**МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ**

**ХЕРСОНСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ**

**Факультет комп'ютерних наук, фізики та математики**

**Кафедра комп'ютерних наук та програмної інженерії**

# РОЗРОБЛЕННЯ ПАРСЕРІВ НАУКОМЕТРИЧНИХ БАЗ ДАНИХ ДЛЯ ІНТЕГРАЦІЇ ДАНИХ В СИСТЕМУ KSU24

**Кваліфікаційна робота**

на здобуття ступеня вищої освіти «бакалавр»

Виконала: студентка 4 курсу 12-431 групи

Спеціальності: 122 Комп'ютерні науки

Освітньо-професійної програми: Комп'ютерні науки

Корягіна Анастасія Анатоліївна

Керівники: старший викладач Черненко І.Є.,

доктор пед. наук, професор Співаковський О.В.

Рецензент: Ємець О.С., Middle Software Engineer, Squad

Херсон – Івано-Франківськ - 2024

# ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

API …… Application Programming Interface

JSON ..... JavaScript Object Notation

НМБД ……. Наукометрична база даних

НПП …… Науково-педагогічний працівник

# ЗМІСТ

# ВСТУП

**Актуальність теми:** Щорічно в Херсонському державному університеті проводиться процес рейтингування Університет. Значна кількість документів, які збираються під час процесу рейтингування, містить повторювану інформацію, що спричинено тим, що інформація збирається за останні три календарних роки. Зі зростанням обсягів наукових публікацій збільшується кількість повторюваної інформації, що збирається для оцінювання наукової діяльності науково-педагогічних працівників. Цей етап вимагає вдосконалення процесу збору та обробки даних з наукометричних баз даних для реалізації системи автоматичного рейтингування. Отримання актуальної наукометричної інформації необхідна умова для функціювання модулю "Рейтингування університету" платформи KSU24.

**Мета дослідження**: Метою кваліфікаційної роботи є розробка та впровадження високоефективних парсерів для НМБД, здатних враховувати особливості структури та форматування наукометричних документів. Реалізація ефективних парсерів є першим етапом для автоматизованого збору та обробки даних з НМБД, що підвищить точність модулю рейтингування платформи KSU24.

Для досягнення поставленої мети перед нами було поставлені наступні **завдання дослідження**:

- Аналіз сервісу Publication;
- Вивчення процесу рейтингування в університеті;
- Аналіз науково-метричних баз даних (НМБД) на предмет парсингу;
- Розробка та написання парсерів;
- Вивчення умов користування API за необхідністю;

**Об’єкт дослідження:** Наукометричні бази даних.

**Предмет дослідження:** Механізми та ефективність реалізації парсерів для збору інформації з наукометричних баз даних.

**Структура роботи:** Робота складається зі змісту, переліку умовних скорочень, вступу, двох розділів, висновків, списку використаних джерел і додатків.

# РОЗДІЛ 1
# АНАЛІЗ ПРОЦЕСУ РЕЙТИНГУВАННЯ УНІВЕРСИТЕТУ

## 1.1 Аналіз сервісу Publication

Першим пунктом у написанні Розділу 1 кваліфікаційної роботи є завдання аналізу сервісу Publication, який було розроблено у співавторстві Співаковського О. В. , Вінника М. О. , Полторацького М. Ю. , Тарасіч Ю. Г. , Лубчук А. О., Круглика В.С. Цей сервіс надає можливість отримувати аналітику публікаційної діяльності Херсонського державного університету на основі показників наукометричних баз даних. Основними наукометричними базами, метрики з яких враховуються під час формування рейтингів та аналітики, є Scopus, Google Scholar, Web of Science, Semantic Scholar. [1] Загальний план аналізу складається з наступних пунктів:

1. Опис сервісу Publication та його функцій. Це пункт передбачає огляд можливостей сервісу Publication у контексті наукометричного аналізу та формування рейтингів відповідних відділів університету: факультетів, кафедр, науково-педагогічних працівників, аналіз інструментів та функціоналу сервісу для підтримки процесу рейтингування.
2. Участь сервісу Publication у процесі рейтингування. Цей пункт передбачає роль сервісу у підготовці та аналізі наукометричних даних, внесок у процес рейтингування шляхом надання актуальних даних для аналізу.
3. Механізми підготовки та аналізу наукометричних даних: огляд процесу збору, аналізу та інтерпретації наукометричних показників
4. Оцінка ефективності сервісу у рамках процесу рейтингування: визначення переваг та недоліків процесу рейтингування з використанням сервісу та самого сервісу Publication.

Університетський сервіс є ключовим інструментом у процесі формування рейтингів відділів та науково-педагогічних працівників Херсонського

державного університету, сервіс дає можливість формувати рейтинги факультетів, кафедр та НПП на основі наукометричих даних, отриманих шляхом парсингу, формування аналітики, відображення кластерного аналізу.

Сервіс Publication надає зручний інтерфейс для перегляду інформації, рейтингів та аналізу наукометричних даних. Головна сторінка сервісу містить посилання на джерела інформації, а також графік, що відображає *«Кількість зареєстрованих осіб за факультетами»*, дозволяючи візуально оцінити розподіл кількості зареєстрованих осіб між різними факультетами.[2]

Наступна сторінка, що представлена у даному сервісі *«Рейтинг»* факультетів, кафедр, науково-педагогічних працівників. Спільно у кожного зі сформованих рейтингів є використання одного набору метрик у кожної НМБД: Scopus, Google Scholar, Semantic Scholar, Web of Science. Також сторінка надає можливість відсортувати рейтинг за зростанням або спаданням наступних метрик: індекс Хірша, кількість документів та цитуваннь. Дана сторінка має функцію пошуку вченого за ПІБ, після пошуку натиснувши на посилання на особистий профіль науковця у цьому сервісі можна перейти до детальної сторінки про НПП. На детальній сторінці присутня інформація з різних сервісів, можливо відслідити як змінювався індекс Хірша науковця у кожній з баз. Також на сторінці можливо побачити в яких базах було опубліковано певну роботу. [2]

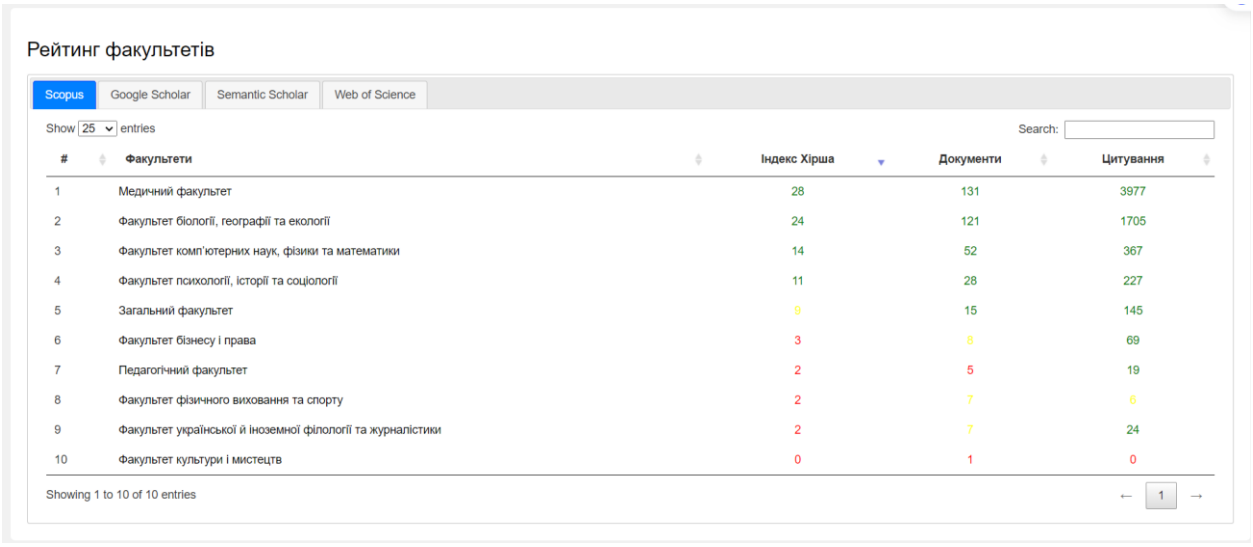

*Рисунок 1.1 Рейтинг факультетів сервісу Publication*

Сторінка *«Документи за галузями знань»* містить звіт, що генерується та діаграму, які вказують, у яких галузях знань знаходиться більше документів. З аналізу, який доступний зараз, можна побачити, що найбільший об’єм документів відноситься до галузі знань Комп’ютерні науки, отже на основі цієї інформації вже можна зробити висновки який відділ найчастіше робить наукові публікації. Графічне відображення діаграми допомагає візуально оцінити розподіл документів між різними галузями знань для зручного та зрозумілого аналізу.[2]

Show 25 entries Search:

| Галузь знань | Кількість |
|---|---|
| Computer Science | 138 |
| Agricultural and Biologic Sciences | 61 |
| Mathematics | 45 |
| Engineering | 24 |
| Business, Management and Accounting | 22 |
| Material Science | 15 |
| Health Professions | 14 |
| Economics, Econometrics and Finance | 13 |
| Environmental Science | 13 |
| Biochemistry, Genetics and Molecular Biologic | 11 |
| Eart and Planetary Sciences | 11 |
| Social Sciences | 10 |
| Chemical Engineering | 8 |
| Physics and Astronomy | 7 |
| Decision Sciences | 6 |
| Energy | 6 |

*Рисунок 1.2 Документи за галузями знань*

Сервіс відображає кількість наукових документів, що були написані у співавторстві з іншими науковими установами та виокремлює це у вкладці «Заклади партнери», що вказує рівень партнерства та міжнародного зв’язку університету з іншими закладами освіти або загалом науковими установами. Відображення цієї інформації в сервісі дозволяє користувачам швидко оцінити рівень активності університету, виявити можливості для майбутньої співпраці та розробити нові стратегії міжнародного партнерства університету.[2]

Сторінка *«Кластерний аналіз»* надає інформативний інструмент для аналізу даних трьох наукометричних баз даних: сторінка відображає впорядковану візуалізацію статистичних даних. Головною перевагою цієї сторінки є можливість отримувати узагальнену інформацію про активність науковців у різних НМБД, для користувачів це є можливістю легко та доступно порівняти статистичні дані.[2]

Сторінка *«Аналітика»* є інтерактивною сторінкою, на якій відображена аналітика факультетів та кафедр, звіти генеруються з можливістю встановити фільтри. Інформація надається по наступним наукометричним базам даних Scopus, Web of Science, Google Scholar.[2]

Після проведеного огляду можливостей та функцій сервісу Publication у контексті наукометричного аналізу та формування рейтингів відповідних відділів університету, аналізу інструментів та функціоналу сервісу для підтримки процесу рейтингування необхідно зазначити роль сервісу у процесі рейтингового оцінювання. Під час вивчення було виявлено, що однією із основних функцій сервісу Publication є надання швидко доступної інформації з профілів науково-педагогічних працівників співробітникам Наукової бібліотеки ХДУ під час перевірки «Індивідуальних рейтингів НПП» ХДУ. Профіль науково-педагогічного працівника містить в собі посилання на його сторінку у відповідній НМБД Scopus, Web of Science, Google Scholar, Semantic Scholar й додатково інформацію про h-індекс, кількість цитувань, кількість документів, а також таблицю з усіма науковими публікаціями з поміткою, в якій з НМБД індексується ця публікація; граф співавторів та у дужках кількість спільних наукових праць. У випадку своєчасної підтримки та оновлення інформації цей сервіс буде мати перевагу – надання достовірної інформації про НПП у порівнянні з ручним пошуком у НМБ науково-педагогічного працівника, що займає значно більший проміжок часу. В такому випадку сервіс забезпечує підвищену ефективність та достовірність процесу рейтингування університету за допомогою своїх функцій та механізмів. [2]

Подальший аналіз сервісу показав, що сервіс має обмежену інформацію по кожному науковцю. Сервіс оперує загальними відомостями, такими як ім’я, ідентифікатор науковця, кількість цитувань та документів, назви статей та співавторів. Цієї інформації достатньо для формування загальних рейтингів але цього не достатньо в повній мірі, щоб проводити оцінювання згідно з критеріями, які описані у документі, який регулює процес рейтингування університету. Наприклад, інформація про індекс Хірша може бути використана при підрахунку балів згідно критерію №2, а ось критерій №1 не може бути обчислено тільки маючи інформацію сервісу. Необхідно буде робити перехід на сторінку науковця в сервісі, до прикладу Scopus і вже звідти брати інформацію про журнал та знов переходити на інший серів, для отримання інформації про квартиль та імпакт фактор. Інформація з сервісів збирається за допомогою парсерів, кожен парсер збирає дані окремо по кожному сервісу. Ці парсери можуть бути взяті до уваги при подальшій розробці парсерів наукометричних баз даних в рамках даної кваліфікаційної роботи для їх вдосконалення та розширення функціональності.

Останнім завданням є оцінка ефективності сервісу в рамках процесу рейтингування університету. Даний сервіс має складну систему підтримки актуальності інформації: необхідно окремо вносити інформацію по НПП. За умови наявності достовірної та актуальної інформації сервіс Publication може бути ефективним інструментом для підвищення рівня якості та прозорості оцінювання в рамках рейтингування університету. Під час аналізу було виявлено, що хоча сервіс й має свої переваги перед ручним пошуком інформації про науково-педагогічних працівників але через наповненості застарілою, неактуальною та неповною інформацією це сповільнює процес рейтингування, та сервіс вже не є настільки достовірним джерелом інформації. Зразком неактуальної інформації є наявність серед науково-педагогічних працівників даних про тих, хто вже не є співробітником Херсонського державного університету.

Даний сервіс вимагає додаткової роботи програмістів, для підтримки актуальності інформації та додатково роботи наукової бібліотеки.

## 1.2 Процес рейтингування в університеті

Під час дослідження процесу рейтингування університету було ознайомлено з документами, які регулюють цей процес в Херсонському державному університеті. Таким документом є «Положення про систему рейтингового оцінювання діяльності науково-педагогічних працівників, кафедр та факультетів Херсонського державного університету». Даний документ спирається на низку законів України «Про вищу освіту», «Про освіту», «Про наукову, науковотехнічну діяльність». Цими законами передбачено впровадження рейтингового оцінювання в освітній процес. [3] Також варто зазначити, що до уваги беруться й внутрішні документи, такі як Статут Херсонського державного університету, Положення про академічну доброчесність учасників освітнього процесу та інші.

Згідно з положенням впровадження рейтингового оцінювання націлене на дотримання напрямку, за яким рухається університет, який зазначено у стратегії розвитку Херсонського державного університету. Дана стратегія прописується на наступні п’ять календарних років. Рейтингове оцінювання націлене також на формування стимулювання діяльності НПП та кафедр з факультетами у відповідності до цього напрямку, що зазначено у пункті положення про основні завдання системи. Також рейтингове оцінювання націлене на забезпечення об’єктивного оцінювання для подальшого самоаналізу відповідно до критеріїв акредитації, вдосконалення системи стимулювання діяльності кафедр, факультетів та науково-педагогічних працівників.[4]

При оцінюванні кафедр оцінка враховує в себе показник наукової діяльності кафедри, показники рейтингового оцінювання та кількість НПП

кафедри. Положення також чітко описує критерії оцінювання кафедри. До критеріїв входять інституційні показники, які складають 43,75% від всіх критеріїв та показники інтегральні, які складають 56,25% від всіх критеріїв.

При оцінюванні факультетів оцінка враховує в себе показник наукової діяльності факультету, показники рейтингового оцінювання та кількість НПП факультету. До критеріїв входять інституційні показники, які складають 48,5% від загальної кількості критеріїв оцінювання факультетів та інтегральні показники, які складають 51,5% від всіх критеріїв оцінювання. Багато критеріїв, як бал за критерій враховують загальну кількість балів по НПП.

Дослідження показало, що при оцінці науково-педагогічних працівників здобувачі вищої освіти також приймають участь та впливають на рейтинг НПП шляхом проходження опитування. Інформація, яка наведена у рейтингах НПП перевіряється та підтверджується багатьма учасниками. Даний підхід має на меті забезпечення точної оцінки проте ускладнює наявний бізнес процес. [4]

Також при процесі рейтингування НПП враховується не лише наукова діяльність але й міжнародна, в склад якої входить участь в міжнародних наукових проєктах, міжнародна академічна мобільність викладача, викладання в закладах вищої освіти закордоном, проведення міжнародних персональних виставок. Також враховується й освітня діяльність, до цього переліку входить: дистанційні курси на KSUonline, робота в складні різних комісій, рад, участь у проведенні акредитаційної експертизи та проведення гостьових лекцій.[4]

Дослідження показало, що більша частина критеріїв відповідає за оцінювання наукової діяльності – 60% всіх критеріїв відносяться саме до цього пункту, 15% критеріїв складають критерії міжнародної діяльності та 25% критеріїв освітньої діяльності.

Першим основним процесом рейтингування університету є процес збору інформації. У Положенні детально описано процедуру рейтингового

оцінювання, у якому виділено пункт «Формування власних індивідуальних рейтингів НПП за формою» на який виділено 15 робочих днів з дати підписання наказу, відповідальною особою є сам науково-педагогічний працівник. Зважаючи на навантаження НПП стає очевидним, що цей етап також займає багато часу: необхідно чітко дотримуватись поставлених вимог, заповнювати форму власноруч. Зазначена тема кваліфікаційної роботи є початковим етапом, який забезпечить зняття даної нагрузки з науково-педагогічного працівника, адже деякі дані наукової діяльності беруться з наукометричних баз. Як покаже дослідження, описане далі у кваліфікаційній роботі, дану інформацію можливо отримувати шляхом розроблення механізмів парсингу. [4]

З огляду на всебічну цифровізацію Університету для спрощення роботи за рахунок автоматизації можна зробити висновок, що певні етапи процедури рейтингування можна автоматизувати. Відповідно до теми кваліфікаційної роботи було визначено розробити ефективні інструменти, щоб автоматизувати етап «Формування власних власних індивідуальних рейтингів НПП за формою» саме за допомогою парсерів наукометричних баз даних, що створить підґрунтя для подальшого вдосконалення та автоматизації цього процесу. Такий висновок було зроблено з огляду на поточний бізнес-процес, який налічує велику кількість залучених осіб, перевірок та інших моментів, що значно ускладню даний процес. На противагу цьому буде розроблено новий бізнес-процес, який буде базуватись на тому, що науково-педагогічні працівники і інші учасники процесу будуть мати достовірну інформацію, зібрану автоматично та подальші дії, які необхідно робити учасникам буде мінімізовано.

# РОЗДІЛ 2
# РОЗРОБКА ПАРСЕРІВ НАУКОМЕТРИЧНИХ БАЗ ДАНИХ

## 2.1 Постановка задачі

В програмній частині кваліфікаційної роботи було розроблено парсери наступних наукометричних баз даних: Scopus, Web Of Science, Google Scholar для подальшої інтеграції зібраних даних у систему KSU24. Основними задачами при розробці парсерів наукометричних баз даних були:

1. Аналіз структури та взаємодії із сервісами. Сюди входило розгляд структур та особливостей сайтів кожного сервісу, оцінка методів взаємодії з фронтендом та вивчення доступних API, аналіз переваг та недоліків кожного сервісу для ефективного парсингу.
2. Написання модулю збору даних з Excel файлу. В цю задачу входило розробка та опис модуля, призначеного для збору інформації з Excel-файлу про викладачів, визначення полів, які були включені у таблицю та методів їх обробки.
3. Розробка структури та формату зберігання отриманих даних. В цю задачу входив опис уніфікованої структури для зберігання інформації про викладачів та їх публікації визначення формату для великого JSON-файлу, який містить дані про всіх викладачів.
4. Розробка механізмів парсингу. В цю задачу входили реалізація механізмів парсингу для кожного сервісу, визначення переваг та недоліків використаних механізмів.

Проект розроблявся у середовищі розробки PyCharm, мовою програмування було обрано мову Python. Обрання Python було обумовлене його ефективністю у вирішенні завдань з парсингу, обробки даних та взаємодії

з веб-сервісами.[5] Основними бібліотеками, які використовувались були бібліотеки json, requests, typing, logging, loguru.

## 2.2 Аналіз структури та взаємодії із сервісами (Scopus, Google Scholar, Web of Science)

Кожен з сервісів має свою унікальну структуру але всі вони спрямовані забезпечувати доступ до наукових даних та їхнього аналізу. Ці сервіси допомагають науковцям та академічним установам знаходити, переглядати та оцінювати наукові публікації, а також визначати їх значення та впливовість у відповідних галузях знань. У межах університету інформація з цих сервісів збирається по кожному науковцю для подальшого формування відповідних документів та проведення рейтингування НПП, кафедр та факультетів.

Кожен науковець має свій унікальний ідентифікатор, який відповідає його профілю у НМБД, це дозволяє однозначно ідентифікувати його в системі, таким чином це є мінімальною необхідною інформацією для подальшого парсингу даних НПП.

### 2.2.1 Scopus

Scopus - найбільша в світі реферативна база даних і науково-метрична платформа, що містить понад п’ятдесят мільйонів записів, база даних надає посилання на повний текст матеріалів, розміщених в ній.[6]

Серед інших сервісів, розглянутих у даній кваліфікаційній роботі, Scopus визначається легкістю доступу до даних, що було виявлено під час аналізу НМБД Scopus, Google Scholar та Web of Science на предмет парсингу за допомогою використання Developer tools, що дозволяє аналізувати мережеві запити та відповіді сервера. Таким чином було виявлено, що дані повертаються з API сервісу.

Сайт має закрите API, яким можна користуватись, та легко та структуровано отримувати данні, доступ до API можна отримати на Elsevier Developer Portal. [7] Використання цього API спрощує інтеграцію та забезпечує легкий доступ до наукометричних даних. Огляд доступних функцій API, особливості доступу, процес отримання ключа доступу та умови використання для наукових досліджень ретельно описані в документації.

Було проведено детальне вивчення документації та виявлено необхідність використання Academic Research в некомерційних цілях, в рамках якого відкривається розширена можливість отримання детальної інформації про автора та його публікації, яка повністю відповідає вимогам та критеріям оцінювання у межах процесу рейтингування університету. [8] Важливою умовою є те, що для того, щоб отримувати відповідь необхідно мати доступ до мережі навчального закладу.

### 2.2.2 Google Scholar

Google Scholar - вільно доступна академічна пошукова система, що надає вільний доступ до повного тексту наукових публікацій з різних напрямків. Однією з функцій, які надає даний сервіс це відомості про документи в межах бази даних, що посилаються на обрану статтю.[6]

Парсинг даних з Google Scholar може бути здійснений за допомогою бібліотеки requests, яка використовує HTTP-запити для отримання HTML-коду сторінок. Цей метод дозволяє витягувати необхідну інформацію про наукового працівника але має значний недолік: у випадку, коли кількість публікацій автора перевищує можливу кількість для відображення на сторінці - інші залишаються прихованими.

Для того, щоб загрузити більшу кількість публікацій необхідно робити автоматизацію за допомогою Selenium. Selenium - це набір інструментів для автоматизації веб-браузерів та виконання різних дій на сторінках, наприклад

натискання клавіш, перевірка елементів, який може бути використаний для витягування даних, які не доступні для безпосереднього парсингу через HTTP-запити. [9] Але цей варіант варто використовувати у крайньому випадку, який може виникнути при обмежених можливостях безпосереднього парсингу через HTTP-запити.

Крім автоматизації Selenium було розглянуто й інші варіанти, наприклад використання Serp API. Serp API - платформа, що дозволяє робити запити до Google Scholar та отримувати відповіді у форматі JSON. [10] Аналіз показав, що хоча Serp API і володіє певними перевагами але не є оптимальним варіантом для використання в рамках поставлених задач. Обмеження безкоштовної версії у кількості 100 запитів на місяць роблять її непрактичною для парсингу даних наукових працівників з великою кількістю публікацій: створюється необхідність підвантажувати дані за допомогою пагінації, що збільшує кількість запитів для одного наукового працівника в залежності від кількості сторінок з публікаціями. Для обходу обмежень потребується створення множини ключів або придбання платної підписки, що не відповідає поставленим задачам дослідження.

На основі проведеного аналізу, було вирішено використовувати відкриту бібліотеку scholarly. [11] Цей вибір було зроблено на основі ефективності цього методу, а також його простоти використання, адже зникала необхідність роботи автоматизацію.

### 2.2.3 Web Of Science

Web of Science, як і Scopus – реферативна наукометрична база даних, однією з ключових концептів якої є індекс впливовості або імпакт-фактор наукового видання.[6]

Аналіз сервісу Web of Science показав, що має API з різним набором даних, які можливо отримати. Існує 4 доступних API-інтерфейси:

1. Web of Science Lite – надає доступ до пошуку по Web of Science Core Collection та отримувати метадані основної статті, доступ для демонстрації публіці, може бути використано для оптимізації репозиторіїв та систем досліджень.
2. Web of Science – розширений. Як можна зрозуміти з назви, інтерфейс надає доступ до більш детальної інформації: повні бібліографічні метадані, поєднує інформацію про автора та іншу інформацію пов’язану з ним, також доступ показу для широкої публіки.
3. InCites – інструмент для оцінки цитувань, який допомагає проводити дослідження продуктивності навчального або наукового закладу та порівнювати з іншими колегами та конкурентами, також доступ показу для широкої публіки.
4. Витяг даних про відповідність статей (AMR) – інтерфейс дає можливість встановити зв’язки між публікаціями, також доступ показу для широкої публіки.[12]

Отже, для отримання інформації про наукового працівника та його публікації, як і для Scopus, можна використовуючи API інтерфейс, отримавши дані обравши згідно до вимог необхідні поля.

### 2.2.4 Semantic Scholar

Під час аналізу сервісу Publication було виявлено врахування інформації про кожного науковця з сервісу Semantic Scholar. [13] Semantic Scholar – розроблена в Інституті штучного інтелекту Аллена пошукова платформа, яка комбінує машинне навчання та обробку природної мови для додавання шару семантичного аналізу окрім методів аналізу цитування. Сервіс виділяє важливі статті та встановлені зв’язки між статтями. [14]

Проте «Положення про систему рейтингового оцінювання діяльності науково-педагогічних працівників, кафедр і факультетів Херсонського

державного університету» не містить критеріїв з урахуванням інформації з цього сервісу. Але було прийняте стратегічне рішення про додавання інформації з цього сервісу у систему KSU24 за рахунок розроблення парсеру.

Аналіз сервісу показав, що існує відкрите API за допомогою якого можна без проблем отримати дані та відсутня необхідність авторизуватись чи генерувати ключ доступу. Доступна інформація про автора: пошук за ім’ям, деталі про автора, деталі про його публікації; про публікації: деталі про публікацію, про автора, цитування, посилань та інше. [15] API дуже зручне у використанні та потребує лише встановлення необхідних параметрів запиту, які використовує браузер щоб виконати додаткові команди і потім віддати вміст. [16] Таким чином можна отримати детальну інформацію по кожному науковцю, яка необхідна, обравши відповідні поля.

## 2.3 Модуль збору даних з Excel-файлу

В ході розробки було розроблено модуль збору даних з Excel-файлу. Дані про наукових працівників збирались відповідно на кожному факультеті у загальний файл, що мав наступну структуру.(таб. 2.1)

*Таблиця 2.1*

Поля файлу про науково-педагогічних працівників

| **worker__ person_id** | **worker__ person__ full_name** | **department __name** | **department__ faculty__ name** | **orcid** | **google_scholar** | **scopus** | **web_of_science** |
|---|---|---|---|---|---|---|---|

Відповідно до структури файлу було визначені наступні поля для формування відповідного JSON-файлу з інформацією про всіх наукових працівників: worker__person_id - id наукового працівника у системі KSU24, та

відповідно до кожного наукового працівника поля orcid, scopus, google_scholar, web_of_science у одноімених наукометричних базах даних.

В початковому варіанті формувались списки по кожній наукометричній базі даних з id наукових працівників, було зроблено висновок, що в такому випадку губиться залежність якому працівнику належить ідентифікатор, тому остаточним став варіант, коли головним ключем словника є дані з worker__person_id. Такий підхід спростить інтеграцію даних у систему KSU24.

Збором та об'єднанням даних з різних сервісів забезпечувалося створення єдиного, комплексного JSON-файлу для кожного працівника, що включав ідентифікатори усіх використовуваних наукометричних баз даних.

Під час збору даних було виявлено наявність науково-педагогічних працівників, дані про яких, були відсутні у кожній з наступних orcid, scopus, google_scholar, web_of_science полів. Тому було зроблено висновок та виконано додавання перевірки існування запису хоча б в одній з комірок, в такому випадку дані про науково-педагогічного працівника записувались, в іншому - ні.

## 2.4 Структура та формат зберігання отриманих даних

В ході розробки було поставлено питання вибору формату файлу для зберігання даних та опису уніфікованої структури файлів для зберігання інформації про науково-педагогічних працівників та їх публікації. Було обрано JSON формат файлів на основі наступних переваг:

JSON - це зручний, універсальний та гнучкий формат для впорядкованого зберігання та обміну даними, що вже є стандартом обміну інформації в Інтернеті, він має переваги у простоті, читабельності: має вигляд звичайного тексту, який легко читається людиною та має зрозумілу структуру: пари «ключ-значення», де ключі є рядками а значення – допустимий тип даних

JSON, що робить найкращим варіантом для збереження великої кількості інформації. [17]

У програмі присутні 3 різних типи JSON файлів, кожен з яких відповідає за збереження окремої, унікальної інформації.

Перший файл містить у собі інформацію про ID наукових працівників, які були зібрані з Excel файлу. Головним ключем кожного наукового працівника є ключ у системі KSU24, в тілі містить інформацію про ID у НМБД.

Файл, що зберігає дані про помилки, які виникли відповідно до ID наукового працівника окремо від основного JSON-файлу. Таким чином це спростить процес відслідковування помилок, та дасть можливість в окремій ітерації зібрати дані.

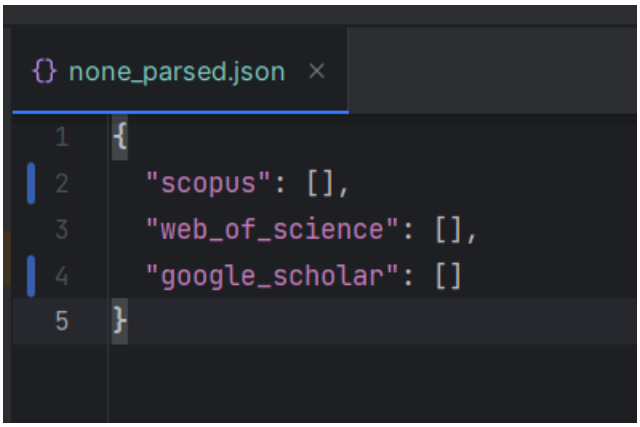


*Рисунок 2.1 Збереження помилок під час парсингу*

Третій тип JSON файлу – є файл зі збереженою інформацією про всіх науковців університету, сформований по кожній наукометричній базі даних окремо.

## 2.5 Розробка механізмів парсингу

Першим було розроблено парсер даних зі НМБД Scopus, парсер являє собою набір трьох класів, один з яких є базовим та містить в собі функцію для надіслання запиту за переданою URL-адресою, цей клас має наступний код:

```python
class ScopusBaseParser:
    """Scopus Base parser
    Makes request and return response"""
    4 usages  rvxt21
    @staticmethod
    def _get_response(url: str) -> requests.Response | None:
        try:
            response = requests.get(url, headers={'Accept': 'application/json'})
            response.raise_for_status()
            return response
        except requests.RequestException as e:
            logging.error(f'Request error on {url}: {e}')
            return None

```

*Рисунок 2.2 Базовий клас парсингу даних зі Scopus*

У цьому коді метод **_get_response** відповідає за надсилання GET-запиту до вказаної URL-адреси та обробку отриманої відповіді. Якщо статус відповіді не вказує на успішне завершення запиту (наприклад, HTTP-помилка), генерується виняток, який логується для подальшого аналізу.

Під час розробки було використано концепцію наслідування об’єктно-орієнтовного програмування – механізм об’єктно-орієнтованого програмування, що дозволяє класу наслідувати властивості та методи іншого класу. Наслідування дозволяє уникати повторне написання коду. [18] Це було зроблено для того, щоб сприяти відповідності принципу ПО «Don`t repeat yourself». Головною ідеєю цього методу є винесення повторюваного коду у окрему функцію, або клас, бо якщо з’явиться необхідність змінити логіку, дотримавшись цього принципу, змінивши код один раз – це застосується в усіх місцях де, викликається ця функція або метод класу, у противному випадку буде необхідно змінювати код в усіх місцях, де він описан. [19]

Діаграма класів парсингу даних з наукометричної бази Scopus виглядає наступним чином:

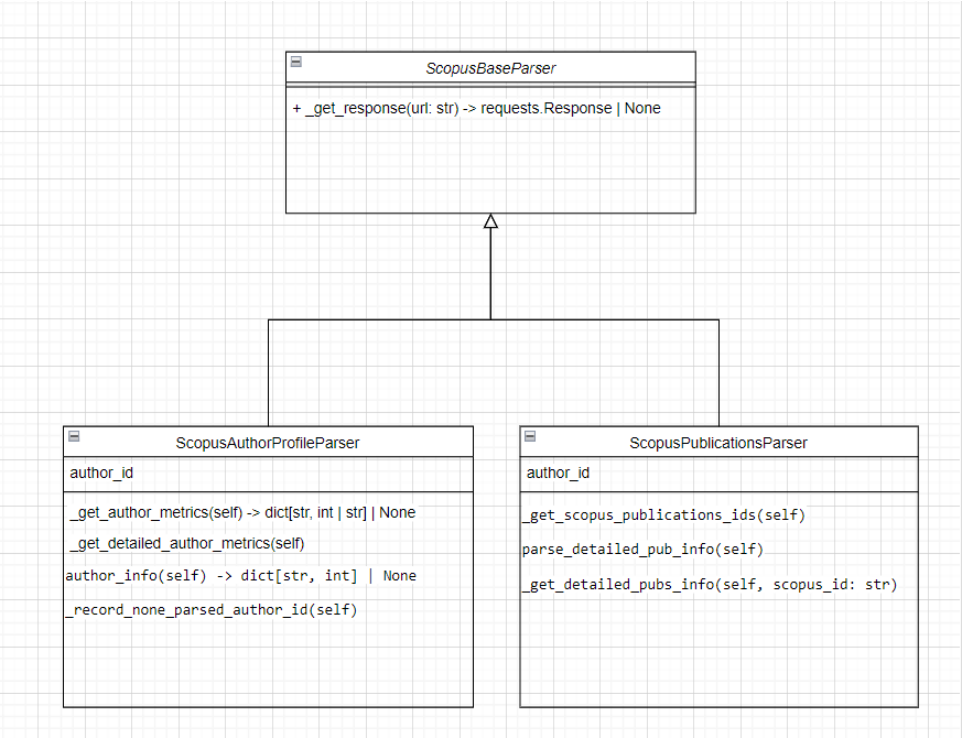


*Рисунок 2.3 UML-діаграма класів Scopus*

Дані збираються по кожному автору окремо і формуються у наступному вигляді :


*Рисунок 2.4 Базовий вигляд збереження інформації про автора*

Спираючись на документ, який регулює рейтингове оцінювання університету було визначено наступний перелік даних, який необхідно зібрати з бази даних Scopus. (таб. 2.2)

Найважливішими метриками та полями, які приймаються до уваги під час рейтингового оцінювання НПП є індекс Хірша, кількість цитувань, виключаючи самоцитування а також важливим пунктом було врахування та збір інформації про предметні області до яких відноситься публікація, адже наступним кроком під час розподілу балів враховується квартиль видання у певній категорії. Це є важливим через те, що журнал, у якому було опубліковано статтю може належати до кількох категорій і згідно цим категоріям можуть бути визначені різні квартилі. [20]

У модулі парсингу даних з наукометричної бази Scopus також було додано функції для обробки інформації та форматування згідно правил PEP8 стиля іменування змінних snake_case ( або lower_case_with_underscores). [21]

Модуль парсингу даних з наукометричної бази даних Google Scholar містить один клас, який відповідає за збір даних по автору ( НПП ), діаграма класів даного класу має наступний вигляд:

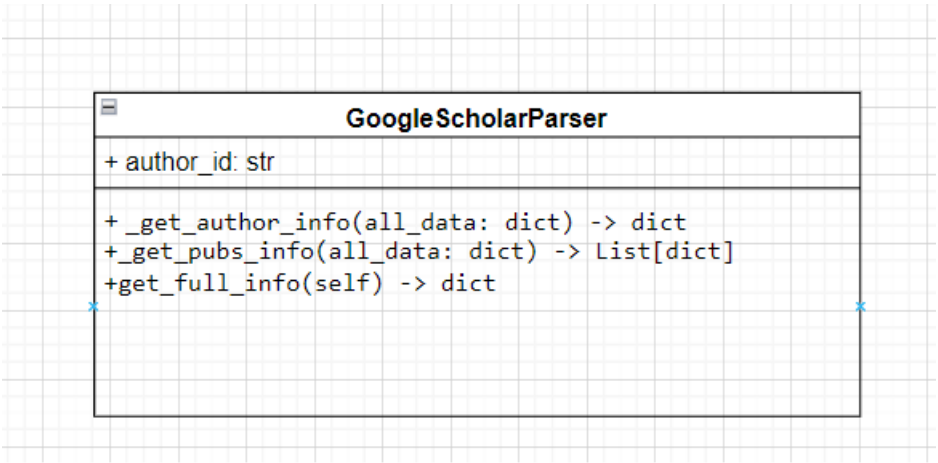


*Рисунок 2.5 UML-діаграма класів Google Scholar*

Як було зазначено у пункті 2.1.2 при аналізі було зроблено вибір на користь відкритої бібліотеки scholarly для парсингу даниз з наукометричної бази даних. Дані збираються у аналогічний формат, як і для Scopus, цей формат є спільним для всіх джерел та є базовим. Варто зазначити, що множина інформації, яку можливо зібрати з даного джерела відрізняється від множини даних Scopus, проте були збережені головні та найважливіші метрики, які беруться до уваги під час рейтингування.

Наступним було виконано розробку парсеру сервісу Semantic Scholar, варто зазначити, що для написання парсеру було затрачено менше часу у порівнянні з іншими сервісами, це зумовлено простотою користування API. Парсер являє собою клас, який приймає ідентифікатор автора як параметр та повертає інформацію по автору, якщо описувати клас за допомогою UML-діаграми він буде мати наступний вигляд

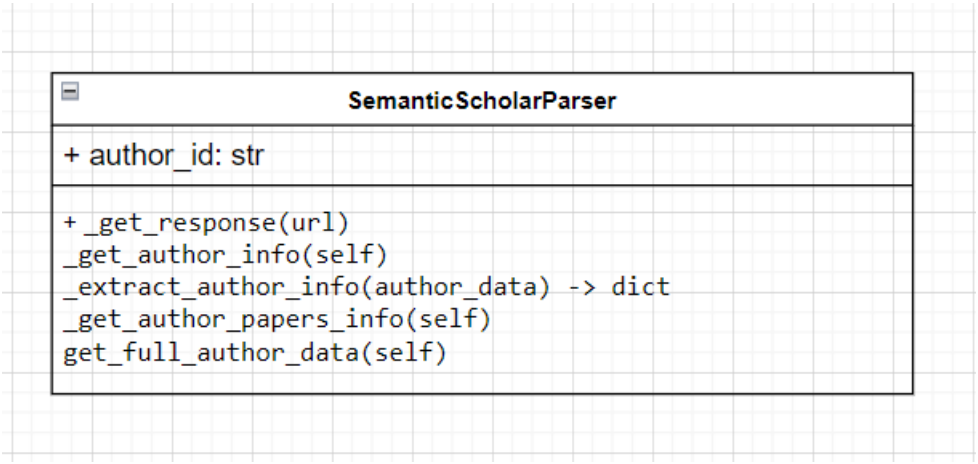


*Рисунок 2.6 UML-діаграма класів Semantic Scholar*

Методи _get_author_info() та _get_author_papers_info() повертають відповідь з інформацією, яка встановлюється за допомогою змінної fields, як параметри запиту, регулювання інфомації, яку необхідно отримати просте – додати поля, які необхідно отримати та прибрати зайві, усі доступні поля описані у відкритому доступі на сторінці з документацією.

Останнім було розроблено парсер наукометричної бази Web of Science. Даний парсер вимагав більших часових затрат на розробку, адже велику кількість часу було відведено на аналіз отримання даних з сервісу. В кінцевому результаті було зроблено парсер на основі надсилання запитів на API, та обробку отриманої відповіді. З надісланого JSON-файлу збиралась необхідна інформація для проведення рейтингування, адже у випадку, коли інформація зберігається у тому вигляді, в якому вона була отримана – це створить велике навантаження на систему за рахунок великої кількості даних.

Результатом роботи є чотири JSON-файли з повною інформацією про науково-педагогічних працівників, зібраною по кожному сервісу окремо. (рис. 2.7, 2.8) Подальшими кроками є інтеграція цих даних у систему KSU24.

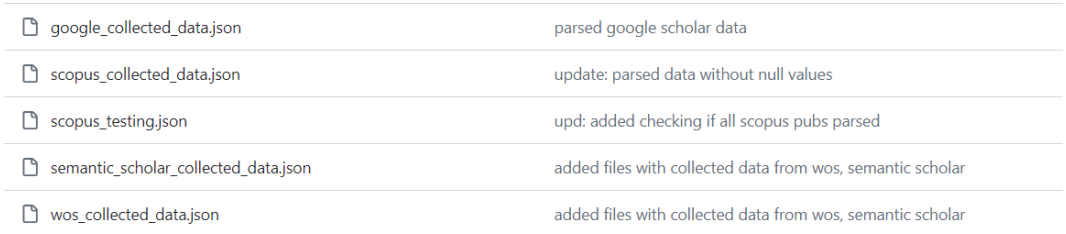


*Рисунок 2.7 JSON-файли зібраної інформації з наукометричних баз*

```
{
  "3eaa8f39-1af3-483f-b8f3-90265bcb8f59": {
    "author": {
      "author_id": 57130603400,
      "citation_count": 11,
      "document_count": 7,
      "cited_by_count": 5,
      "prism_url": "http://api.elsevier.com/content/author/author_id/57130603400",
      "h_index": 2,
      "eid": "9-s2.0-57130603400",
      "orcid": "0000-0003-2153-2367",
      "subject_areas": [
        {
          "category": "Medicine (all)",
          "flag": "true",
          "abbreviation": "MEDI",
          "code": "2700"
        },
        {
          "category": "Neuroscience (all)",
          "flag": "true",
          "abbreviation": "NEUR",
          "code": "2800"
        },
        {
          "category": "Physiology (medical)",
          "flag": "true",
          "abbreviation": "MEDI",
          "code": "2737"
        },
        {
          "category": "Physical Therapy, Sports Therapy and Rehabilitation",
          "flag": "true",
          "abbreviation": "HEAL",
          "code": "3612"
        },
        {
          "category": "Physiology",
          "flag": "true",
          "abbreviation": "BIOC",
          "code": "1314"
        }
```


*Рисунок 2.8 JS0N-файл Scopus*

Розроблені парсери мають перевагу перед сервісом Publication в тому, що вони не обмежуються збором загальної інформації про науковця, а додатково збирають розширену інформацію, як про публікації так і про науковця. Такий підхід надає повний об’єм інформації, яка необхідна для проведення підрахунку балів за критеріями, в яких враховується публікаційна активність науково-педагогічного працівника. До того ж розробка парсерів дає інформацію, для подальшого вдосконалення процесу рейтингування та розробки елементів контролю.

Дані в базі модель наукометричного профілю рейтингування НПП має в собі наступні поля, які містять інформацію, зібрану за допомогою парсерів: ID профілю, індекс Хірша, кількість документів, кількість цитувань, платформа з переліку Scopus, Google Scholar, Web of Science, Semantic Scholar, ORCID. Модель стаття науковця має наступні поля: назва, автори, DOI – ідентифікатор

цифрового об’єкту. Дані імпортуються завдяки трьом різним файлам, які відповідають за імпорт даних з Scopus, Web of Science, Google Scholar.

Сформовані сторінки наукометричного профілю рейтингування НПП має подібне наповнення до наповнення наукометричного профілю НПП у сервісі Publication.(рис.2.9)

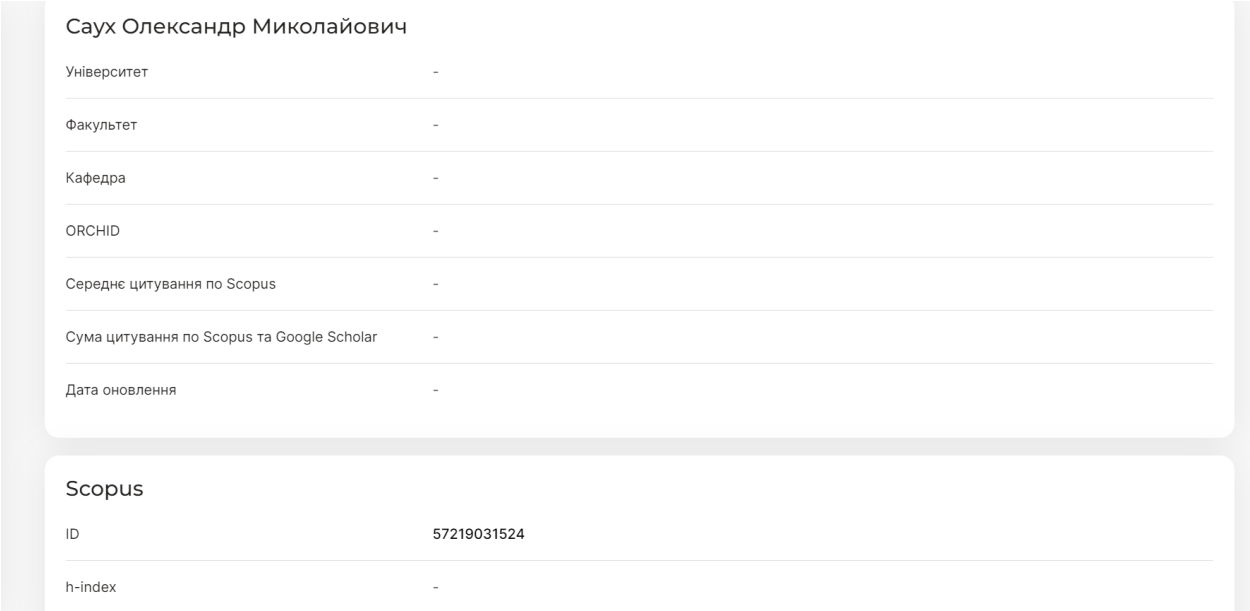


*Рисунок 2.9 Наукометричний профіль НПП*

# ВИСНОВКИ

Під час виконання кваліфікаційної роботи на тему «Розроблення парсерів наукометричних баз даних для інтеграції даних у KSU24» було виконано аналіз сервісу рейтингування Publication, проведено дослідження процесу рейтингування університету, аналіз взаємодії з сервісами Scopus, Google Scholar, Web of Science, Semantic Scholar, аналіз цих сервісів на предмет парсингу, вибір інструментів парсингу, побудування UML діаграм класів та розробка парсерів наукометричних баз.

Проведене дослідження процесу рейтингування показало, що процес може бути вдосконалений шляхом автоматизації певних процесів, для спрощення роботи працівників університету, економії часу та підтримки цифровізації університету, яка активно розвивається від початку заснування платформи KSU24.

Розроблені парсери наукометричних баз даних спростять процес збору та обробки інформації відповідними відділами, зокрема покращать та спростять роботу науково-педагогічних працівників, які щорічно заповнюють певні форми з інформації про їх наукову діяльність та публікування у наукометричних базах даних. Під час збору інформації було заміряно час, який витрачається на отримання інформації по всім наданим НПП: процес займав близько 1 години для збору інформації про 106 науково-педагогічних працівників, що у порівнянні із ручним введенням має значну перевагу.

Подальше вдосконалення може мати у собі виправлення помилок, які виникали під час парсингу, обробку відповідей та форматування інформації, виявлення повторень та встановлення нових зв’язків

Варто також зазначити, що у рамках цифровізації університету та вдосконалення і розширення можливостей сервісу KSU24 Херсонського державного університету розроблення парсерів є вдалим кроком, який спрощує роботу учасників процесу рейтингування університету на першому

етапі – етапі збору інформації. Подальшими кроками є інтеграція цих даних у систему та обробка, нарахування балів та формування рейтингів.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Електронний сервіс рейтингування Publication. URL: https://www.kspu.edu/About/DepartmentAndServices/Library/Actual/ScientometricActivity/ElectronicRatingServicePublication.aspx (дата звернення 24.02.2024)
2. Сервіс Publication. URL: http://publication.kspu.edu/ (дата звернення: 25.03.2024)
3. Закон України «Про вищу освіту» [Редакція від 24.03.2024 р.]; [Електронний ресурс] // Сайт Верховної Ради України. – Режим доступу: https://zakon.rada.gov.ua/laws/show/1556-18.
4. «Положення про систему рейтингового оцінювання діяльності науково-педагогічних працівників, кафедр та факультетів Херсонського державного університету» Івано-Франківськ – Херсон, 2023
5. Python. (n.d.). Офіційний веб-сайт мови програмування Python. URL: https://www.python.org/ (дата звернення 20.02.2024)
6. Наукометричні бази даних. URL: http://www.nbuv.gov.ua/node/1367 (дата звернення 20.02.2024)
7. Elsevier Developer Portal. URL: https://dev.elsevier.com/ (дата звернення 09.03.2024)
8. Elsevier Developer Portal. URL: https://dev.elsevier.com/academic_research_scopus.html (дата звернення 09.03.2024)
9. Що таке Selenium. URL: https://drukarnia.com.ua/articles/sho-take-selenium-0KhP- (дата звернення 08.03.2024)
10. Google Scholar API. URL: https://serpapi.com/google-scholar-api (дата звернення 21.02.2024)
11. ScholarlyORG Quickstart. URL: https://scholarly.readthedocs.io/en/stable/quickstart.html (дата звернення 01.02.2024)

12. API-інтерфейси Web of Science Group. URL: https://clarivate.com/cis/solutions/xml-%d0%b8-api-%d0%b8%d0%bd%d1%82%d0%b5%d1%80%d1%84%d0%b5%d0%b9%d1%81%d1%8b/ (дата звернення 26.03.2024)

13. Semantic Scholar. URL: https://www.semanticscholar.org/ (дата звернення 09.03.2024)

14. Нове відео для пошуку літератури: Semantic Scholar URL: https://entc.com.ua/uk/1131-nove-video-dlia-poshuku-literatury-semantic-scholar (дата звернення 26.03.2024)

15. Academic Graph API (1.0) URL: https://api.semanticscholar.org/api-docs/ (дата звернення 26.03.2024)

16. Що таке URL адреса сайту? URL: https://hostiq.ua/blog/ukr/what-is-url/ (дата звернення 26.03.2024)

17. Що таке JSON. Усе про цей формат передачі даних в інтернеті. URL: https://apix-drive.com/ua/blog/useful/scho-take-json (дата звернення 24.03.2024)

18. Креневич А.П. Python у прикладах і задачах Частина 2. Об’єктно-орієнтоване програмування: навч. посібник Київ, 2020. 152 с.

19. David Thomas, Andrew Hunt The Pragmatic, 2019

20. Що таке квартиль журналу та де і як його знайти? URL: https://science.snau.edu.ua/wp-content/uploads/2020/02/RCO-informs_3_%D0%A9%D0%9E-%D0%A2%D0%90%D0%9A%D0%95-%D0%9A%D0%92%D0%90%D0%A0%D0%A2%D0%98%D0%9B%D0%AC-%D0%96%D0%A3%D0%A0%D0%9D%D0%90%D0%9B%D0%A3-%D0%A2%D0%90-%D0%94%D0%95-%D0%86-%D0%AF%D0%9A-%D0%99%D0%9E%D0%93%D0%9E-%D0%97%D0%9D%D0%90%D0%99%D0%A2%D0%98.pdf (дата звернення 29.02.2024)

21. PEP8 – Style Guide for Python Code. URL: https://peps.python.org/pep-0008/ (дата звернення 25.03.2024)