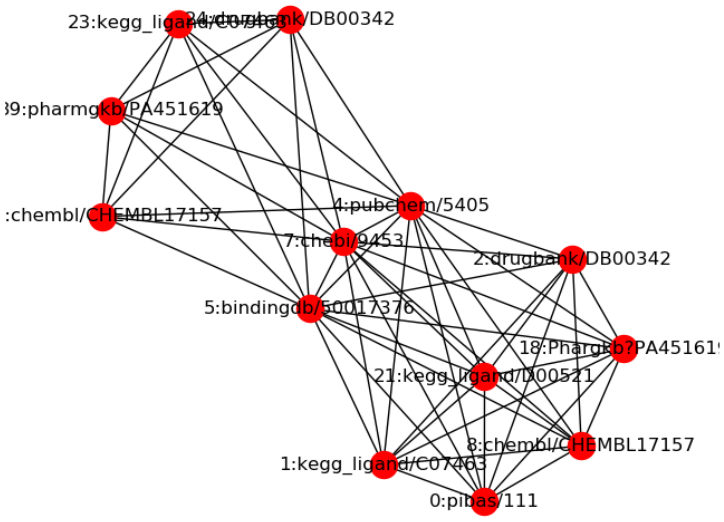
Због јасније слике о самим принципима рада софтверске платформе, у претходном тексту се није обраћала посебна пажња на начине на које се графови могу повезати. Тренутни приступ, повезивање графова помоћу једног чвора (слепљивањем), се показао као приступ који нам доноси највише бенефита у контексту броја релевантних скупова података, исправности упита и броја враћених резултата. У овом поглављу посебно ће бити прокоментарисани различити начини повезивања графова и објашњене разлике које они доносе у односу на тренутни приступ. Такође, у овом поглављу биће речи и о графовима који нису комплетни, а који се могу проследити као улаз у алгоритам, и променама које они у алгоритму доносе. На крају поглавља биће дат преглед ограничења софтверске платформе, као и предлози за њено будуће унапређивање.

9.1 Различити начини повезивања графова

У кораку 6, Процедуре 1, повезивање графа се може извести на више начина. На пример, поред спајања (слепљивања) два чвора у један чвор, као што је то урађено на Слици 7.4, спајање два графа се може урадити и по више чворова истовремено (Слика 9.1). У овом случају за спајање два графа су искоришћени чворови *5:bindingdb/50017376*, *7:chebi/9453* и *4:pubchem/5405*.

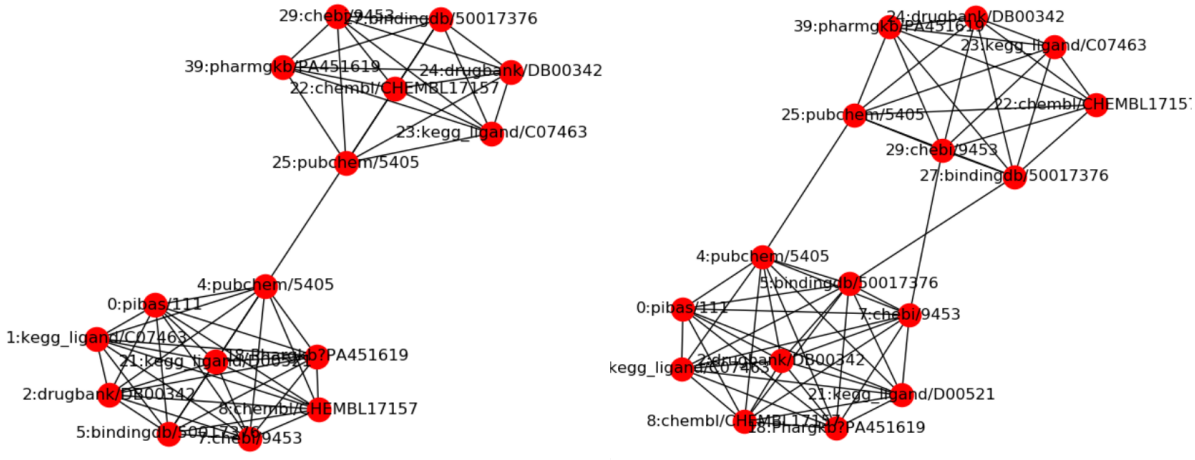
Суштина алгоритма остаје непромењена, јер Теорема А.4 даје одговор на питање који чвор припада којем репозиторијуму. Међутим, основни недостатак овог приступа је све мањи број чворова ван пресека који смањује вероватноћу спајања путања у једном чвору. Додатно, велики број чворова у пресеку са великим степеном повећава могућност за мимоилажење путања са две стране приликом њиховог спајања. Зато је овај приступ избегнут и замењен оним приступом који је представљен у Алгоритму 3, где се повезивање графова увек ради преко једног чвора. У случају да путање не могу да се повежу, конструише се нови граф (у петљи) који ради слепљивање по другом чвору. Другим речима, боље решење представља петља у којој се графови увек повезују преко једног чвора, него да се повезивање графова ради у више чворова одједном и тако ризикује фаза спајања путања.

На другој страни, поред повезивања графова помоћу чворова, два графа се могу повезати и помоћу гране (грана) (видети Слику 9.2). Као и до сада, за два „иста” чвора која



Слика 9.1: Повезивање два графа слепљивањем више чворова.

су повезана граном циљ је направити путање са обе стране управо до тих чворова. Ако се мало боље сагледају ствари, овакав начин повезивања је само једна варијанта онога што је већ и урађено. Међутим, једна битна разлика је та што када имамо повезивање преко једног чвора, тај чвор омогућава ограничавање у броју враћених резултата (међурезултата) у наредном обрасцу (шаблону графа) који је позван, док се на овај начин број враћених резултата повећава. У сарадњи са истраживачима из Лабораторије донесен је закључак да је повезивање преко једног чвора ипак оптималније решење са аспекта брзине и количине враћених резултата.



Слика 9.2: Повезивање два графа помоћу једне или више грана.

Даље, два графа се могу повезати и граном која спаја чворове који не припадају истом скупу података. На пример, грана се може додати између чворова *KEGG* (репозиторијум 1) и *DrugBank* (репозиторијум 2), при чему се на овај начин занемарује ограничавање претраге у наредном подупиту, а смањење међурезултата је једна од

битних предности коју платформа треба да пружи. На исти начин се могу додати и друге гране, као што је то био случај са чворовима у пресеку, али се из истог разлога увек преферира повезивање било преко једног чвора, било преко једне гране.

9.2 Графови који нису комплетни

Чињеница на којој се доста потенцира и која се провлачи од самог почетка докторске дисертације је рад са комплетним графовима. Природно питање које се намеће је да ли почетни графови могу да буду некомплетни и како би онда алгоритам изгледао. Одговор је потврдан. Природа проблема је довела до тога да се све време ради са комплетним графовима, јер је циљ добијање што је могуће више релевантних података који су везани за само једну супстанцу. Супстанца која се претражује је представљена у више репозиторијума, појављујући се у више скупова података истовремено, али под различитим именом. Како је увек реч о истој супстанци, отуда и комплетан граф који се посматра. Ово свакако није једина опција коју можемо да очекујемо током претраживања репозиторијума. На пример, некада је потребно пронаћи све супстанце које су тестиране на истој ћелијској линији и које су дале добре резултате. Добијени граф неће бити комплетан, јер су супстанце различите, и тада треба наћи одговарајућу метрику на основу које одлучујемо да ли су два чвора повезана или не.

Најчешћи случај претраживања репозиторијума са графовима који нису комплетни се јавља у ситуацији када се посматрају сличне, а не исте супстанце. Ово је случај који је од посебног интереса за моделе квантитативног односа структуре и активности (енг. *Quantitative structure–activity relationship*), који су још познати и под називом QSAR модели. QSAR модели су регресиони или класификациони модели који се користе у хемији и биологији када је потребно установити везе које постоје између хемијске структуре супстанце и њене активности коју је показала у одређеним експериментима. Када се ради са комплетним графовима довољно је израчунати један сопствени вектор матрице суседства графа и тада на основу њега одредити припадност скупа репозиторијуму. Међутим, када имамо некомплетне графове или више репозиторијума које треба повезати, тада је најбоље користити неку од верзија алгоритма спектралног кластеровања који су описани у Подсекцији 7.4.1. Тада се за одређивање који чвор припада којем кластеру користи више сопствених вектора Лапласове матрице графа, који се даље користе као улаз у k-means алгоритам за кластеризацију података.

Пример креирања *QSAR* модела и његова примена се могу видети у раду [114]. Аутори рада су описали процедуру креирања предикционог модела који треба да предвиди цитотоксичне активности 17-пиколил и 17-пиколинилиден андростанских деривата на ћелијској линији канцера простате (PC-3) са негативним андрогенским рецептором. За креирање модела коришћене су вештачке неуронске мреже (енг. *Artificial Neural Networks*) на бази молекулских дескриптора. Коришћењем *stepwise selection* приступа у комбинацији са *partial least squares* методом показано је да су *skin permeability (SP)*, *Madin–Darby canine kidney cell permeability (MDCK)* и *universal salt solubility factor (S + SF)* најважнији предиктори у предложеном моделу. Моделовање је спроведено помоћу неуронских мрежа како би се добио реални модел који може да олакша наредне синтезе андростанских деривата са високом антипролиферативном активношћу у односу на PC-3 ћелијску линију. Касније ће у оквиру закључка бити описани конкретни

проблеми који ће се у склопу будућих истраживачких активности решавати на сличан начин.

Поред регресионих и класификационих проблема који су примарни у QSAR анализи, могуће је урадити и кластеризацију одређеног скупа хемијских структура на основу различитих параметара. Најчешће се кластеризација ради на основу бинарног потписа молекула (енг. *molecular fingerprints*), особина графова (молекулски дескриптори) или максималних заједничких подструктура. Молекуларни потписи представљају начине енкодирања структуре молекула. Најчешће коришћени потпис је низ бинарних цифара (битови) које представљају појављивање одређене подструктуре у молекулу. Уопштено говорећи кластеризација се користи за детектовање група једињења која се различито понашају, али и за лакше разумевање понашања једињења у оквиру групе. Кластеризација супстанци такође може бити прилично корисна када је потребно у датом скупу једињења детектовати нетипичне вредности (хемијске структуре које испољавају другачију активност од очекиване). Примери различитих алгоритама кластеризације и њихове примене у области хемије и биологије се могу видети у [115, 116, 117].

9.3 Непотпуна кластеризација

Када је реч о повезивању комплетних графова, као што је то рађено кроз читаву дисертацију, ту су ствари прилично јасно дефинисане што се тиче одређивања припадности чворова репозиторијумима. Фидлеров вектор у потпуности одређује припадност сваког чвора репозиторијуму, као и који чвор се користи за њихово повезивање, јер је структура графа таква да се постиже потпуна тачност (на основу Теореме 6.15). Фидлеров вектор заправо одређује три кластера: један је одређен позитивним координатама, други негативним, а трећи нултом координатом (кластер са само једним чвором).

Међутим, ситуација постаје нешто компликованија када се у екперименти ма користе графови који нису комплетни. Тада само Фидлеров вектор није довољан за одређивање припадности чворова кластерима, већ је потребно применити неки од алгоритама спектралног кластеровања. У овом случају, ако је задати број кластера k једнак броју 2 (два репозиторијума), након примене k-means алгоритма на одређене векторе као резултат се добијају тачно два кластера. Тада корисник мора да води рачуна о томе који чвор је искоришћен за слепљивање графова и да ту додатну информацију проследи Алгоритму 3.

Оно што додатно може да забрине је могућност лоше кластеризације, тј. да су неки чворови додељени погрешном кластеру. Другим речима, ако је један скуп података додељен погрешном репозиторијуму, за њега ће бити употребљен неадекватан подупит (шаблон). Иако ће се ово ретко када дешавати због самог изгледа графа, овакав сценарио је ипак могућ. Једини начин да се проблем реши у овом тренутку је да се алгоритам мало прилагоди новој ситуацији. Ако се након кретања од чвора до чвора унутар једног кластера примети да постоји грана ка чвору који припада другом кластеру (грана која повезује два чвора из различитих кластера), тада треба урадити додатну верификацију постојања ове гране, па самим тим и њеног почетног и крајњег чвора. Након урађене провере се одређује који подупит је валидан избор. Ова провера ће незнатно да увећа време извршавања упита, јер се увек ради о једном или два чвора.

9.4 Мање познате хемијске структуре

Графови се разликују од супстанце до супстанце, а могу се разликовати и од временског тренутка до тренутка. Претходна евалуација софтверске платформе је урађена само са супстанцама чије се информације чувају на оба репозиторијума, јер је ово неопходан предуслов за њихово повезивање. Међутим, ово није увек случај, јер се мање познате супстанце могу јавити само у једном репозиторијуму или скупу. Постоје два објашњења за постојање оваквог случаја. Прво, такве супстанце су показале јако лоше цитотоксично дејство на неким ћелијама канцера, тако да друге лабораторије нису наставиле истраживања. Или друго, супстанце су се тек појавиле на тржишту и нова испитивања се тек очекују. Иако су овакви случајеви ретки, ипак су могући и зато су кориснику приказани само подаци из скупова чија тематика одговара постављеном питању. Пример таквих супстанци су Phenylethyl alcohol (InChIKey=WRMNZCZEMHIOCP-UHFFFAOYSA-N) који се налази само у *Chembl*-у, или Prazosin (InChIKey=IENZQIKPVF-GBNW-UHFFFAOYSA-N) који се налази само у *PubChem*-у.

Платформа је направљена тако да може да функционише и када су информације о супстанци интегрални део само једног репозиторијума. У таквим случајевима путању је могуће креирати само на основу PageRank вредности чворова, али овај приступ губи смисао, јер се супстанца најчешће налази у највише два скупа.

9.5 Ограничења софтверске платформе

У овом делу су набројани недостаци тренутног решења и на основу њих дате директиве за будући рад.

Ендпоинти нису активни: SpecINT доста зависи од досупности SPARQL ендпоинта. Често локалне копије података није могуће направити, јер се величина скупова података мери у стотинама гигабајта. Генерисани упити могу да прескоче ендпоинте који нису активни користећи кључну реч *SILENT*, међутим то није највећи проблем, јер на овај начин неке гране могу да буду избрисане из диграфа (недоступан ендпоинт) и самим тим спрече формирање путање.

Избор артикулационог чвора: Избор различитог артикулационог чвора може да смањи или повећа број изабраних чворова у путањама репозиторијума. Тражење trade-off решења између избора одговарајућег чвора и ”повољне” путање је тежак задатак, што су експерименти и показали. Иако је разлика у резултатима незнатна, будући рад би требало да обухвати и тражење решења које доводи до најбољих перформанси за дати граф. Ово ће свакако захтевати мало другачије кораке препроцесирања података и дуже време извршавања.

Путање не могу да се споје: Проблем настаје оног тренутка када путање са два репозиторијума није могуће повезати, тј. када се оне не споје у одређеном чвору. Тада се претрага одвија независно за сваки репозиторијум, а резултати се спајају и у таквом облику приказују. Овај начин претраге, који је до сада био стандард, као одговор доноси доста велики број резултата, као и спорије извршавање целокупне процедуре. У екстремним случајевима, и овај начин претраге може да буде користан.

Два репозиторијума: Основни недостатак тренутне верзије платформе је рад са два репозиторијума. Сличне теореме које су показане за два, могу се показати да важе за више репозиторијума. Ипак, ово проширење на светло доноси нове изазове који се, на пример, тичу избора артикулационих чворова између репозиторијума. Такође, очекује се да репозиторијуми буду сличне тематике и имају неке заједничке скупове података како би се лакше урадило њихово повезивање. Ипак, ови приступи нису тестирани у реалним ситуацијама због репозиторијума који често нису били доступни, а један од њих је већ поменути *LOOD*.

„Свежи" подаци: Добијене резултате које на излазу даје Платформа увек треба узети са резервом, јер се дешава да поједини скупови података нису допуњени новим подацима. На пример, *Chem2Bio2RDF* и *Bio2RDF* иницијативе користе податке из *PubChem RDF* базе података [118] која се мења више пута на дневном нивоу. Ове иницијативе освежавање података не раде тако често због превеликог обима посла.

Глава 10

Закључак

Главни резултат ове докторске дисертације је софтверска платформа, *SpecINT (Spectral Integration)* [119]. *SpecINT* представља скалабилно софтверско решење које користи предности које пружају технологије Семантичког Веба и резултати спектралне теорије графова за интеграцију и претраживање више репозиторијума истовремено. Методологија која је коришћена за постизање пуне функционалности Платформе се заснива на сопственим векторима графа који се користе за избор релевантних скупова података и надовезивање шаблона (енг. *patterns*), што омогућава смањивање броја дупликата и брисање непотребних информација у враћеним резултатима. Сврха платформе је да унапреди рад истраживачких тимова у области природних наука, пре свега биолошких и хемијских наука, тако да је највећа пажња приликом њеног развоја била усмерена ка добијању адекватних резултата на постављени упит. Платформа је развијена у сарадњи са истраживачима Лабораторије за ћелијску и молекуларну биологију, и она пре свега иде у сусрет њиховим захтевима, али се може посматрати и кроз много шири контекст, јер се базира на опште прихваћеним стандардима и концепту „Отворених података".

У ери Интернета и велике количине података (енг. *Big Data*), природно окружење у којем се подаци чувају је дистрибуирано што, поред стандардних изазова, додатно отежава процес њихове интеграције и претраживања. Да би одређени подаци били доступнији истраживачкој заједници, најпре их све треба интегрисати у једну или више целина, а онда ту целину учинити доступном за претраживање. Тако интегрисани скуп података треба да повезује све ентитете између којих постоји веза, било да је она слаба или јака, и ближе опише односе између њих како би се омогућило доношење закључака на основу њих. Због константног генерисања нових података, креирање свих веза је јако дуг процес који захтева пуно ангажовање свих страна, од ИТ стручњака до доменских експерата. Процес интеграције података је по природи ствари итеративни процес, јер је неопходно наставити са даљим увезивањем података, али и ревидирањем постојећих веза које су раније креиране. Креирана Платформа представља компромисно решење које процес повезивања података убрзава, користећи само мапирање између кључних ентитета, али и доменско знање које ће ићи у сусрет постављеним захтевима на улазу. Додатно, поред повезивања ентитета, интеграција два велика репозиторијума која су формирана за различиту намену је један од циљева ове докторске дисертације. На овај начин се ствара виртуелни репозиторијум који ствара привид да

се претраживање обавља над једним складиштем података. У тренутку писања ове докторске дисертације, једино познато теоријско решење које обрађује проблем интеграције више репозиторијума је описано у раду [57]. У раду је предложена архитектура за претраживање дистрибуираних RDF складишта проширивањем постојећег *SESAME* система. Такође, предложене су одређене структуре за индексирање података, као и алгоритми за обраду и оптимизацију упита у тако дистрибуираном окружењу.

Најпопуларнија решења у литератури која се баве тематиком претраживања семантички базираних скупова се заснивају на статистичким прорачунима. Полазећи од претпоставке да су подаци већ интегрисани, ова решења за потребе претраживања у обзир узимају број појављивања одређених ентитета или предиката, као и начине на који су они повезани, било да се ради о директним или индиректним везама. Међутим, таква решења су креирана за аутоматско генерисање Federated SPARQL упита чије се постојање везује искључиво за брзо креирање бенчмарк упита за проверу брзине извршавања различитих стратегија. Готово да се нигде у њиховим циљевима не помињу теме као што су интеграција података, корисничко искуство и резултати који се добијају након њиховог извршавања. Оно ка чему се једним делом ишло развојем ове Платформе је увођење човека (доменског знања) као саставног дела генератора упита, а са циљем да се на излазу добију валидни упити који након извршавања треба да врате употребљиве и релевантне резултате за постављено питање. Ручно навођење процесом креирања упита треба да донесе велике бенефите у односу на цену која је плаћена додатним ангажовањем експерта.

10.1 Постигнути резултати

Основна идеја око које се Платформа развијала, у недостатку сличних решења, је интеграција и претраживање семантички базираних репозиторијума који су од велике важности за истраживачку заједницу. На овај начин се могу уштедети ресурси, време постаје битан фактор о којем се води рачуна, а тражене информације доступне у сваком тренутку. На пример, једна од функционалности коју Платформа пружа је лак и брз приступ информацијама о супстанцама које су добар инхибитор за одређене ћелијске линије канцера, тј. чија IC_{50} вредност (концентрација која инхибира 50% ћелија) задовољава прописана ограничења у пре-клиничкој фази испитивања. Такође, поред увида у тако интегрисане податке, подаци се потенцијално могу искористити за прављење математичких модела који омогућавају да се предвиди IC_{50} вредност (регресија) за нову супстанцу, да се сличне супстанце поделе у кластере према структури или резултатима (кластеризација), или пак да се одређене супстанце доделе унапред задатим класама (класификација).

Поред на почетку постављених и остварених циљева који су описани кроз ову дисертацију, остаје да се још једном таксативно помену постигнути резултати и наведе шта је све остало доступно истраживачима након развоја Платформе.

- Након што је Платформа развијена, једна од додатних вредности која је остала након тога је креирана база података са одговарајућим подупитима (енг. *patterns*), за сваки скуп података у оквиру репозиторијума. Ова база података омогућава креирање валидних и сврсисходних Federated SPARQL упита. У духу претход-

но поменутих технологија, развијена је онтологија *RepoIntegration.owl* са краћим описом сваког скупа података која нам пружа могућност прецизнијег претраживања репозиторијума у зависности од постављеног питања.

- Архитектура на којој је равијена Платформа је скалабилна и конструисана на такав начин да омогућава лаку интеграцију нових скупова података. На овај начин друге институције могу веома једноставно да презентују и учине јавно доступним своја истраживања. Додавање нових и брисање старих скупова података је веома једноставно и не утиче на рад Платформе, нити је потребно радити њено додатно конфигурисање.

  На пример, за потребе Лабораторије, сви подаци из базе података су мигрирани у складиште RDF података у чијој се основи налази PIBAS (*Preclinical Investigation of BioActive Substances*) онтологија која осликава начин рада Лабораторије. Комплетна структура нове, семантички базиране базе података, је представљена научној заједници на међународним конференцијама и детаљно описана у научним радовима [49, 50].

- У оквиру Платформе је развијена процедура за конструисање графова чији су чворови различите репрезентације исте (сличне) супстанце у репозиторијумима, док су гране дефинисане постојећим триплетима у RDF графу који повезују те исте репрезентације супстанце. Представљање података у облику графа омогућава примену различитих резултата спектралне теорије графова и доношење закључака о његовој структури [88].

- Развијен алгоритам за аутоматско креирање Federated SPARQL упита који обухватају више репозиторијума на основу изабране путање из претходно добијених графова.

- Тестиране функционалности Платформе за различите супстанце од стране истраживача у Лабораторији (доменских експерата). Урађена провера релевантности враћених резултата и њихова анализа.

- Извршено поређење Платформе са другим доступним платформама. За поређење враћених резултата је коришћена *Open PHACTS* платформа на којој тренутно ради 27 партнера. Циљ овог поређења није био да се покаже која је платформа боља, већ да се укаже на постојање комплементарних података и да су даља сарадња и интеграција, у светлу све веће експанзије података на Вебу, неопходни кораци.

- Подаци и резултати су постали доступни за целокупну истраживачку заједницу. Комплетан код је јавно доступан и може се репродуковати[1].

[1] https://github.com/malibanekg/SpecINT

10.2 Смернице за даља истраживања

Поред раније наведених ограничења Платформе (Подсекција 9.5) које треба превазићи уколико је то изводљиво, постоји још неколико предлога који имају потенцијал да унапреде тренутну верзију Платформе. У наставку су прво дате смернице за решавање тренутних ограничења Платформе, а након тога и предлози за њено проширење.

10.2.1 Предлози за превазилажење ограничења Платформе

Недостатак информација о доступности и техничким ограничењима ендпоинта доводи то тога да није сваки Federated SPARQL упит у потпуности изводљив, већ само неки његови делови. Због свега тога, резултати некада могу бити некомплетни. Међутим, чак и да имамо ове информације, тешко је контролисати начин и редослед извршавања упита. Као једно могуће решење се намеће креирање више реплика ендпоинта за сваки скуп.

Описани приступ који се овде користи подразумева рад са семантички базираним скуповима података, тако да у овом тренутку није могуће прикупљати податке из других извора података (који нису у RDF формату). Тренутно Платформа не подржава рад са репозиторијумима који се ослањају на API као начину за приступ подацима. Пример репозиторијума који податке нуде кроз REST API су *PubChem*, *Open PHACTS*, *ChemSpider* итд. И ови подаци су драгоцени, јер многи од њих нису представљени кроз семантичке скупове података, и зато резултатима Платформе у будућности треба придружити и ове резултате.

Платформа има могућност да интегрише два репозиторијума у исто време. Међутим, коришћењем алгоритма за спектралну кластеризацију има смисла размишљати о укључивању више репозиторијума истовремено. Тада треба имати у виду комплексност и дужину извршавања упита, што отвара питање да ли толики број добијених резултата заиста доноси бенефите или само додатно компликује издвајање потребних од непотребних података.

Величина међурезултата који су смештени у меморију је недвосмислено смањена, јер добијена вредност из претходног подупита у низу смањује број резултата у наредном подупиту и тако ограничава смештање свих података у меморију. Међутим, некада је неопходно додатно водити рачуна о меморији која се користи, и због тога је битан редослед повезивања подупита у велики Federated SPARQL упит. Због ограничених ресурса овај проблем није разматран, али свакако и ове информације треба узети у обзир.

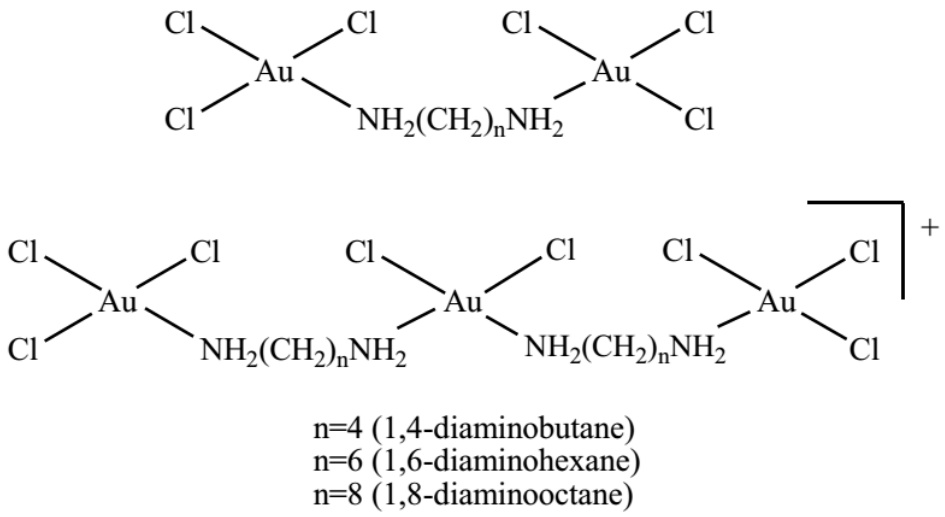
10.2.2 QSAR модул

Вероватно најважнија функционалост која ће у неком тренутку бити додата Платформи је модул за QSAR анализу. Истраживања су показала да овај део може да донесе више бенефита него константно проширивање броја добијених резултата из Платформе. QSAR модул има потенцијал да помогне у откривању обећавајућих антиканцерогених једињења користи већ добијене податке да би усмерио истраживаче и показао им начин да синтетишу ефикаснија једињења без потребе да их све тестирају. Овај приступ ће уштедети време, средства и изнад свега усмерити истраживаче ка откривању и синтези најефикаснијих једињења против малигнитета. У овом тренутку су прецизно дефинисана два пројекта на којима ће бити фокус у наредном периоду.

Тренутно у свету постоји пуно истраживачких пројеката који се баве анализом потенцијалних антиканцерогених активности многих биолошки активних једињења. Једињења која садрже платину припадају групи молекула са веома високим потенцијалом за лечење многих врста малигних обољења укључујући рак простате, рак дојке, рак плућа, колоректални карцином, рак грлића материце, итд. Анализа квантитативног односа структуре и активности молекула (QSAR), као један од главних хемометријских приступа, широко је применљива у предвиђању биолошке активности једињења која може да укључује и антиканцерогено деловање. Главни циљ предложеног пројекта, под називом „Синтеза, карактеризација, антиканцерогена процена и QSAR моделовање новосинтетисаних комплекса платине", је синтетисање и карактеризација 25 комплекса платине (II) и, потом, примена различитих хемометријских мултиваријантних алата за корелацију структурних карактеристика са биолошким одговором према ћелијама малигних тумора. У плану је коришћење низа стандардних *in vitro* метода: MTT тест, Анексин V-PI тест, анализа ћелијског циклуса и одређивање кључних протеина који су укључени у апоптозу (bax, bcl-2, каспаза 3 и цитохром c). Постоје две информатичке технике које се могу користити за испитивање односа структуре и цитотоксичности једињења: *класични QSAR* (односи квантитативне структуре и активности) и *3-D QSAR* (тродимензионални квантитативни однос структуре и активности). Обе технике имају и предности и ограничења, па ће се у овом пројекту обе технике и користити, а на крају изабрати она која даје најбоље резултате.

Други пројекат се односи на цитотоксичну активност коју испољавају одређени комплекси злата у односу на људску ћелијску линију канцера груди (MDA-MB-231), колоректални канцер (HCT-116) и фибробласт плућа (MRC-5). На Институту за хемију, Природно-математичког факултета у Крагујевцу, је синтетисан низ од dinuclear и trinuclear комплекса злата опште формуле $[Au_2(N-N)Cl_6](1-3)$ за dinuclear и $[Au_3(N-N)_2Cl_8]^+(4-6)$ за trinuclear супстанце, редом, где је за везу $N-N$ коришћен *bidentate ligand (1,4-diaminobutane; 1,6-diaminohexane or 1,8-diaminooctane).* Примери ових комплекса су приказани на Слици 10.1. У Лабораторији је показано да сви синтетисани комплекси смањују вијабилност нормалних ћелија и ћелија канцера, са значајним цитотоксичним ефектом (IC_{50} < 25 μM) за trinuclear gold(III) комплексе (4, 5) на HCT-116 ћелијској линији [120].

Први корак у оба пројекта обухвата претраживање одређених репозиторијума са циљем да се пронађу слични комплекси који су такође показали добро цитотоксично дејство на HCT-116 ћелијској линији. Идеја је да се након тога креирају предикциони модели који могу да сугеришу које комплексе треба даље синтетисати.



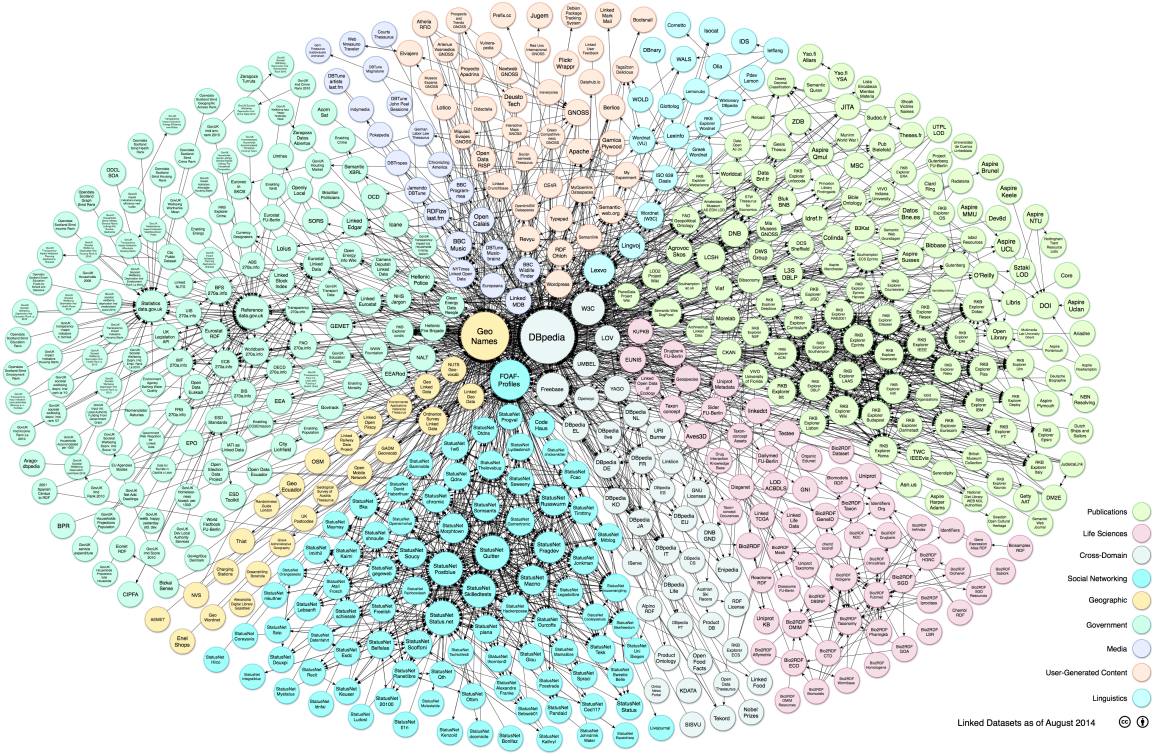
Слика 10.1: Структурне формуле за шест различитих комплекса злата.

10.2.3 Проширење Платформе на друге домене

Тренутна верзија Платформе је специјализована за природне науке, али се под одређеним условима може проширити и на друге области. Проширивост предложене методологије за интеграцију и претраживање више репозиторијума веома је битан моменат којем треба посветити пажњу. Посматрајући *LOD Cloud* са Слике 10.2, методологија се веома лако може проширити и на друге домене као што су влада (енг. *government*), кругови обележени жутом бојом, затим географија (плава боја), медији (љубичаста боја) итд. У основи, приступ би садржао следеће кораке:

- Детектовање најважнијих појмова који најближе одређују тематику и који се могу искористити за повезивање скупова.
- Констурисање онтологије која описује тематику скупова.
- Сакупљање подупита за одговарајућа питања од интереса.

Адаптација Платформе на други домен би свакако захтевала додатно време за подешавање иницијалних поставки, јер се она ослања на доменско знање о скуповима без којег не бисмо били у могућности да неке ствари предупредимо унапред, као што су повезивање путања, избор иницијалног чвора и слично.



Слика 10.2: LOD Cloud облак (преузето из Schmachtenberg et al [2]).

Додатак А

Додатни резултати у кластеровању одређених класа графова

Као што је већ поменуто, два графа не морају бити повезана само слепљивањем чворова или додавањем нове гране, већ је то могуће урадити и на друге начине. У овом додатку је описана једна од таквих процедура која има и математичку потврду да ју је могуће искористити за одређивање припадности скупова репозиторијумима.

Као што је већ познато, процес слепљивања (енг. *coalescence*) два графа G_1 и G_2 започиње избором два произвољна чвора, v_1 у графу G_1 и чвора v_2 у графу G_2. Након почетног избора два чвора се претварају у један чвор v који је суседан са свим чворовима у G_1 који су суседни чвору v_1 и сваким чвором из G_2 суседним чвору v_2.

Дефиниција А.1. Под резом графа G (енг. *cut*) сматраћемо скуп грана C за које партиција $N = (N_1, N_2)$ скупа чворова N графа G постоји тако да C садржи само гране из G за које један чвор припада скупу N_1, а други припада скупу N_2. Подграфови графа G индуковани подскуповима N_1 и N_2 се називају блокови пресека (C-блокови).

Дефиниција А.2. Блок је или максимално 2-повезан подграф, мост (заједно са својим чворовима) или изоловани чвор. Различити блокови графа G се преклапају у највише једној тачки која је тада артикулациони чвор (енг. *articulation point, cut-vertex*) графа G.

Дефиниција А.3. За дати граф G, подела скупа чворова $V(G) = V_1 \dot\cup V_2 \dot\cup \ldots \dot\cup V_k$ се назива *равноправна подела* (енг. *equitable partition*) ако сваки чвор из V_i има исти број суседа у V_j, за $i, j \in \{1, 2, \ldots, k\}$.

Дефиниција А.4. Нека је дата $s \times s$ матрица $B = (b_{ij})$, и нека је скуп чворова графа G подељен на непразне подскупове $X_1, X_2, \ldots, X_s$ тако да је за било које $i, j = 1, 2, \ldots, s$ сваки чвор из X_i суседан са тачно b_{ij} чворова из X_j (равноправна подела). Мултиграф H са матрицом суседства B се назива делиоцем (енг. *divisor*) графа G.

Напомена А.1. За доказ Теореме А.4 биће коришћена математичка техника која се ослања на делиоце графа, што нам омогућава да израчунавања радимо над мањим матрицама и тако добијемо потребне сопствене вредности и сопствене векторе на лакши и бржи начин. Основна идеја која је употребљена у доказу се заснива на одређивању сопственог вектора делиоца графа, а затим се на основу њега одређују који је чвор графа G позитиван, негативан, или једнак нули у односу на одређени сопствени вектор

графа G. Пре него што прикаћемо формулацију Теореме А.4, прво ћемо да прикажемо неке познате резултате који ће бити искоришћени за њено доказивање.

Координате сопственог вектора графа G се лако могу одредити на основу координата сопственог вектора делиоца графа G. Ако је $x^T = (x_1, x_2, \ldots, x_s)$ произвољан сопствени вектор делиоца графа, тада се одговарајући сопствени вектор z графа G добија на следећи начин:

$$z(G) = (\underbrace{x_1, x_1, \ldots, x_1}_{n_1}, \underbrace{x_2, x_2, \ldots, x_2}_{n_2}, \ldots, \underbrace{x_s, x_s, \ldots, x_s}_{n_s})$$

где су $n_i = |X_i|, i = 1, 2, \ldots, s$ бројеви чворова у одговарајућој равномерној подели.

Теорема А.1. (видети [84]) Произвољан делилац графа G садржи индекс од G као сопствену вредност.

За сопствену вредности λ_i се каже да је главна сопствена вредност (енг. *main eigenvalue*) ако $\epsilon(\lambda_i)$ није ортогоналан на вектор $(1, 1, \ldots, 1)$. Главним делом спектра графа G се назива скуп свих главних сопствених вредности и означава са $\mathcal{M}$.

Теорема А.2. (видети [84]) Спектар произвољног делиоца H графа G садржи главни део спектра од G.

Теорема А.3 (Декартово правило знакова)**.** (видети [91], страна 282) Нека је $f(x) = a_n x^n + a_{n-1}x^{n-1} + ... + a_1 x + a_0$ полином где су $a_n, a_{n-1}, ..., a_0$ реални коефицијенти.

Број позитивних нула од f је или једнак броју промена знакова између узастопних не-нула коефицијената од $f(x)$ или мањи од тог броја за паран цео број.

Број негативних нула од f је или једнак броју промена знакова између узастопних не-нула коефицијената од $f(-x)$ или мањи од тог броја за паран цео број.

Нека су K_n и K_m комплетни графови реда n и m, редом. Посматрајмо корак у којем је један чвор иницијалног графа K_n повезан са свим чворовима графа K_m. За иницијлани граф K_n ова процедура може да се понови све до предзадњег чвора, без брисања грана у међувремену. На овај начин се добије повезани граф G у којем све додате гране фомирају рез C_k графа G (где је k број корака), док су блокови реза индуковани чворовима графова K_n и K_m.

Лема А.1. Нека је граф G добијен у k-том кораку ($k < n$) користећи претходно описану процедуру. Тада важи да је $\lambda_2(G) > 0$.

Доказ. У карактеристичном полиному делиоца графа G се могу уочити две промене знака између узастопних чланова полинома, што значи да је највећи могући број позитивних нула управо 2. За $-\lambda$ постоји једна промена знака између узастопних чланова полинома што имплицира да је највећи могући број негативних нула једнак јединици. Из Теореме А.3 следи да је $\lambda_2(G) > 0$. ■

Теорема А.4. Нека је граф G добијен у k-том кораку ($k < n$) користећи претходно описану процедуру и нека је $z = (z_i)$ сопствени вектор који одговара другој највећој сопственој вредности матрице суседства графа G. Могућа су два случаја:

(1) за $\boldsymbol{n \leq m}$: у било којем кораку, чворови који припадају скупу $N(z)$ су у једном C_k-блоку, а чворови који припадају скупу $P(z)$ су у другом C_k-block.

(2) за $\boldsymbol{n > m}$: ако је испуњен услов $n - k < m$, тада чворови који припадају скупу $N(z)$ су у једном C_k-блоку, а чворови који припадају скупу $P(z)$ су у другом C_k-block.

Доказ. Посматрајмо једну равноправну поделу $\Pi : \{v_1, v_2, ..., v_{n-k}\}, \{v_{n-k+1}, v_{n-k+2}, ...,$-$v_n\}, \{v_{n+1}, v_{n+2}, v_{n+3}, ..., v_{n+m}\}$ графа G, где је k број узетих чворова из иницијалне компоненте. Из матрице суседства A делиоца графа G се добијају следеће једнакости:

$$\begin{aligned}
(n-k-1)x_1+ \quad & kx_2 \quad && = \lambda x_1 && \text{(A.1)}\\
(n-k)x_1- \quad & (k-1)x_2+ \quad && mx_3 = \lambda x_2 && \text{(A.2)}\\
& kx_2+ \quad && (m-1)x_3 = \lambda x_3 && \text{(A.3)}
\end{aligned}$$

Означимо са $\lambda = \lambda_2(G)$ (на основу Леме A.1 важи да је $\lambda_2(G) > 0$) и са $P_D(\lambda)$ карактеристични полином од A. Из (A.1) и (A.2) се добија

$$\begin{aligned}
kx_2 &= (\lambda - n + k + 1)x_1 && \text{(A.4)}\\
kx_2 &= (\lambda - m + 1)x_3 && \text{(A.5)}
\end{aligned}$$

(1) Ако је $\lambda = n-k-1$, тада важи $P_D(n-k-1) = k(k-n)(n-m-k) > 0$ за произвољно n и k. На основу теореме о преплитању важи да је $n-k-1 < \lambda_1$. Користећи ове чињенице, из (A.4) се добије да x_1 и x_2 имају исти знак тј. да сви чворови из првог блока реза C_k, $v_i \in \{v_1, v_2, ..., v_n\}$, су или сви позитивни или негативни.

Даље, за произвољно $n \leq m$ важи $P_D(m-1) = -km(k+m-n) < 0$. Познато је да је $m - 1 > 0$, одакле се даље добије да важи $\lambda < m - 1$. Пошто је $k > 0$ и $\lambda - m + 1 < 0$, из (A.5) следи да x_2 и x_3 имају различите знаке тј. сви чворови из другог блока реза C_k, $v_i \in \{v_{n+1}, v_{n+2}, ..., v_{2n}\}$, припадају различитим скуповима ($P(z)$ и $N(z)$) у односу на чворове из првог C_k-блока, чиме је завршен први део доказа.

(2) У другом случају се могу посматрати два подслучаја: $n - k = m$, $n - k > m$ и $n - k < m$. Користећи потпуно исти приступ, лако се доказује да теорема важи и овом случају. ■

Литература

[1] Vladimir P. Petrović, Marko N. Živanović, Dušica Simijonović, Jelena Đorović, Zorica D. Petrović, and Snežana D. Marković. Chelate N, O-palladium (II) complexes: synthesis, characterization and biological activity. *RSC Advances*, 5(105):86274–86281, 2015.

[2] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. State of the LOD Cloud 2014. *University of Mannheim, Data and Web Science Group. August*, 30, 2014.

[3] David J. Wild, Ying Ding, Amit P. Sheth, Lee Harland, Eric M. Gifford, and Michael S. Lajiness. Systems chemical biology and the Semantic Web: What they mean for the future of drug discovery research. *Drug Discovery Today*, 17(9-10):469–474, 2012. ISSN 13596446. doi: 10.1016/j.drudis.2011.12.019.

[4] Alberta Bergamo and Gianni Sava. Linking the future of anticancer metal-complexes to the therapy of tumour metastases. *Chemical Society Reviews*, 44(24):8818–8835, 2015. ISSN 14604744. doi: 10.1039/c5cs00134j.

[5] Petar Čanović, Jovana Bogojeski, Jelena V. Košarić, Snežana D. Marković, and Marko N. Živanović. Pt(IV), pd(II), and rh(III) complexes induced oxidative stress and cytotoxicity in the HCT-116 colon cancer cell line. *Turkish Journal of Biology*, 41(1):141–147, 2017. ISSN 13036092. doi: 10.3906/biy-1605-77.

[6] Marko N. Živanović, Jelena V. Košarić, Biljana Šmit, Dragana S. Šeklić, Radoslav Z. Pavlović, and Snežana D. Marković. Novel seleno-hydantoin palladium(II) complex - antimigratory, cytotoxic and prooxidative potential on human colon HCT-116 and breast MDA-MB-231 cancer cells. *General Physiology and Biophysics*, 36(2):187–196, 2017. ISSN 13384325. doi: 10.4149/gpb_2016036.

[7] Aidan Hogan and Supervisor Axel Polleres. *Exploiting RDFS and OWL for Integrating Heterogeneous , Large-Scale , Linked Data Corpora*. PhD thesis, National University of Ireland, Galway, 2011. URL `http://sw.deri.org/{~}aidanh/docs/thesis/thesis-one-sided.pdf`.

[8] Theodor H. Nelson. Complex information processing: a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th national conference*, pages 84–100. ACM, 1965. doi: 10.1145/800197.806036.

[9] Tim Berners-Lee, James Hendler, Ora Lassila, and Others. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[10] Ian Horrocks, Bijan Parsia, Peter Patel-Schneider, and James Hendler. Semantic Web architecture: Stack or two towers? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 3703 LNCS, pages 37–41, 2005. ISBN 3540287930. doi: 10.1007/11552222_4.

[11] Tim Berners-Lee, Roy Fielding, and Larry Masinter. Uniform resource identifiers (URI): Generic syntax, 2005. URL https://tools.ietf.org/html/rfc3986.

[12] Mark Needleman. The unicode standard. *Serials Review*, 26(2):51–54, 2000. ISSN 00987913. doi: 10.1080/00987913.2000.10764582.

[13] Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau, and John Cowan. Extensible Markup Language (XML) 1.1 (Second Edition), 2006.

[14] Patrick J. Hayes and Peter F. Patel-Schneider. RDF 1.1 Semantics. *W3C Recommendation 25 February 2014*, 2014. URL https://www.w3.org/TR/rdf11-mt/.

[15] Eric Prud Hommeaux and Andy Seaborne. SPARQL Query Language for RDF. *W3C Recommendation*, 2008. URL http://www.w3.org/TR/rdf-sparql-query/.

[16] The W3C SPARQL Working Group. SPARQL 1.1 Overview. W3C Recommendation 21 March 2013, 2013. URL https://www.w3.org/TR/sparql11-overview/.

[17] Dan Brickley. RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation 10 February 2004, 2004. URL https://www.w3.org/TR/2004/REC-rdf-schema-20040210/.

[18] Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview, 2004. URL https://www.w3.org/TR/owl-features/.

[19] Michael Kifer and Harold Boley. RIF Overview (Second Edition) - W3C Working Group Note 5, 2013. URL http://www.w3.org/TR/rif-overview.

[20] Claudio Gutierrez, Carlos A. Hurtado, Alberto O. Mendelzon, and Jorge Pérez. Foundations of Semantic Web databases. *Journal of Computer and System Sciences*, 77(3): 520–541, 2011. ISSN 00220000. doi: 10.1016/j.jcss.2010.04.009.

[21] Draltan Marin. RDF Formalization. Technical report, Universidad de Chile, Santiago de Chile, 2004. URL https://users.dcc.uchile.cl/{~}cgutierr/ftp/draltan.pdf.

[22] Christian Bizer, Richard Cyganiak, and Tom Heath. How to Publish Linked Data on the Web. *Publish*, 2007. URL http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/.

[23] Brian McBride. The Resource Description Framework (RDF) and its Vocabulary Description Language RDFS. In *Handbook on Ontologies*, pages 51–65. Springer Berlin Heidelberg, 2004. doi: 10.1007/978-3-540-24750-0_3.

[24] Jeremy J. Carroll, Ian Dickinson, Chris Dollin, Dave Reynolds, Andy Seaborne, and Kevin Wilkinson. Jena: implementing the semantic web recommendations. *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 74–83, 2002.

[25] Zoi Kaoudi, Iris Miliaraki, and Manolis Koubarakis. RDFS reasoning and query answering on top of DHTs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5318 LNCS, pages 499–516, 2008. ISBN 3540885633. doi: 10.1007/978-3-540-88564-1-32.

[26] Syed Muhammad Ali Hasnain. *Cataloguing and linking publicly available biomedical SPARQL endpoints for federation - addressing aPosteriori data integration.* PhD thesis, National University of Ireland Galway, 2017. URL `https://aran.library.nuigalway.ie/handle/10379/6518`.

[27] Kendall G. Clark, Lee Feigenbaum, and Elias Torres. SPARQL protocol for RDF, 2005. URL `https://www.w3.org/TR/rdf-sparql-protocol/`.

[28] Damien Graux. *On the efficient distributed evaluation of SPARQL queries.* PhD thesis, Université Grenoble Alpes, France, 2016. URL `https://tel.archives-ouvertes.fr/tel-01618366/`.

[29] Tim Berners-Lee. Design Issues: Linked Data, 2006. URL `http://www.w3.org/DesignIssues/LinkedData.html`.

[30] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global, 2011.

[31] Christian Bizer and Andreas Schultz. The Berlin SPARQL Benchmark. *International Journal on Semantic Web and Information Systems*, 5(2):1–24, 2011. ISSN 1552-6283. doi: 10.4018/jswis.2009040101.

[32] Tom Heath and Christian Bizer. Linked Data: Evolving the Web into a Global Data Space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011. ISSN 2160-4711. doi: 10.2200/s00334ed1v01y201102wbe001.

[33] Chris Bizer, Anja Jentzsch, and Richard Cyganiak. State of the LOD Cloud. *Version 0.3 (September 2011)*, 1803:1–9, 2011. URL `http://lod-cloud.net/state/`.

[34] Carlos Buil-Aranda, Aidan Hogan, Jürgen Umbrich, and Pierre Yves Vandenbussche. SPARQL web-querying infrastructure: Ready for action? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8219, pages 277–293, 2013. ISBN 9783642413377. doi: 10.1007/978-3-642-41338-4_18.

[35] François Belleau, Marc Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716, 2008. ISSN 15320464. doi: 10.1016/j.jbi.2008.03.004.

[36] Yanli Wang, Jewen Xiao, Tugba O. Suzek, Jian Zhang, Jiyao Wang, and Stephen H. Bryant. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Research*, 37(SUPPL. 2):623–633, 2009. ISSN 03051048. doi: 10.1093/nar/gkp456.

[37] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam MacIejewski, David Arndt, Michael Wilson, Vanessa Neveu, Alexandra Tang, Geraldine Gabriel, Carol Ly, Sakina Adamjee, Zerihun T. Dame, Beomsoo Han, You Zhou, and David S. Wishart. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Research*, 42(D1):1091–1097, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1068.

[38] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):9–14, 2012. ISSN 03051048. doi: 10.1093/nar/gkr988.

[39] Carolyn J. Mattingly, Glenn T. Colby, John N. Forrest, and James L. Boyer. The Comparative Toxicogenomics Database (CTD). *Environmental Health Perspectives*, 111(6):793, 2003. ISSN 00916765. doi: 10.1289/ehp.6028.

[40] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N. Jorissen, and Michael K. Gilson. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35(SUPPL. 1):198–201, 2007. ISSN 03051048. doi: 10.1093/nar/gkl999.

[41] T. E. Klein, J. T. Chang, M. K. Cho, K. L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D. E. Oliver, D. L. Rubin, F. Shafa, J. M. Stuart, and R. B. Altman. Integrating genotype and phenotype information: An overview of the PharmGKB project. *Pharmacogenomics Journal*, 1(3):167–170, 2001. ISSN 1470269X. doi: 10.1038/sj.tpj.6500035.

[42] Stefan Günther, Michael Kuhn, Mathias Dunkel, Monica Campillos, Christian Senger, Evangelia Petsalaki, Jessica Ahmed, Eduardo Garcia Urdiales, Andreas Gewiess, Lars Juhl Jensen, Reinhard Schneider, Roman Skoblo, Robert B. Russell, Philip E. Bourne, Peer Bork, and Robert Preissner. SuperTarget and Matador: Resources for exploring drug-target relationships. *Nucleic Acids Research*, 36(SUPPL. 1), 2008. ISSN 03051048. doi: 10.1093/nar/gkm862.

[43] QSAR sets. URL `http://www.cheminformatics.org`.

[44] Antony J. Williams, Lee Harland, Paul Groth, Stephen Pettifer, Christine Chichester, Egon L. Willighagen, Chris T. Evelo, Niklas Blomberg, Gerhard Ecker, Carole Goble, and Barend Mons. Open PHACTS: Semantic interoperability for drug discovery. *Drug Discovery Today*, 17(21-22):1188–1198, 2012. ISSN 13596446. doi: 10.1016/j.drudis.2012.05.016.

[45] Alasdair J.G. Gray, Paul Groth, Antonis Loizou, Sune Askjaer, Christian Brenninkmeijer, Kees Burger, Christine Chichester, Chris T. Evelo, Carole Goble, Lee Harland, Steve Pettifer, Mark Thompson, Andra Waagmeester, and Antony J. Williams. Applying linked data approaches to pharmacology: Architectural decisions and implementation. *Semantic Web*, 5(2):101–113, 2014. ISSN 22104968. doi: 10.3233/SW-2012-0088.

[46] Satya S. Sahoo, Wolfgang Halb, Sebastian Hellmann, Kingsley Idehen, Ted Thibodeau Jr, Sören Auer, Juan Sequeda, and Ahmed Ezzat. A survey of Current approaches for mapping of relational databases to RDF. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2009. ISSN 1570-8268. doi: 10.1016/j.websem.2007.11.011.

[47] Aleksandar Stanimirović, Miloš Bogdanović, and Leonid Stoimenov. Methodology and intermediate layer for the automatic creation of ontology instances stored in relational databases. *Software - Practice and Experience*, 43(2):129–152, 2013. ISSN 00380644. doi: 10.1002/spe.2103.

[48] Larisa N. Soldatova and Ross D. King. An ontology of scientific experiments. *Journal of the Royal Society Interface*, 3(11):795–803, 2006. ISSN 17425662. doi: 10.1098/rsif.2006.0134.

[49] Vladimir Cvjetković, Marija Đokić, Branko Arsić, and Milena Ćurčić. The ontology supported intelligent system for experiment search in the scientific Research center. *Kragujevac Journal of Science*, 36(36):95–110, 2014. ISSN 1450-9636. doi: 10.5937/kgjsci1436095c.

[50] Branko Arsić, Marija Đokić, Vladimir Cvjetković, Petar Spalević, and Marko Živanović. Integration of bioactive substances data for preclinical testing with Cheminformatics and Bioinformatics resources. In *Proceedings of 23nd International Electrotechnical and Computer Science Conference, ERK 2014*, pages 146–149, 2014.

[51] Paula de Matos, Rafael Alcántara, Adriano Dekker, Marcus Ennis, Janna Hastings, Kenneth Haug, Inmaculada Spiteri, Steve Turner, and Christoph Steinbeck. Chemical Entities of Biological Interest: an update. *Nucleic acids research*, 38:D249–54, 2010. ISSN 1362-4962. doi: 10.1093/nar/gkp886.

[52] A. Patrícia Bento, Anna Gaulton, Anne Hersey, Louisa J. Bellis, Jon Chambers, Mark Davies, Felix A. Krüger, Yvonne Light, Lora Mak, Shaun McGlinchey, Michal Nowotka, George Papadatos, Rita Santos, and John P. Overington. The ChEMBL bioactivity database: An update. *Nucleic Acids Research*, 42(D1):083–090, 2014. ISSN 03051048. doi: 10.1093/nar/gkt1031.

[53] Anja Jentzsch, Matthias Samwald, and Bo Andersson. Linking Open Drug Data. In *I-Semantics '09: Proceedings of the International Conference on Semantic Systems*, pages 3–6, 2009.

[54] Bin Chen, Xiao Dong, Dazhi Jiao, Huijun Wang, Qian Zhu, Ying Ding, and David J Wild. Chem2Bio2RDF: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC bioinformatics*, 11:255, 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-255.

[55] Simon Jupp, James Malone, Jerven Bolleman, Marco Brandizi, Mark Davies, Leyla Garcia, Anna Gaulton, Sebastien Gehant, Camille Laibe, Nicole Redaschi, Sarala M. Wimalaratne, Maria Martin, Nicolas Le Novère, Helen Parkinson, Ewan Birney, and Andrew M. Jenkinson. The EBI RDF platform: Linked open data for the life sciences. *Bioinformatics*, 30(9):1338–1339, 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btt765.

[56] Harry E. Pence and Antony Williams. Chemspider: An online chemical information resource. *Journal of Chemical Education*, 87(11):1123–1124, 2010. ISSN 00219584. doi: 10.1021/ed100697w.

[57] Heiner Stuckenschmidt, Richard Vdovjak, Geert-Jan Houben, and Jeen Broekstra. Index structures and algorithms for querying distributed RDF repositories. In *Proceedings of the 13th international conference on World Wide Web*, pages 631–639, 2004. doi: 10.1145/988672.988758.

[58] Giovanni Tummarello, Renaud Delbru, and Eyal Oren. Sindice.com: Weaving the open linked data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 4825 LNCS, pages 552–565, 2007. ISBN 3540762973. doi: 10.1007/978-3-540-76298-0_40.

[59] Jürgen Umbrich, Aidan Hogan, Axel Polleres, and Stefan Decker. Improving the Recall of Live Linked Data Querying through Reasoning. In *Web Reasoning and Rule System*, pages 188–204, 2012. doi: 10.1007/978-3-642-33203-6_14.

[60] Olaf Hartig. Zero-knowledge query planning for an iterator implementation of link traversal based query execution. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6643 LNCS, pages 154–169. Springer Berlin Heidelberg, 2011. ISBN 9783642210334. doi: 10.1007/978-3-642-21034-1_11.

[61] Jeen Broekstra, Arjohn Kampman, and Frank van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. In Ian Horrocks and James Hendler, editors, *International Semantic Web Conference (ISWC2002)*, pages 54–68, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-48005-1.

[62] Nur Aini Rakhmawati, Jürgen Umbrich, Marcel Karnstedt, Ali Hasnain, and Michael Hausenblas. Querying over Federated SPARQL Endpoints —A State of the Art Survey. Technical Report June, 2013. URL `http://arxiv.org/abs/1306.1723`.

[63] Andreas Schwarte, Peter Haase, Katja Hose, Ralf Schenkel, and Michael Schmidt. FedX: Optimization techniques for federated query processing on linked data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7031 LNCS, pages 601–616. Springer Berlin Heidelberg, 2011. ISBN 9783642250729. doi: 10.1007/978-3-642-25073-6_38.

[64] Olaf Görlitz and Steffen Staab. SPLENDID: SPARQL endpoint federation exploiting voiD descriptions. In *CEUR Workshop Proceedings*, volume 782, page 12, 2011.

[65] Renzo Angles, Ioan Toma, Peter Boncz, Josep Larriba-Pey, Irini Fundulaki, Thomas Neumann, Orri Erling, Peter Neubauer, Norbert Martinez-Bazan, and Venelin Kotsev. The linked data benchmark council. *ACM SIGMOD Record*, 43(1):27–31, 2014. ISSN 01635808. doi: 10.1145/2627692.2627697.

[66] Michael Schmidt, Olaf Görlitz, Peter Haase, Günter Ladwig, Andreas Schwarte, and Thanh Tran. FedBench: A benchmark suite for federated semantic data query processing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in*

Artificial Intelligence and Lecture Notes in Bioinformatics), volume 7031 LNCS, pages 585–600. Springer Berlin Heidelberg, 2011. ISBN 9783642250729. doi: 10.1007/978-3-642-25073-6_37.

[67] Muhammad Saleem, Axel Cyrille Ngonga Ngomo, Josiane Xavier Parreira, Helena F. Deus, and Manfred Hauswirth. DAW: Duplicate-AWare federated query processing over the web of data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8218 LNCS, pages 574–590. Springer Berlin Heidelberg, 2013. ISBN 9783642413346. doi: 10.1007/978-3-642-41335-3_36.

[68] Olaf Görlitz, Matthias Thimm, and Steffen Staab. SPLODGE: Systematic Generation of SPARQL Benchmark Queries for Linked Open Data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7649 LNCS, pages 116–132. Springer Berlin Heidelberg, 2012. ISBN 9783642351754. doi: 10.1007/978-3-642-35176-1_8.

[69] Bastian Quilitz and Ulf Leser. Querying distributed RDF data sources with SPARQL. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5021 LNCS:524–538, 2008. ISSN 03029743. doi: 10.1007/978-3-540-68234-9_39.

[70] Maribel Acosta, Maria Esther Vidal, Tomas Lampo, Julio Castillo, and Edna Ruckhaus. ANAPSID: An adaptive query processing engine for SPARQL endpoints. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7031 LNCS, pages 18–34. Springer Berlin Heidelberg, 2011. ISBN 9783642250729. doi: 10.1007/978-3-642-25073-6_2.

[71] Muhammad Saleem, Yasar Khan, Ali Hasnain, Ivan Ermilov, and Axel Cyrille Ngonga Ngomo. A fine-grained evaluation of SPARQL endpoint federation systems. *Semantic Web*, 7(5):493–518, 2016. ISSN 22104968. doi: 10.3233/SW-150186.

[72] Muhammad Saleem, Qaiser Mehmood, and Axel Cyrille Ngonga Ngomo. FEASIBLE: A feature-based SPARQL benchmark generation framework. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9366, pages 52–69. Springer Berlin Heidelberg, 2015. ISBN 9783319250069. doi: 10.1007/978-3-319-25007-6_4.

[73] Muhammad Saleem, Claus Stadler, Qaiser Mehmood, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. SQCFramework: SPARQL Query Containment Benchmark Generation Framework. In *K-Cap*, 2017.

[74] Heiko Dietze and Michael Schroeder. GoWeb : a semantic search engine for the life science web Abstract Background. *BMC bioinformatics*, 10(10):1–14, 2014.

[75] Dominik Schweiger, Zlatko Trajanoski, and Stephan Pabinger. SPARQLGraph: A web-based platform for graphically querying biological Semantic Web databases. *BMC Bioinformatics*, 15(1), 2014. ISSN 14712105. doi: 10.1186/1471-2105-15-279.

[76] Alexander De Leon Battista, Natalia Villanueva-Rosales, Myroslav Palenychka, and Michel Dumontier. SMART: A web-based, ontology-driven, semantic web query answering application. In *CEUR Workshop Proceedings*, volume 295, 2007.

[77] María Jesús García-Godoy, Ismael Navas-Delgado, and José Aldana-Montes. Bioqueries. In *ACM International Conference Proceeding Series*, pages 24–31, 2012. ISBN 9781450310765. doi: 10.1145/2166896.2166906.

[78] Wei Hu, Honglei Qiu, Jiacheng Huang, and Michel Dumontier. BioSearch: a semantic search engine for Bio2RDF. *Database : the journal of biological databases and curation*, 2017, 2017. ISSN 17580463. doi: 10.1093/database/bax059.

[79] Marija Đokić Petrović, Vladimir Cvjetković, Jeremy Yang, Marko Živanović, and David J. Wild. PIBAS FedSPARQL: A web-based platform for integration and exploration of bioinformatics datasets. *Journal of Biomedical Semantics*, 8(1), 2017. ISSN 20411480. doi: 10.1186/s13326-017-0151-z.

[80] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4): 395–416, 2007. ISSN 09603174. doi: 10.1007/s11222-007-9033-z.

[81] Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56(18):3825–3833, 2012. ISSN 13891286. doi: 10.1016/j.comnet.2012.10.007.

[82] Alon Altman and Moshe Tennenholtz. Ranking systems: the PageRank axioms. In *Proceedings of the 6th ACM conference on Electronic commerce*, pages 1–8, 2005.

[83] Dragoš M. Cvetkovic, Michael Doob, Horst Sachs, and Others. *Spectra of graphs*, volume 10. Academic Press, New York, 1980.

[84] Dragoš Cvetković, Peter Rowlinson, and Slobodan Simić. *Eigenspaces of graphs*. Cambridge University Press, 1997. doi: 10.1017/cbo9781139086547.

[85] Dragoš Cvetković, Peter Rowlinson, and Slobodan Simić. *An Introduction to the Theory of Graph Spectra*. Cambridge University Press, 2009. ISBN 0521118395.

[86] Reinhard Diestel. *Graph Theory (Graduate Texts in Mathematics)*. Springer Berlin Heidelberg, 2000. ISBN 3540261834. doi: 10.1109/IEMBS.2010.5626521.

[87] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

[88] Branko Arsić, Dragoš Cvetković, Slobodan Simić, and Milan Škarić. Graph spectral techniques in computer sciences. *Applicable Analysis and Discrete Mathematics*, 6(1): 1–30, 2011. ISSN 1452-8630. doi: 10.2298/aadm111223025a.

[89] Dragoš Cvetković and Peter Rowlinson. The largest eigenvalue of a graph: A survey. *Linear and Multilinear Algebra*, 28(1-2):3–33, 1990. ISSN 0308-1087. doi: 10.1080/03081089008818026.

[90] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *Proceedings of the IEEE Symposium on Reliable Distributed Systems*, pages 25–34, 2003. ISBN 0769519555. doi: 10.1109/RELDIS.2003.1238052.

[91] Piet Van Mieghem, Jasmina Omic, and Robert Kooij. Virus spread in networks. *IEEE-/ACM Transactions on Networking*, 17(1):1–14, 2009. ISSN 10636692. doi: 10.1109/TNET.2008.925623.

[92] Michael D. König, Stefano Battiston, Mauro Napoletano, and Frank Schweitzer. On Algebraic Graph Theory and the Dynamics of Innovation Networks. *Networks and Heterogeneous Media*, 2008.

[93] Michael D. König, Stefano Battiston, Mauro Napoletano, and Frank Schweitzer. The efficiency and stability of R&D networks. *Games and Economic Behavior*, 75(2):694–713, 2012. ISSN 08998256. doi: 10.1016/j.geb.2011.12.007.

[94] Dragoš Cvetković, Peter Rowlinson, and Slobodan Simić. Graphs with least eigenvalue -2: Ten years on. *Linear Algebra and Its Applications*, 484:504–539, 2015. ISSN 00243795. doi: 10.1016/j.laa.2015.06.012.

[95] Nair Maria Maia de Abreu. Old and new results on algebraic connectivity of graphs. *Linear Algebra and Its Applications*, 423(1):53–73, 2007. ISSN 00243795. doi: 10.1016/j.laa.2006.08.017.

[96] Steve Kirkland and Shaun Fallat. Perron components and algebraic connectivity for weighted graphs. *Linear and Multilinear Algebra*, 44(2):131–148, 1998. ISSN 0308-1087. doi: 10.1080/03081089808818554.

[97] Steve Kirkland. A note on limit points for algebraic connectivity. In *Linear Algebra and Its Applications*, volume 373, pages 5–11, 2003. doi: 10.1016/S0024-3795(02)00413-5.

[98] Teh-Hsing Wei. *Algebraic foundations of ranking theory*. PhD thesis, University of Cambridge, 1952. URL `https://www.repository.cam.ac.uk/handle/1810/250988`.

[99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999. ISSN 00045411. doi: 10.1145/324133.324140.

[100] Amy N. Langville and Carl D. Meyer. A Survey of Eigenvector Methods for Web Information Retrieval. *SIAM Review*, 47(1):135–161, 2005. ISSN 0036-1445. doi: 10.1137/s0036144503424786.

[101] Miroslav Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25(4):619–633, 1975.

[102] Stephen T. Barnard and Horst D. Simon. Fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. *Concurrency: Practice and Experience*, 6(2):101–117, 1994. ISSN 10969128. doi: 10.1002/cpe.4330060203.

[103] Alex Pothen, Horst D. Simon, and Kan-Pu Liou. Partitioning Sparse Matrices with Eigenvectors of Graphs. *SIAM journal on matrix analysis and applications*, 11(3):430–452, may 1990. ISSN 0895-4798. doi: 10.1137/0611030.

[104] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. ISSN 01628828. doi: 10.1109/34.868688.

[105] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 975–982, 1999. doi: 10.1109/iccv.1999.790354.

[106] Christos Gkantsidis, Milena Mihail, and Ellen Zegura. Spectral analysis of Internet topologies. In *Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies*, pages 364–374, 2004. doi: 10.1109/infcom.2003.1208688.

[107] Damien Fay, Hamed Haddadi, Andrew Thomason, Andrew W. Moore, Richard Mortier, Almerima Jamakovic, Steve Uhlig, and Miguel Rio. Weighted spectral distribution for internet topology analysis: Theory and applications. *IEEE/ACM Transactions on Networking*, 18(1):164–176, 2010. ISSN 10636692. doi: 10.1109/TNET.2009.2022369.

[108] R. B. Bapat, S. J. Kirkland, and S. Pati. The perturbed laplacian matrix of a graph. *Linear and Multilinear Algebra*, 49(3):219–242, 2001. ISSN 0308-1087. doi: 10.1080/03081080108818697.

[109] Simon J. Coles, Nick E. Day, Peter Murray-Rust, Henry S. Rzepa, and Yong Zhang. Enhancement of the chemical semantic web through the use of InChI identifiers. *Organic and Biomolecular Chemistry*, 3(10):1832–1834, 2005. ISSN 14770520. doi: 10.1039/b502828k.

[110] Jon Chambers, Mark Davies, Anna Gaulton, Anne Hersey, Sameer Velankar, Robert Petryszak, Janna Hastings, Louisa Bellis, Shaun McGlinchey, and John P. Overington. UniChem: A unified chemical structure cross-referencing and identifier tracking system. *Journal of Cheminformatics*, 5(1):3, 2013. ISSN 17582946. doi: 10.1186/1758-2946-5-3.

[111] Charu C Aggarwal. *Data classification: algorithms and applications*. CRC press, 2014.

[112] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002. ISBN 0262042088.

[113] Gene H. Golub and Charles F. van Loan. *Matrix Computations*. Johns Hopkins University Press, fourth edition, 2013. ISBN 1421407949 9781421407944.

[114] Strahinja Z Kovačević, Sanja O Podunavac-Kuzmanović, Lidija R Jevrić, Evgenija A Djurendić, and Jovana J Ajduković. Non-linear assessment of anticancer activity of 17-picolyl and 17-picolinylidene androstane derivatives–chemometric guidelines for further syntheses. *European Journal of Pharmaceutical Sciences*, 62:258–266, 2014.

[115] Nolen Joy Perualila-Tan, Ziv Shkedy, Willem Talloen, Hinrich WH Göhlmann, Quantitative Structure Transcription Assay Relationships (QSTAR) Consortium, Marijke Van Moerbeke, and Adetayo Kasim. Weighted similarity-based clustering of chemical structures and bioactivity data in early drug discovery. *Journal of bioinformatics and computational biology*, 14(04):1650018, 2016.

[116] Fabian A Grimm, Yasuhiro Iwata, Oksana Sirenko, Grace A Chappell, Fred A Wright, David M Reif, John Braisted, David L Gerhold, Joanne M Yeakley, Peter Shepard, et al. A chemical–biological similarity-based grouping of complex substances as a prototype approach for evaluating chemical alternatives. *Green chemistry*, 18(16):4407–4419, 2016.

[117] Raimund Mannhold, Povl Krogsgaard-Larsen, and Hendrik Timmerman. *Advanced computer-assisted techniques in drug discovery*, volume 3. John Wiley & Sons, 2008.

[118] Gang Fu, Colin Batchelor, Michel Dumontier, Janna Hastings, Egon Willighagen, and Evan Bolton. Pubchemrdf: towards the semantic annotation of pubchem compound and substance databases. *Journal of cheminformatics*, 7(1):34, 2015.

[119] Branko Arsić, Marija Đokić-Petrović, Petar Spalević, Ivan Milentijević, Dejan Rančić, and Marko Živanović. Specint: a framework for data integration over cheminformatics and bioinformatics rdf repositories. *Semantic Web*, 10(4):795–813, 2019.

[120] Snežana Radisavljević, Dušan Ćoćić, Snežana Jovanović, Biljana Šmit, Marijana Petković, Nevena Milivojević, Nevena Planojević, Snežana Marković, and Biljana Petrović. Synthesis, characterization, dft study, dna/bsa-binding affinity, and cytotoxicity of some dinuclear and trinuclear gold (iii) complexes. *JBIC Journal of Biological Inorganic Chemistry*, 24(7):1057–1076, 2019.

Биографија аутора

Бранко Ј. Арсић је рођен 17.08.1984. године у Крагујевцу. Основну школу „Јован Поповић” и „Прву крагујевачку гимназију” у Крагујевцу, завршио је као носилац дипломе Вук Караџић. Основне студије на Природно-математичком факултету Универзитета у Крагујевцу уписао је школске 2003/2004 године, које је завршио 2008. године са просечном оценом 9.50/10. Дипломски рад под називом “Информациони систем за праћење каријере студената – клијентска и администраторска веб апликација” одбранио је са највишом оценом. Током школовања био је стипендиста Министарства просвете и Министарства науке и технолошког развоја Републике Србије, и добитник награде Конгреса српског уједињења - Задужбине Студеница и награде Српске народне одбране - фонда „Михаило Пупин“. Докторске студије на Електронском факултету у Нишу уписао је школске 2014/2015, при чему је положио све испите предвиђене студијским програмом.

Као истраживач, ангажован је од фебруара 2009. године у оквиру пројеката Министарства науке и технолошког развоја Републике Србије, под називом „Теорија графова и математичко програмирање са применама у хемији и рачунарству” (174033) и „Развој нових информационо-комуникационих технологија, коришћењем напредних математичких метода, са применама у медицини, телекомуникацијама, енергетици, заштитити националне баштине и образовању” (ИИИ44006). Током 2017. године боравио је на Темпл Универзитету, у Филаделфији, Сједињене Америчке Државе, у центру Center for Data Analytics and Biomedical Informatics. Том приликом је био део пројекта „DARPA-GRAPHS: Prospective Analysis of Large and Complex Partially Observed Temporal Social Networks” чији је носилац Темпл Универзитет. Циљ боравка је обухватао истраживања из области машинског учења, где је аутор радио на примени стандардних и добро познатих алгоритама структурне регресије на графовима и комплексним мрежама.

Бранко Арсић је аутор седам научних радова (три рада су на SCI листи), једног техничког решења, и дванаест саопштења на међународним конференцијама. Његова истраживања обухватају развој алгоритама за машинско учење (надгледано и ненадгледано учење), као и интеграцију података на Интернету коришћењем технологија Семантичког Веба и спектралне теорије графова.

Додатно, аутор је ангажован као Data Scientist у Истраживачко развојном центру за биоинжењеринг - BioIRC, у Крагујевцу, где тренутно ради на примени дубоких мрежа на медицинским сликама. Аутор је такође и сарадник на Институту за филозофију и друштвену теорију, Универзитета у Београду, где ради на примени дубоких мрежа у анализи текста (Natural Language Processing).

Библиографија

Радови у међународним и домаћим часописима

- R. Miković, B. Arsić, Đ. Gligorijević, M. Gačić, D. Petrović, and N. Filipović. The Influence of Social Capital on Knowledge Management Maturity of Nonprofit Organizations–Predictive Modelling Based on a Multilevel Analysis. IEEE Access, 7, pp. 47929-47943, 2019. DOI: 10.1109/ACCESS.2019.2909812
- B. Arsić, M. Đokić-Petrović, P. Spalević, I. Milentijević, D. Rančić, and M. Živanović. SpecINT: A framework for data integration over cheminformatics and bioinformatics RDF repositories. Semantic Web - Interoperability, Usability, Applicability, Vol. 10, No. 4, pp. 795-813, 2019. DOI: 10.3233/SW-180327
- B. Arsić, Lj. Bojić, I. Milentijević, P. Spalević, and D. Rančić. SYMBOLS: software for social network analysis. Facta Universitatis, Series: Automatic Control and Robotics, vol. 17, no. 3, pp. 205-222, 2018. DOI:10.22190/FUACR1803205A
- V. Cvjetković, M. Đokić, B. Arsić, and M. Ćurčić. The ontology supported intelligent system for experiment search in the scientific research center. Kragujevac Journal of Science, 36(2014), pp. 95-110.
- A. Alwardi, B. Arsić, I. Gutman and N. D. Sonera. The common neighborhood graph and its energy. Iranian Journal of Mathematical Sciences and Informatics, vol. 7, issue 2 (2012), pp.1-8.
- B. Arsić, I. Gutman, K. Ch. Das, and K. Xu. Relations between Kirchhoff index and Laplacian–energy–like invariant. Bulletin T.CXLIV de l'Acadèmie serbe des sciences et des arts 2012 Classe des Sciences mathèmatiques et naturel les Sciences mathèmatiques, vol. 37, pp. 61-72.
- B. Arsić, D. Cvetković, S. K. Simić, and Milan Škarić. Graph spectral techniques in computer sciences. Applicable Analysis and Discrete Mathematics, 6(2012), issue 1, pp. 1-30.

Радови у зборницима конференција међународног значаја

- B. Arsić, M. Obrenović, M. Anić, A. Tsuda, N. Filipović (accepted). Image segmentation of the pulmonary acinus imaged by synchrotron X-ray tomography. 19th annual IEEE International Conference on Bioinformatics and Bioengineering (BIBE), October 28-30, 2019, Athens, Greece.
- M. Bašić and B. Arsić. Dynamic updates of hierarchical clustering operations (Meeting abstract). 5th International Scientific Conference, Analysis, Topology, Algebra: Theory and Applications (ATA2016), Serbia, July 6-9, 2016, pp. 7.
- B. Arsić, M. Bašić, P. Spalević, M. Ilić, M. Veinović. Facebook profiles clustering. 6th International Conference on Information Society and Technology (ICIST 2016), Serbia, 28 February - 2 March, 2016, pp. 154-158 (ISBN: 978-86-85525-16-2).
- A. Ljajić, E. Ljajić, P. Spalević, B. Arsić, D. Vučković. Sentiment analysis of textual comments in field of sport. 24th International Electrotechnical and Computer Science Conference (ERK 2015), pp. 35-38, IEEE, Slovenia, September 21-23, 2015 (ISSN:1581-4572).

- B. Arsić, P. Spalević, Lj. Bojić, A. Crnišanin. Social networks in logistics system decision-making. 2nd Logistics International Conference (LOGIC 2015), pp. 166-171, Serbia, May 21-23, 2015 (ISSN: 978-86-7395-339-7).
- B. Arsić, M. Đokić, V. Cvjetković, P. Spalević, S. Ilić. 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, March 8-11, 2015 (ISBN: 978-86-85525-16-2).
- B. Arsić, M. Đokić, V. Cvjetković, P. Spalević, M. Živanović, M. Mladenović. Integration of bioactive substances data for preclinical testing with cheminformatics and bioinformatics resources. 23nd International Electrotechnical and Computer Science Conference (ERK 2014), IEEE, Vol. 1, issue 1, pp. 146-149, Slovenia, September 22-24, 2014 (ISSN:1581-4572).
- B. Arsić, M. Đokić, N. Stefanović. Mapping ebXML standards to ontology. 4th International Conference on Information Society and Technology (ICIST 2014), vol. 1, pp. 198-203, Kopaonik, Serbia, March 9-13, 2014 (ISBN: 978-86-85525-14-8).
- V. M. Cvjetković, M. Đokić, B. Arsić. Semantically based customized search on local web site. The 2nd Virtual International Conference on Advanced Research in Scientific Areas (ARSA 2013), Vol. 2, issue 1, pp. 453-458, Slovakia, December 2 - 6, 2013 (ISBN: 978-80-554-0825-5, ISSN:1338-9831).
- V. M. Cvjetković, M. Đokić, B. Arsić. Wikipedia browsing with DBpedia. The 2nd Electronic International Interdisciplinary Conference (EIIC 2013), vol. 2, issue 1, pp. 470-475, Slovakia, September 02-06, 2013 (ISBN: 978-80-554-0762-3, ISSN: 1338-7871).
- V. M. Cvjetković, M. Đokić, B. Arsić. Ontology visualization. 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 (ISBN: 978-80-554-0606-0, ISSN: 1338-9831).
- V. M. Cvjetković, M. Đokić, B. Arsić. OWL based modeling and visualization of arbitrary semantic data structures. 5th International Conference Science and Higher Education in Function of Sustainable Development, Bussines Technical College Uzice, pp. 2:13-19, October 04 - 05, 2012 (ISBN 978-86-83573-26-4).

Техничка решења и постери

Техничко решење: Семантички базирана кастомизована претрага на локалном веб сајту. – М84

Постер: B. Arsić, B. Stojanović, N. Filipović. Automated IVUS contour detection using normalized cuts[1]. US-Serbia and West Balkan Data Science Workshop, Београд, август 2018.

[1]https://imi.pmf.kg.ac.rs/pub/58f1d1ef71a7ec2e63d47cfe3c933d35_08292018_114114/nsf_ivus_poster_final.pdf

ИЗЈАВА О ИСТОВЕТНОСТИ ШТАМПАНОГ И ЕЛЕКТРОНСКОГ ОБЛИКА ДОКТОРСКЕ ДИСЕРТАЦИЈЕ

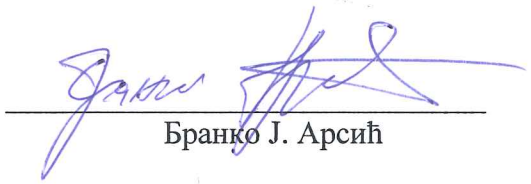
Наслов дисертације:

СКАЛАБИЛНА СОФТВЕРСКА ПЛАТФОРМА ЗА ПРЕТРАЖИВАЊЕ ХЕМИЈСКИХ И БИОЛОШКИХ РЕПОЗИТОРИЈУМА

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

У Нишу, _____ године

Потпис аутора дисертације



Бранко Ј. Арсић

ИЗЈАВА О АУТОРСТВУ

Изјављујем да је докторска дисертација, под насловом

СКАЛАБИЛНА СОФТВЕРСКА ПЛАТФОРМА ЗА ПРЕТРАЖИВАЊЕ ХЕМИЈСКИХ И БИОЛОШКИХ РЕПОЗИТОРИЈУМА

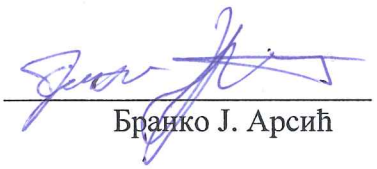
која је одбрањена на Електронском факултету Универзитета у Нишу:

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

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

У Нишу, _____ године

Потпис аутора дисертације



Бранко Ј. Арсић

ИЗЈАВА О КОРИШЋЕЊУ

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

СКАЛАБИЛНА СОФТВЕРСКА ПЛАТФОРМА ЗА ПРЕТРАЖИВАЊЕ ХЕМИЈСКИХ И БИОЛОШКИХ РЕПОЗИТОРИЈУМА

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

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

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

У Нишу, _____ године

Потпис аутора дисертације

_____

Бранко Ј. Арсић