Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)
Кафедра комп'ютерних наук
(повна назва кафедри)

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня

магістр
(назва освітнього ступеня)

на тему: Дослідження платформ керування даними та інструментів візуалізації для аналітики наукових досліджень

Виконав: студент VI курсу, групи САм-61
спеціальності 124 Системний аналіз
(шифр і назва спеціальності)

_____ Бєлоусов К.К.
(підпис) (прізвище та ініціали)

Керівник _____ Дуда О.М.
(підпис) (прізвище та ініціали)

Нормоконтроль _____ Мацюк О.В.
(підпис) (прізвище та ініціали)

Завідувач кафедри _____ Боднарчук І.О.
(підпис) (прізвище та ініціали)

Рецензент _____ Жаровський Р.О.
(підпис) (прізвище та ініціали)

Тернопіль
2022

Міністерство освіти і науки України
Тернопільський національний технічний університет імені Івана Пулюя

Факультет комп'ютерно-інформаційних систем і програмної інженерії
(повна назва факультету)

Кафедра комп'ютерних наук
(повна назва кафедри)

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Боднарчук І.О.
(підпис) (прізвище та ініціали)
«_19_» _грудня_ 2022 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

на здобуття освітнього ступеня Магістр
(назва освітнього ступеня)

за спеціальністю 124 Системний аналіз
(шифр і назва спеціальності)

Студенту Бєлоусову Казимиру Казимировичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження платформ керування даними та інструментів візуалізації для аналітики наукових досліджень

Керівник роботи Дуда Олексій Михайлович, к.т.н., доцент кафедри КН
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджені наказом ректора від «_22_» _листопада_ 2022 року № _4/7-947_

2. Термін подання студентом завершеної роботи 16 грудня 2022р.

3. Вихідні дані до роботи Наукові публікації про зберігання даних, платформи керування даними, інструменти візуалізації даних та аналітичне опрацювання даних

4. Зміст роботи (перелік питань, які потрібно розробити)
Вступ. 1 Стан досліджень в галузі платформ керування даними. 2 Системний аналіз типів інформаційних колекцій та наборів даних, методів їх аналітичного опрацювання в наукових Дослідженнях. 3 Моделювання та використання платформ керування даними та інструментів візуалізації для аналітики наукових досліджень. 4 Охорона праці та безпека в надзвичайних ситуаціях. Висновки. Додатки.

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень, слайдів)
1 Титульна сторінка. 2 Тема, Мета, Об'єкт, Предмет дослідження. 3 Завдання дослідження. 4 Актуальність дослідження. 5 Усереднена тривалість наукових проектів в різних областях досліджень. 6 Перелік термінів пошукових запитів. 7 Ієрархічна діаграма публікацій інформаційно-технологічні платформи та набори даних в наукових дослідженнях. 8 Критерії оцінювання інформаційно технологічних платформ для зберігання даних в наукових дослідженнях. 9 Критерії оцінювання інформаційно технологічних платформ для зберігання даних в наукових дослідженнях. 10 Поширені методи аналітичного опрацювання колекцій та наборів великих даних у наукових дослідженнях. 11 Ескіз архітектури системи на основі мікросервісного підходу. 12 Використання інструментів графічного подання даних. 13 Висновки. 14 Завершальний слайд.

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Мацюк О.В., доцент	08.12.2022	09.12.2022
Безпека в надзвичайних ситуаціях	Клепчик В.М., ст. викладач	10.12.2022	11.12.2022

7. Дата видачі завдання 14 листопада 2022 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1.	Ознайомлення з завданням до кваліфікаційної роботи	14.11.2022-15.11.2022	*Виконано*
2.	Підбір наукових джерел про зберігання даних,	16.11.2022-20.11.2022	*Виконано*
	платформи керування даними, інструменти		
	візуалізації даних та аналітичне опрацювання даних		
3.	Переклад та опрацювання наукових джерел про	21.11.2022-23.11.2022	*Виконано*
	зберігання даних, платформи керування даними,		
	інструменти візуалізації даних та аналітичне		
	опрацювання даних		
4.	Виконання дослідження щодо платформ керування	24.11.2022-27.11.2022	*Виконано*
	даними, інструментів візуалізації даних та		
	аналітичного опрацювання даних для наукових		
	досліджень		
5.	Оформлення розділу «Стан досліджень в галузі	28.11.2022-30.11.2022	*Виконано*
	платформ керування даними»		
6.	Оформлення розділу «Системний аналіз типів	01.12.2022-04.12.2022	*Виконано*
	інформаційних колекцій та наборів даних, методів їх		
	аналітичного опрацювання в наукових дослідженнях»		
7.	Оформлення розділу «Моделювання та використання	05.12.2022-07.12.2022	*Виконано*
	платформ керування даними та інструментів		
	візуалізації для аналітики наукових досліджень»		
8.	Виконання завдання до підрозділу «Охорона праці»	08.12.2022-09.12.2022	*Виконано*
9.	Виконання завдання до підрозділу «Безпека в	10.12.2022-11.12.2022	*Виконано*
	надзвичайних ситуаціях»		
10.	Оформлення кваліфікаційної роботи	12.12.2022-13.12.2022	*Виконано*
11.	Нормоконтроль	14.12.2022-15.12.2022	*Виконано*
12.	Перевірка на плагіат	15.12.2022	*Виконано*
13.	Попередній захист кваліфікаційної роботи	16.12.2022	*Виконано*
14.	Захист кваліфікаційної роботи	20.12.2022	

Студент _____ Бєлоусов К.К.
(підпис) (прізвище та ініціали)

Керівник роботи _____ Дуда О.М.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Дослідження платформ керування даними та інструментів візуалізації для аналітики наукових досліджень // Кваліфікаційна робота освітнього рівня «Магістр» // Бєлоусов Казимир Казимирович // Тернопільський національний технічний університет імені Івана Пулюя, факультет комп'ютерно-інформаційних систем і програмної інженерії, кафедра комп'ютерних наук, група САм-61 // Тернопіль, 2022 // С. 71, рис. – 25, табл. – 3, кресл. – 14, додат. – 1, бібліогр. – 61.

Ключові слова: аналіз, інтеграція, дані, дослідження, доступність, пошук, сумісність, робочий процес.

Кваліфікаційна робота присв'ячена дослідженню платформ керування даними та інструментів візуалізації для аналітики наукових досліджень. В першому розділі кваліфікаційної роботи описано прогресивні методи збирання даних в процесах наукових досліджень. Виконано пошук та аналіз наукових публікацій щодо платформ даних в наукових дослідженнях. Проаналізовано кількісні показники наукових публікацій щодо платформ даних в наукових дослідженнях. Розглянуто критерії оцінювання інформаційно технологічних платформ для зберігання даних в наукових дослідженнях. В другому розділі кваліфікаційної роботи описано типи інформаційних колекцій та наборів даних, що використовуються в сучасних наукових дослідженнях. Досліджено поширені методи аналітичного опрацювання колекцій та наборів великих даних у наукових дослідженнях. Розглянуто системну платформу для інтеграції даних, засобів візуалізації та аналітичного опрацювання. В третьому розділі кваліфікаційної роботи описано встановлення та налаштування програмно-алгоритмічних елементів системних платформ керування даними, візуалізації та аналітичного опрацювання. Проаналізовано результати наукових розвідок.

ANNOTATION

Research on Data Management Platforms and Visualization Tools for Science Analytics // Qualification work of the educational level "Master" // Bielousov Kazymyr Kazymyrovych // Ternopil National Technical University named after Ivan Pulyuy, Faculty of Computer Information Systems and Software Engineering, Department of Computer Science, SAm-61 group // Ternopil, 2022 // P. 71, fig. - 25, tables - 3, chair. - 14, annexes - 1, references. - 61.

Key words: analysis, integration, data, research, accessibility, search, interoperability, workflow.

The qualification work is dedicated to researching data management platforms and visualization tools for scientific research analytics. In the first section of the qualification work, progressive methods of data collection in the processes of scientific research are described. A search and analysis of scientific publications on data platforms in scientific research was carried out. Quantitative indicators of scientific publications regarding data platforms in scientific research were analyzed. The criteria for evaluating information technology platforms for data storage in scientific research are considered. The second section of the qualification paper describes the types of information collections and data sets used in modern scientific research. Common methods of analytical processing of collections and sets of big data in scientific research are studied. A system platform for data integration, visualization tools and analytical processing is considered. The third section of the qualification work describes the installation and configuration of software and algorithmic elements of system platforms for data management, visualization and analytical processing. The results of scientific investigations were analyzed.

ПЕРЕЛІК СКОРОЧЕНЬ І ТЕРМІНІВ

AUC (англ. Area Under ROC Curve) – Площа під кривою ROC.

API (англ. Application Programming Interface) – Прикладний програмний інтерфейс.

LIMS (англ. Laboratory information management system) – Система управління лабораторною інформацією.

RMI (англ. Remote Method Invocation) – Програмний інтерфейс виклику віддалених методів.

SAN (англ. Storage Area Network) – Мережа зберігання даних.

SSH (англ. Secure SHel) – Мережевий протокол рівня застосунків, що дозволяє проводити віддалене управління комп'ютером і тунелювання TCP-з'єднань.

QC (англ. Quality Control) – Контроль якості.

WSDL (англ. Web Services Description Language) – Мова опису інтерфейсів вебсервісу.

XFS – Високопродуктивна журнальована файлова система.

ЗМІСТ

ВСТУП

Актуальність теми. Навчальні заклади мають пропонувати системні інформаційно-технологічні рішення, які найкраще задовольняють потреби користувачів і забезпечують їм конкурентну перевагу при проведенні наукових досліджень. Тому потрібно спроєктувати, розробити, запровадити та підтримувати обширний перелік обчислювальних інструментів для створення великих обсягів даних з забезпеченням обслуговування потреб та запитів тисяч активних користувачів та науково-дослідних проєктів. Водночас особливу увагу слід приділити відтворюваності результатів. На даний час всі процеси наукових досліджень, зокрема, формування дослідницької гіпотези, проведення експериментів, вимірювань, власне саме дослідження та аналіз даних, керується невеликою кількістю експертів у різних наукових галузях. Проте здатність виконувати повне дослідження даних у режимі реального часу з використанням окремого персонального комп’ютера часто обмежена завдяки неоднорідності програмного забезпечення, форматів та структури даних і величиною їх розмірів. Це суттєво впливає на структуру та інформаційно-технологічну архітектуру використовуваного стеку програмно-алгоритмічних засобів. Тому сучасним науковим напрямком є дослідження платформ керування даними та інструментів візуалізації для аналітики наукових досліджень.

Мета і задачі дослідження. Метою даної кваліфікаційної роботи освітнього рівня «Магістр» є підвищення рівня повноти подання інформації платформ керування даними та інструментів візуалізації при проведенні науково-дослідних робіт. Для досягнення поставленої мети було потрібно виконати наступні завдання:

– Проаналізувати стан досліджень в предметній області платформ керування даними та інструментів візуалізації.

– Дослідити критерії оцінювання інформаційно технологічних платформ для зберігання даних в наукових дослідженнях.

– Проаналізувати типи інформаційних колекцій та наборів даних, що використовуються в сучасних наукових дослідженнях.

– Виконати порівняння існуючих системи ідентифікації відбитків пальців.

– Розробити архітектуру систем та платформи для інтеграції даних, засобів візуалізації та аналітичного опрацювання.

Об’єкт дослідження процеси збирання, зберігання та опрацювання даних при виконання науково-дослідних робіт.

Предмет дослідження: методи зберігання, аналітичного опрацювання та візуалізації науково-дослідних даних.

Наукова новизна одержаних результатів кваліфікаційної роботи полягає у тому, що отримав подальший розвиток метод формування узагальненої архітектури систем збирання, опрацювання та зберігання науково-дослідних даних.

Практичне значення одержаних результатів. Виконано макетування та прототипування системи для збирання, опрацювання та зберігання науково-дослідних даних.

Апробація результатів магістерської роботи. Основні результати проведених досліджень обговорювались на X науково-технічної конфції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя (м. Тернопіль, 2022 р.).

Публікації. Основні результати кваліфікаційної роботи опубліковано у двох працях конференції (Див. додатки А).

Структура й обсяг кваліфікаційної роботи. Кваліфікаційна робота складається зі вступу, чотирьох розділів, висновків, списку літератури з шістдесят одного найменування та одного додатку. Загальний обсяг кваліфікаційної роботи складає 71 сторінку, з них 46 сторінок основного тексту, який містить 25 рисунків та 3 таблиці.

1 СТАН ДОСЛІДЖЕНЬ В ГАЛУЗІ ПЛАТФОРМ КЕРУВАННЯ ДАНИМИ

1.1 Прогресивні методи збирання даних в процесах наукових досліджень

Основні інформаційно-технологічні сутності та об’єкти спрямовані на підтримку наукових досліджень, щороку регулярно створюють терабайти необроблених даних, які потрібно збирати, зберігати, опрацьовувати та в подальшому можна архівувати. Ці дані анотуються, попередньо обробляються [1], контролюються на відповідність якісних характеристик [2] та аналізуються засобами розроблених та реалізованих на даний час конвеєрів. Крім того, для забезпечення цілісності та повноти всього спектру інформаційно-технологічних процесів підтримки та супроводження, створюються та формуються звіти, котрі потім надсилаються клієнтам. Зазвичай, цикли збирання, попереднього опрацювання управління та аналізу даних проводяться паралельно для декількох науково-дослідницьких проектів з різними дослідницькими питаннями. При цьому залучають декількох експертів і вчених з різних наукових галузей, наприклад в царині біології, медицини, хімії, статистики, інформатики та аналітики. На рисунках 1.1 – 1.4 зображено ключові етапи типових робочих процесів проведення наукових досліджень в «omics»-галузях – це сучасні галузі біології в яких проводиться аналіз груп подібних елементів. При цьому відбувається відбір, зберігання та аналітичне опрацювання великих за обсягом наборів та колекцій даних. На рис. 1.1 [3] проілюстровано підготовку біологічного матеріалу та зразків. Зображено процеси підготовки зразків різних біологічних організмів і тканин, зокрема подано, свіжозаморожені людські тканини з визначеними раковими та доброякісними ділянками тканин, модельні організми [4].

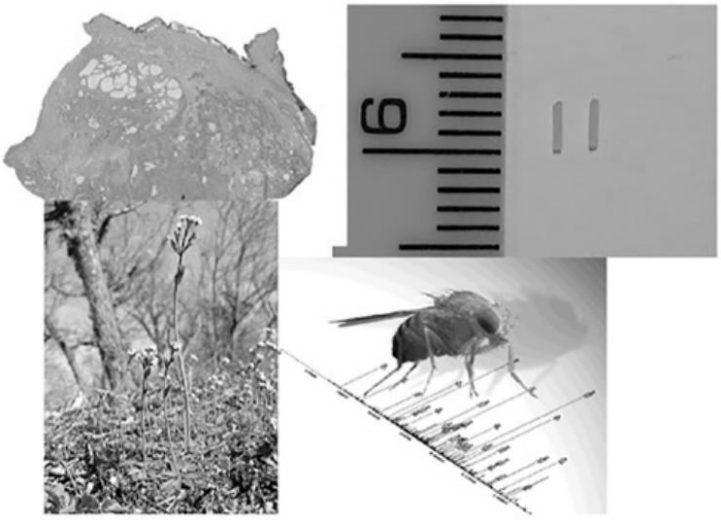


Рисунок 1.1 – Процеси підготовки біологічного матеріалу та зразків

А на рис. 1.2 [3] проілюстровано проведення наукових експериментів в іноваційних «omics»-галузях. Зокрема, зображено процеси секвенування та мас-спектрометрії, котрі є типовими методами відбору даних для виконання вимірювань на основі попередньо підготовлених зразків у «omics»-дослідженнях.

Рисунок 1.2 – Процеси проведення наукових експериментів в іноваційних «omics»-галузях науки

На рис. 1.3 [3] подано ключові етапи анотування та архівування даних в процесі іноваційних наукових досліджень.

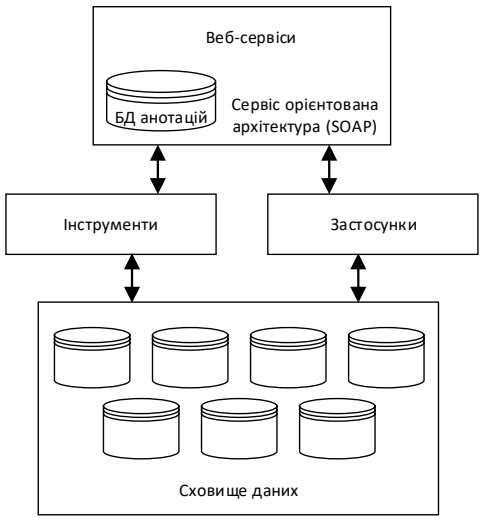


Рисунок 1.3 – Структурна схема процесу відбирання, анотування та архівування даних під час іноваційних наукових досліджень

При цьому обробка та архівування даних та метаданих є важливими етапами наукових досліджень з особливим наголосом на відтворюваності результатів вимірювань

А на рис. 1.4 [3] проілюстровано процес видобування корисних даних та знань в іноваційних «omics»-галузях науки.

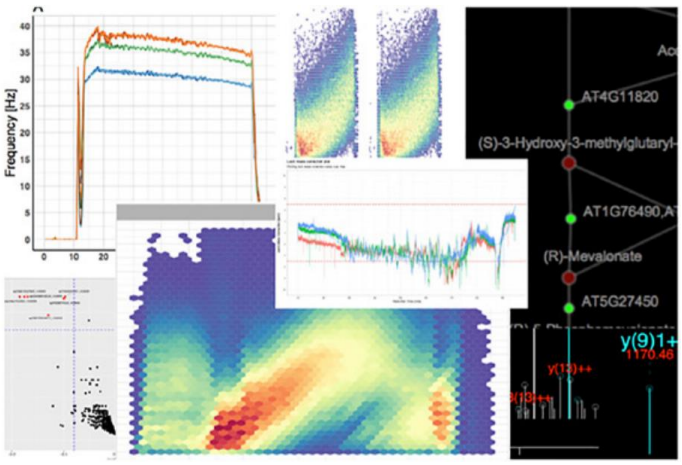


Рисунок 1.4 – Видобування корисних даних та знань в іноваційних «omics»-галузях науки

При цьому відбувається спеціалізоване застосування загальновідомих інструментів біоінформатики та методів візуалізації даних, розроблених науковою спільнотою, з використанням відображення не лише вихідних наборів та колекцій даних, але й пов’язаних з ними метаданих.

Доцільно зазначити, що переважна більшість сунаукових науково-дослідних проектів тривають впродовж декількох років. На рисунку 1.5 зображено типову тривалість «omics»-проєктів [3].

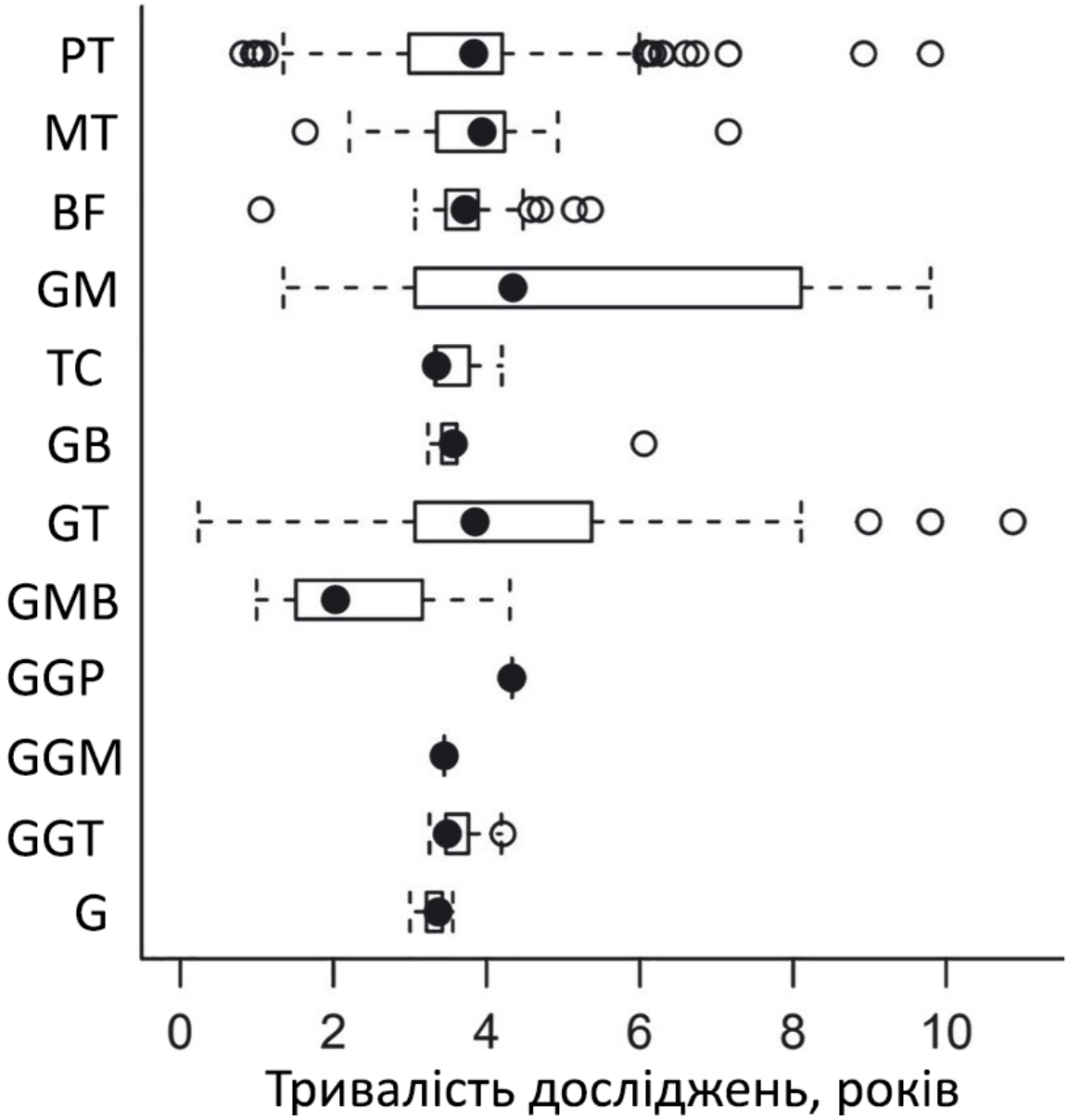


Рисунок 1.5 – Усереднені тривалість наукових проектів в різних областях досліджень

При цьому позначено інноваційні галузі наукових досліджень:

«PT» – протеоміка;

«MT» – метаболоміка;

«BF» – біофізика;

«GM» – геноміка;

«TC» – транскриптоміка;

«GB» – геноміка/транскриптоміка + метаболоміка/біофізика;

«GT» – геноміка + транскриптоміка;

«GMB» – загальна метаболоміка/біофізика;

«GGP» – загальна геноміка/транскриптоміка + протеоміка;

«GGM» – загальна геноміка/транскриптоміка + метаболоміка/ біофізика;

«GGT» – загальна геноміка/транскриптоміка;

«G» – загальні.

А на рисунку 1.6 зображено кількість учасників науково-дослідного проєкту [3], усереднену внаслідок спостережень протягом майже двох десятиліть.

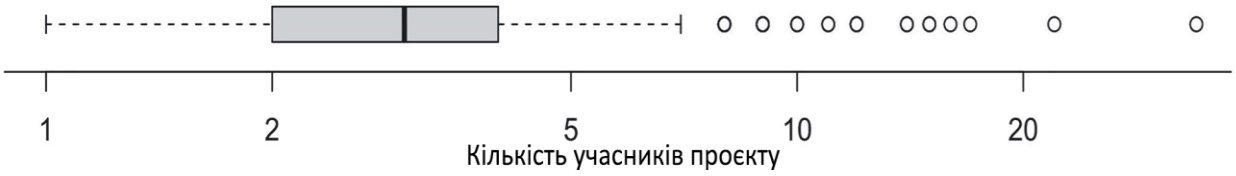


Рисунок 1.6 – Усереднені кількість учасників наукових проектів у різних областях дослідження

Міждисциплінарний характер сучасних наукових проєктів потребує використання потужних та гнучких інформаційно-технологічних платформ для високопродуктивного збирання, анотування та зберігання даних, оперативного спілкування між учасниками проєкту, ефективного забезпечення процесів аналітичного опрацювання та дослідження даних.

Інноваційні інтерактивні методи візуалізації та подання інформації [5] підходять для підтримки дослідницької фази сучасних наукових проєктів. Однак загальна підтримка сучасними загальнодоступними інформаційно-технологічними платформами широкого переліку програм такого роду часто ускладнюється з різних причин, зокрема:

– Неоднорідність структури накопичених та оброблюваних даних, відсутність ефективного впровадження стандартів. Це не дозволяє легко інтегрувати та об’єднувати існуючі на даний час програмно-алгоритмічні засоби.

– Програмно-алгоритмічні застосунки-прототипи здебільшого не можливо ефективно розгортати та обслуговувати з використанням розширених налаштувань інформаційно-технологічних платформ та систем. Зазвичай це необхідно для масового використання та запровадження у широкосерійному виробництві.

– Високий рівень початкових порогових витрат у відношенні на одного користувача-дослідника [6].

– Сучасні фахівці та вчені в окремих наукових галузях в переважній більшості зазвичай мають обмежені навички програмування.

На рисунку 1.7 відображено графік, сформований на основі сукупних даних проведених циклів наукових досліджень за кількістю паралельних дослідницьких проектів опрацювання блоків мас-спектрометрії [3] на основі побудованого щомісячного ковзного вікна.

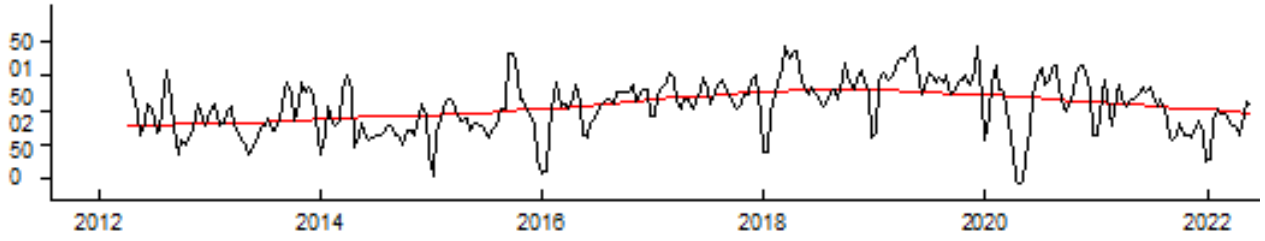


Рисунок 1.7 – Графік опрацювання блоків мас-спектрометрії в наукових дослідженнях

На рисунку 1.8 відображено графік обсягів агрегованих файлів даних накопичених внаслідок проведених циклів наукових досліджень [3] на основі побудованого щомісячного ковзного вікна.

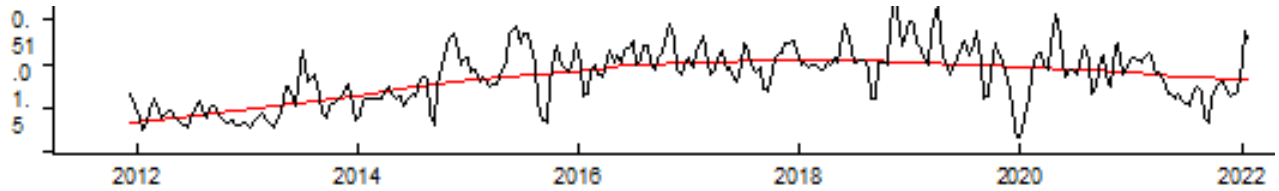


Рисунок 1.8 – Обсяги агрегованих файлів даних, терабайт (Тб)

На рисунку 1.9 подано графік обсягів архівованих даних внаслідок проведених циклів наукових досліджень [3] на основі побудованого щомісячного ковзного вікна.

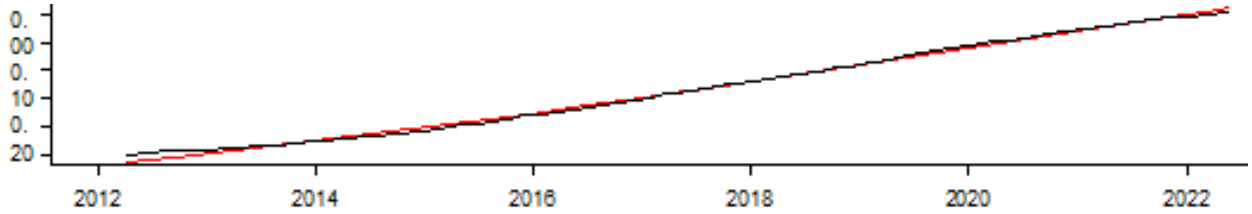


Рисунок 1.9 – Обсяги архівованих даних, петабайь (Рб)

Впродовж останнього періоду часу спостерігалось явище, що традиційні базові IT-розробки не можуть впоратися зі швидкістю зміни циклів розвитку програмних засобів аналітичного опрацювання даних [1]. Водночас застарівають існуючі інструменти аналітичного опрацювання даних, оскільки вони не повністю інтегровані в функціонуючі IT-системи та платформи. Тому, основні команди IT-розробників наукових проєктів мають ретельно оцінювати останні тенденції в галузі аналітичного опрацювання даних та спрогнозувати, які з них слід першочергово запроваджувати, враховуючи науково-дослідницькі та економічні фактори. Адже дослідники стикаються з фундаментальними складнощами щодо обробки та аналізу

даних через складність та різнорідність форматів накопичуваних даних в різних наукових областях.

1.2 Пошук та аналіз наукових публікацій щодо платформ даних в наукових дослідженнях

В процесі дослідження проведемо огляд академічної літератури на основі обмежених пошукових запитів до наукометричних баз даних Scopus та Web of Science (WoS).

В процесі пошуку виявлено понад сто п’ятдесят опублікованих наукових статей. Для первинного перегляду оберемо п’ятдесят найбільш відповідних статей. Це зроблено з метою забезпечення адекватного контексту та наукових розвідок щодо решти літературних джерел щодо джерел та платформ даних в наукових дослідженнях. Для систематизації огляду наукових публікаій потрібно визначити надійні методи формування пошукових запитів, протоколи виконання запитів та сформувати структуру оцінювання статей.

Протокол запиту для оцінювання наукових публікацій про інновації в галузі платформ даних та великих даних ппри проведенні наукових досліджень, доцільно сформувати на комплексному наборі умов запиту. Щоб забезпечити керований обсяг вибірки наукових публікацій, доцільно визначити часові рамки обмеживши їх останніми п’ятьма роками. Крім того, сформуємо набір термінів запитів, що стосуються платформ даних та великих даних в проведенні наукових досліджень на основі початкового огляду літератури.

Статистична комісія ООН [7] описує типи джерел великих даних та інформаційно-технологічні платформи для розміщення даних як сукупність різнотипових адміністративних наборів даних та інформаційно-технологічних засобів для їх зберігання. Зокрема, наприклад:

– персональні записи громадян (електронні медичні записи, медичні карти, відвідування лікарень, персональні страхові записи, банківські записи);

– комерційні або транзакційні дані (банківських та кредитних карток, онлайн-транзакції, у тому числі з мобільних пристроїв);

– дані сенсорних мереж (наприклад, супутникові зображення, дорожні та кліматичні давачі);

– дані пристроїв спостереження (наприклад, дані відстеження з мобільних телефонів та GPS);

– поведінкові дані (наприклад, онлайн пошук щодо продуктів, послуги або будь-якого іншого типу інформації, переглядів онлайн-сторінок);

– дані про думку чи поведінку громадян (наприклад, дані соціальних мереж).

Дослідники ідентифікують типи джерел великих даних, а також відповідні терміни запиту [8]. На основі проведеного огляду літератури сформуємо набір термінів для пошукових запитів, що стосуються інформаційно-технологічних платформ для зберігання даних наукових досліджень типів і джерел великих даних (див. таблицю 1.1).

Таблиця 1.1 – Перелік термінів пошукових запитів щодо інформаційно-технологічних платформ для зберігання даних і джерел великих даних

Джерело	**Терміни запиту (Назва, Анотація, Ключові слова)**
1	**2**
Загальне	«platform» або «big data» або «big earth data»
Давачі	«remote sensing» або «satellite» або «geospatial» або «earth observation» або «digital earth» або «sensor» або «smart meter»

Продовження таблиці 1.1

1	2
Пристрої спостереження	«smartphone»або «mobile phone» або «cell phone» або «GPS» або «global positioning system» або «digital science»
Трансакційний	«scanner data» або «transaction data»
Дані про поведінку чи думку	«data mining» або «text mining» або «social media data» або «Facebook data» або «Twitter data» або «Instagram data»

Використовуючи подані в табл. 1.1 терміни запиту, а також умовні пошукові терміни, пов’язані з інформаційно-технологічними платформами для зберігання наукових даних, проведемо початковий пошук у накометричних базах даних Scopus та WoS. В результаті виконання процедур пошуку сформуємо початковий список знайдених наукових публікацій із понад п’ятсот статей.

Після перевірки релевантності статтей, виокремимо дослідження, які спеціально розробляли нові інформаційно-технологічні платформи для зберігання та формування наборів даних для підтримки наукових досліджень. В процесі аналізу наукових публікацій виявлено додаткові високорелевантні або часто цитовані статті за допомогою методу сніжної кульки ключових цитат з ключових наукових публікацій. Загалом відіберемо сотню опублікованих наукових статей для більш детального аналізу, який проводитимемо з використанням узгодженої системи оцінювання, із формуванням шаблоні результатів. Зазначимо, що переглянуті та оцінені статті було опубліковано в понад сорока п’яти наукових виданнях. Видання охоплюють широкий перелік наукових напрямків та дисциплін, зокрема:

– геопросторові науки;

– науки про Землю;

– соціальні науки;

– медичні науки;

– наука про дані;

– інформаційні та комунікаційні технології.

Загалом понад сімдесят відсотків проаналізованих статей відповідали дванадцяти основним групам наукових журналів (див. рисунок 1.10) [9], у кожній з яких було опубліковано принаймні дві статті, включені до огляду.

Розміри та номери блоків відповідають кількості проаналізованих статей з кожного наукового журналу. Кольори відповідають основним дисциплінам:

– аква – багатопрофільні;

– оранжевий – геопросторові науки;

– зелений – науки про землю та навколишнє середовище;

– червоний – медичні науки.

На рисунку подано категорії:

– «ISPRS» відповідає журналам, виданим Міжнародним товариством фотограмметрії та дистанційного зондування;

– «Наука про землю» включає журнали «Nature», «Nature Climate Change» і «Nature Communications»;

– «Наукові дані» стосується журналу «Nature Scientific Data»;

– «PNAS» посилається на матеріали академій наук;

– «Науки про землю» містить наукові публікації щодо «морських наук» та «наук про довкілля».

Після класифікації та категоризації публікацій доцільно виокремити критерії оцінювання платформ для зберігання та опрацювання даних в наукових дослідженнях.

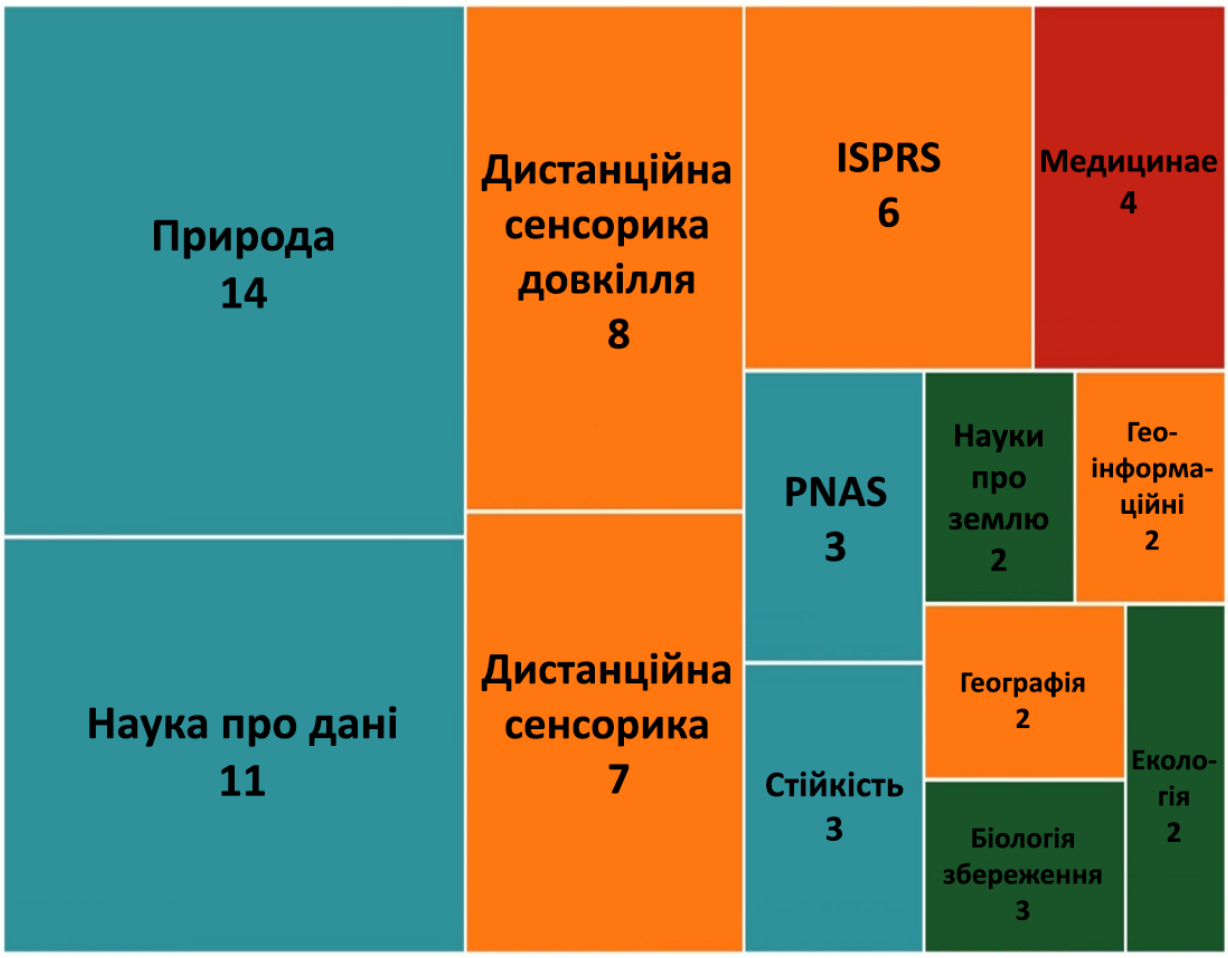
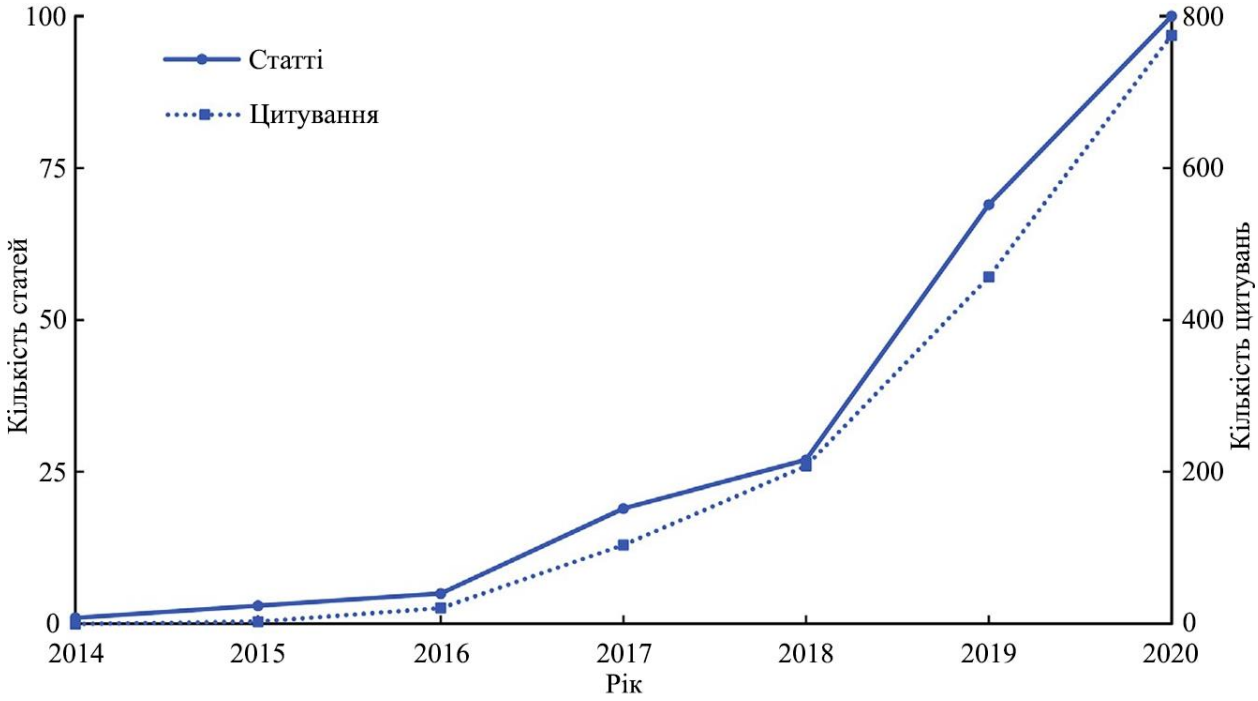


Рисунок 1.10 – Ієрархічна діаграма публікацій інформаційно-технологічні платформи та набори даних в наукових дослідженнях

1.3 Аналіз кількісних показників наукових публікацій щодо платформ даних в наукових дослідженнях

Кількість опублікованих робіт та їх цитування в наукометричних базах базах даних щодо платформ керування даними, інструментів візуалізації та аналітики наукових досліджень експоненційно зростає впродовж останнього періоду часу (див. рисунок 1.11).

У 2020 році зросла кількість опублікованих статей щодо платформ керування даними, інструментів візуалізації та аналітики наукових досліджень до сотні з приблизно сімсот семидесяти цитуваннями.



Рисцнок 1.11 – Кількість публікацій та цитувань результатів досліджень щодо платформ керування даними, інструментів візуалізації та аналітики наукових досліджень [10]

Згідно з інформацією про авторів, опубліковано результати досліджень в понад тридцяти країнах на шести континентах (див. рисунок 1.12).

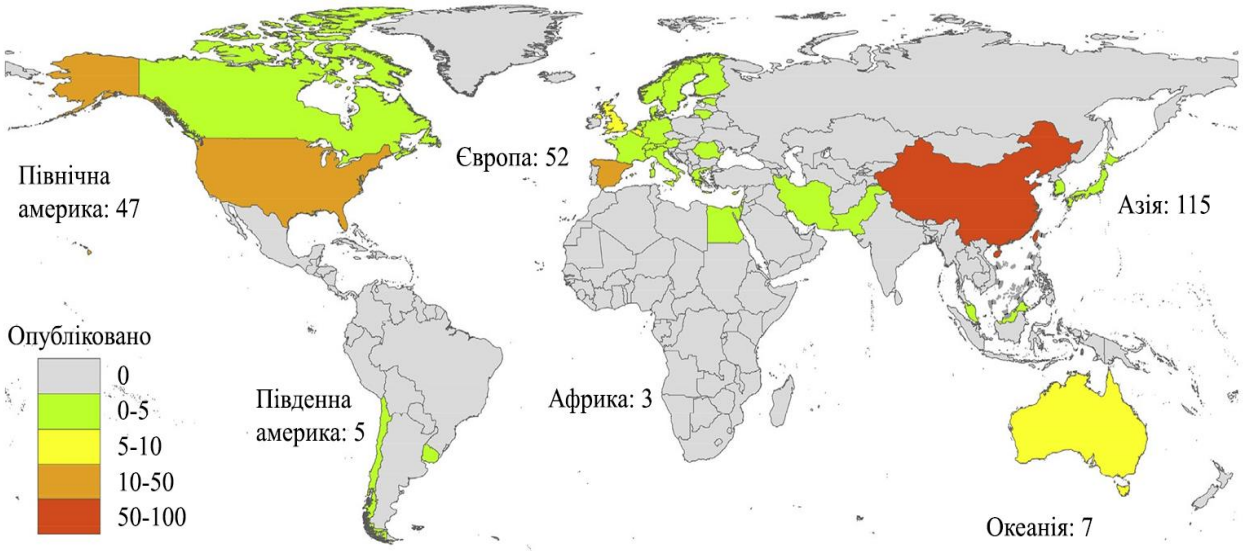


Рисунок 1.12 – Просторовий розподіл публікацій щодо платформ керування даними, інструментів візуалізації та аналітики наукових досліджень [10]

Найбільше результатів досліджень щодо платформ керування даними, інструментів візуалізації та аналітики опубліковано азіатськими вченими. Зокрема, 115 публікацій – це приблизно 50% від загальної кількості статей. На другому місці за кількістю публікацій, європейські вчені, зокрема, 52 статті – це приблизно 23%, та північноамериканські вчені, зокрема, 46 статей, або 21%. В перерізі окремих країн переважно автори з Китаю – 97 статей та США – 41 стаття. Це становить 43,3% та 18,3% від загальної кількості статей відповідно.

Статті щодо платформ керування даними, інструментів візуалізації та аналітики наукових досліджень були опубліковані в майже сто двадцяти наукових журналах. Наукові видання з найбільшою кількістю відповідних статей здебільшого зосереджені на галузях геонауки, науки про навколишнє середовище та екології (див. рисунок 1.13).

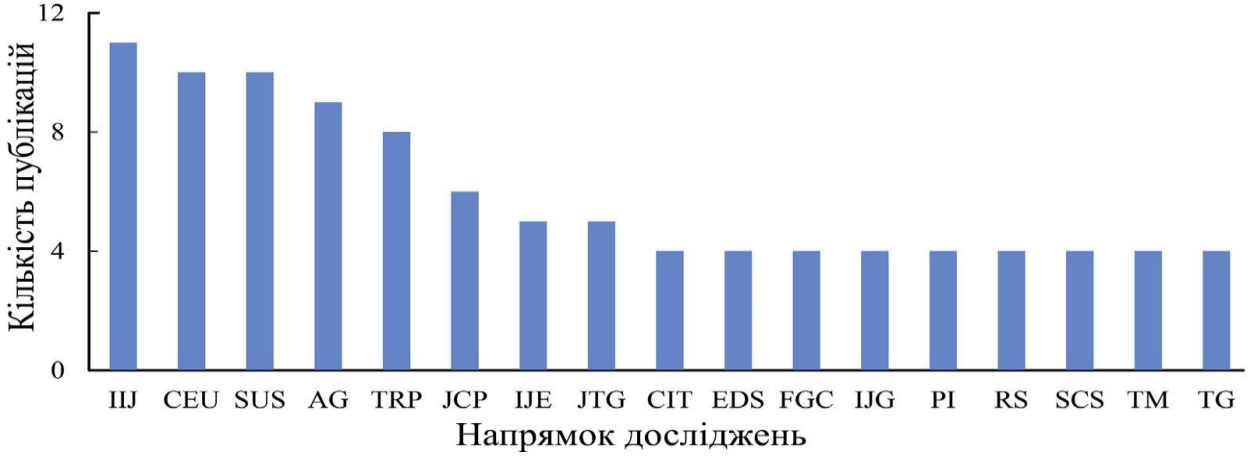


Рисунок 1.13 – Розподіл статей між різними тематичними напрямками наукових видань [10]

На рис. 1.3 позначено загальні тематичні напрямки наукових видань:

– «IIJ» – Міжнародний журнал геоінформації;

– «CEU» – комп’ютери, навколишнє середовище та міські інформаційні системи;

– «SUS» – стійкість інформаційних систем;

– «AG» – прикладна географія;

– «TRP» – дослідження в галузі транспорту;

В тому числі з категорії інноваційних технології позначено підкатегорії наукових видань:

– «JCP» – екологічно-чисте виробництво;

– «IJE» – дослідження навколишнього середовища та в галузі охорони здоров’я;

– «JTG» – транспортна географія;

– «CIT» – цитати;

– «EDS» – наука про дані («Data Science»);

– «FGC» – інноваційні комп’ютерні системи наступних поколінь;

– «IJG» – міжнародні журнали географічної інформації;

– «PI» – наукові праці IEEE;

– «RS» – дистанційне зондування;

– «SCS» – сталий розвиток міст та суспільних формацій;

– «TM» –туризм та менеджмент;

– «TG» – транзакційні та геоінформаційні системи.

Серед проаналізованих видань Міжнародний журнал геоінформації мав найбільшу кількість публікацій – приблизно п’ять відсотків від загальної кількості опублікованих та проаналізованих статей, зокрема понад десять статей. В категоріях комп’ютери, навколишнє середовище та міські інформаційні системи та стійкість інформаційних систем опубліковано по десять статей, тобто по 4,5%.

1.4 Критерії оцінювання інформаційно технологічних платформ для зберігання даних в наукових дослідженнях

На основі проведеного аналізу опублікованих статей сформовано логічну структуру (див. рис. 1.13) для видобування інформації про ключові

характеристики інформаційно-технологічних платформ для зберігання та опрацювання наборів та колекцій даних і великих даних. При цьому слід звернути увагу на:

– відповідні цілі та індикатори тематичних наукових досліджень;

– типи використаних джерел великих даних;

– застосованих аналітичних методів та підходів до опрацювання даних;

– залучення зовнішніх співробітників та спонсорів;

– комплексне оцінювання кожного набору даних на основі сформованого набору загальних критеріїв.

Щоб забезпечити основу для систематичного аналізу, спираючись на проведений початковий огляд наукових публікацій, виокремимо ключові фактори або критерії для оцінювання інформаційно-технологічних платформ, наборів та колекцій даних:

– географічний масштаб та покриття;

– зернистість;

– актуальність;

– періодичність/своєчасність;

– доступність, вартість та відтворюваність;

– точність та валідність;

– надійність;

– стійкість.

Розглянемо детальніше критерії, що використовуються для оцінювання інформаційно-технологічних платформ, наборів та колекцій даних.

Географічний масштаб та покриття відноситься до географічного масштабу, починаючи від місцевого, субнаціонального та національного, до регіонального та глобального. В контексті наукових досліджень цікавими є інформаційно-технологічні платформи та набори даних, які можуть

підтримувати національний моніторинг з більшим географічним охопленням, або глобальним охопленням.

Зернистість використовується для оцінювання рівня просторової дезагрегації з перевагою інформаційно-технологічних платформ, які можуть зберігати та опрацьовувати детальніші набори даних для звітності на нижчих адміністративних рівнях в межах окремих країн на основі просторового розподілу на регіональному, національному та нижчому адміністративному рівнях.

Актуальність відноситься до відповідності інформаційно-технологічних платформ для узгодження наборів даних з офіційними визначеннями тематичних критеріїв та індикаторів наукових досліджень і того, чи можна їх використовувати для формування та моніторингу офіційних індикаторів або надання часткових, проміжних або додаткових наборів даних.

Періодичність та своєчасність відноситься до частоти продукування, ґенерації та зберігання наборів даних засобами інформаційно-технологічних платформ, з перевагою щонайменше щорічного продукування. Частота оновлення наборів та колекцій даних може бути включена щорічно або частіше, тобто щоквартально, щомісяця, щотижня або впродовж менших періодів часу.

Доступність, вартість та відтворюваність стосується надання інформаційно-технологічними платформами вільного та відкритого доступу до накопичуваних та похідних наборів даних. При цьому також слід враховувати доступність необроблених вхідних наборів даних та вихідних програмних кодів моделей. Там, де це може бути доступно, також доцільно збирати інформацію про фінансові витрати та робоче навантаження.

Точність та валідність відноситься до чіткості забезпечуваних інформаційно-технологічними платформами методів і використання процедур перевірки, пошуку стандартних помилок або формування

статистики та накопичення точності, надійності, невизначеності даних та показників продуктивності використовуваними аналітичними засобами.

Надійність використовується при оцінюванні наукових публікацій, які використовуватимуться для пошуку інформаційно-технологічних платформ. При цьому вона відноситься до релевантності або репутації наукових видань, інтерпретованих на основі імпакт-факторів журналів.

Стійкість інформаційно-технологічних платформ характеризує, чи будуть набори даних підтримуватися та оновлюватися в перспективному майбутньому, інтерпретуватися на основі доступності, чітких планів оновлення та дослідницької співпраці з міжнародними організаціями, технічними партнерами та спонсорами.

На основі «Фундаментальних принципів офіційної статистики» [11], статистичні агенції обирають відповідні джерела даних зважаючи на їх якість, своєчасність, витрати та навантаження на респондентів. Тім та Ван Халдерен [12] описують набір міркувань та критеріїв для прийняття рішень щодо використання інформаційно-технологічних платформ для накопичення та опрацювання великих даних. При цьому оцінюється їхня порівняльна вартість, робоче навантаження, достовірність та стабільність статистичних результатів і довгострокову перспективу. Флореску [13] опублікував відомості щодо використання інформаційно-технологічних платформ для зберігання та постачання наборів та колекцій великих даних. Автор розклядає множину критріїв їх оцінювання, зокрема, якість даних, їх точність, актуальність, узгодженість, інтерпретацію та своєчасність тощо. підкреслюють, що використання великих даних має приносити. Витрати характеризуються структурованістю та обширним переліком очевидних переваг статистичних параметрів якості та даних, ступеня доступності та вартість якості. Водночас Ван ден Хомберг [14] наводить множину факторів для оцінювання інформаційно-технологічних платформ для роботи з даними,

зокрема, своєчасність, надійність джерела, структура вмісту та описує метод оцінювання точності, деталізації та географічного покриття.

Подані вище вісім критеріїв, використано для формування аналітичної структури з метою системного огляду наукових досліджень для видобування інформації щодо кожного дослідження. Це дозволить сформувати доволі простий підхід до оцінювання критеріїв якості інформаційно-технологічних платформ для роботи з даними з використанням шкали від одного до чотирьох або відповіді на кшталт «так/ні». Структуру доцільно включити до шаблонів звітності, який забезпечить логіку процесів видобування ключової інформації та характеристик щодо кожної інформаційно-технологічної платформи та набору даних і для підтримки процесів подальшого оцінювання та синтезу результатів.

1.5 Висновок до першого розділу

В першому розділі кваліфікаційної роботи освітнього рівня «Магістр» описано прогресивні методи збирання даних в процесах наукових досліджень. Розглянуто процес пошуку та аналізу наукових публікацій щодо платформ даних в наукових дослідженнях. Проаналізовано кількісні показники наукових публікацій щодо платформ даних в наукових дослідженнях. Подано критерії оцінювання інформаційно технологічних платформ для зберігання даних в наукових дослідженнях.

2 СИСТЕМНИЙ АНАЛІЗ ТИПІВ ІНФОРМАЦІЙНИХ КОЛЕКЦІЙ ТА НАБОРІВ ДАНИХ, МЕТОДІВ ЇХ АНАЛІТИЧНОГО ОПРАЦЮВАННЯ В НАУКОВИХ ДОСЛІДЖЕННЯХ

2.1 Типи інформаційних колекцій та наборів даних, що використовуються в сучасних наукових дослідженнях

Інформаційні колекції та набори даних, що використовуються в сучасних наукових дослідженнях можна класифікувати на різні типи відповідно до множини їх характеристик та різних критеріїв, зокрема джерел даних, середовища отримання даних та структури даних [15]. Вони використовуються для опису стану факторів наукових досліджень, котрі часто проводяться для міського середовища в режимі реального часу, економічної діяльності та моделей поведінки громадян. Запропонована дослідниками [15] класифікаційна схема використана для розрізнення двох широких категорій великих даних:

– дані про середовище;

– дані про поведінку людини.

В процесі аналізу опублікованих в статтях даних у цих двох категоріях визначено десятки інформаційно-технічних платформ – джерел великих даних. Вони відрізнялися змістом та характеристиками даних (див. рисунок 2.1) [10]. На рисунку позначено:

– «POI» – визначні місця;

– «OSM» – відкриті географічні мапи;

– «RSI» – зображення дистанційного зондування;

– «SVI» – візуальні зображення фізичних сутностей та архітектурних об’єктів;

– «OBED» – інші дані про архітектурне середовище;

– «CND» – дані стільникових мереж;

– «SMD» – дані соціальних мереж;
– «V-GPS» – дані GPS-позиціонування транспортних засобів;
– «BSCD» – дані громадського транспорту;
– «OHBD» – інші дані про поведінку громадян.

Рисунок 2.1 – Великі дані, що використовуються в сучасних наукових дослідженнях [10]

З лівого боку рис. 2.1 позначено даці щодо середовища. З правого боку позначено людино-залежні дані.

Розглянемо ключові характеристики великих даних щодо середовища. Інформація про «POI» включає координати широти та довготи, назви та адреси. Дані «POI» відображають різні галузі та функції і широко використовуються для оцінки життєздатності різнотипових систем, місць їх функціонування, перевірки обсягів та типів використовуваних ресурсів [16].

Платформа даних «OSM» [17] – це відкрита інформаційно-технологічна платформа даних з веб-інтерфейсом, яка дозволяє користувачам

легко завантажувати дані про дорожні мережі [17]. Платформу «OSM» можна використовувати для вивчення процесів функціонування та доступності міського транспорту і класифікації процесів землекористування.

Платформи «RSI» в основному використовуються зберігання даних зображень дистанційного зондування високої роздільної здатності («HRSI») і даних нічного освітлення («NLD»). «HRSI» можуть отримувати дані про земної поверхні з високою роздільною здатністю. «NLD» може ідентифікувати урбаністичні території та відображати інтенсивність людської діяльності [18]. Дані «RSI»-платформ важливі для широкомасштабних наукових досліджень сталого розвитку в різних галузях.

Дані в категоріях «SVI» доволі точно відображають особливості 360-градусного антропогенного урбаністичного середовища [19]. Однак такі дані доволі важко збирати, отримувати та обробляти засобами існуючих інформаційно-технологічних платформ.

«OBED»-дані в основному включають тривимірну інформацію про будівлі та споруди. Зокрема державні та корпоративні дані щодо споживання енергії та ресурсів, планування урбаністичного будівництва, документації щодо земельних ресурсів та корисних копалин тощо, великі екологічні дані та великі метеорологічні дані [20].

Колекції «CND»-даних записує час кожного окремого голосового виклику та розташування зони базової приймально-передавальної станції мобільного зв’язку, де відбувається обмін даними [21]. «CND»-даних мають ряд переваг, зокрема, широке охоплення населення та висока роздільна часова здатність. Проте просторова роздільна здатність, приблизно 200м в урбаністичному ланшафті, залежить від розташування базової приймально-передавальної станції.

Колекції та набори даних з категорії «SMD» в основному надходять від «Facebook», «Twitter», «Flickr» і «Sina Weibo» – це китайський еквівалент «Twitter». Вони містять реєстраційні дані користувачів, твіти, фотографії та

іншу інформацію. «SMD»-дані характеризуються точним гео-позиціонування та високу просторову роздільну здатність у режимі реального часу, але часова роздільна здатність залежить від частоти публікацій користувача. Недоліком колекцій та наборів «SMD»-даних є недостатнє та нерівномірне представництво різних груп населення. Ці дані можуть бути ефективним джерелом для аналізу емоцій та формування прогнозів щодо особистої думки [22].

Колекції та набори «V-GPS»-даних можна умовно розділитись на дві категорії: дані траєкторії та дані «OD» – початку та місця призначення, які записують місцезнаходження транспортного засобу, час, швидкість, маршрут поїздки та інші атрибути. Ці колекції даних мають відносно високу просторову та часову роздільну здатність. Проте переважній більшості дослідників важко отримати доступ до них та обробити [23].

Інформаційно-технологічні платформи, що використовуються для накопичення, зберігання та опрацювання колекцій «BSCD»-даних записують ідентифікатори власників транспортних карток, тип, час і місце посадки та виходу з маршрутного транспортного засобу [24]. Ці дані мають ряд переваг, зокрема, хороші показники безперервності, широке охоплення та динамічне оновлення. Однак просторова роздільна здатність доволі сильно залежить від розташування платформи даних, а самі дані доволі важко отримати й обробити.

Колекції та набори «OHBD»-даних ґенеруються в основному внаслідок людської діяльності, зокрема, споживання продуктів та послуг з використанням Інтернет-ресурсів, записи щодо пошукових запитів, дані споживання UnionPay, відеоспостереження, інформація щодо здоров’я громадян, споживання ресурсів та енергії [20]. Отримати такі дані дослідникам здебільшого дуже складно.

Інтелектуальне середовище – це *«простір, створений людиною, у якому люди живуть, працюють і відпочивають щодня»* [25]. Дані про архітектурне

середовище – це об’єктивні дані, що описують різні атрибути урбаністичного архітектурного середовища [26]. Такі дані отримують в основному з відкритих платформ наукових даних, дистанційного зондування, урядів, підприємств. Вони використовуються для дослідження та розуміння основних характеристик урбаністичного середовища, зокрема для формування моделей міської інфраструктури, транспортних потоків, динамічного розвитку урбаністики.

2.2 Поширені методи аналітичного опрацювання колекцій та наборів великих даних у наукових дослідженнях

Аналітичне опрацювання колекцій та наборів великих даних безпосередньо починається з процесу виявлення потенційної цінності елементів даних [27]. Основні методи аналітичного опрацювання великих даних:

– класифікація;
– кластеризація;
– регресія;
– аналіз правил асоціації;
– аналіз соціальних мереж [28].

Різні методи аналітичного опрацювання надають дослідникам можливість збирати дані з різних точок та напрямків досліджеень і можуть забезпечити різні результати. Вони часто асоціюються з певними типами великих даних і мають конкретні цілі [10] (див. таблицю 2.1). Опишемо кожен з методів.

Класифікація – це метод поділу наукових наборів та колекцій великих даних на попередньо визначені групи відповідно до певних правил або атрибутів [28]. Зокрема, дерева рішень, опорні векторні машини та наївні правила Байєса на даний час є відносно зрілими алгоритмами класифікації.

Таблиця 2.1 – Методи аналітичного опрацювання великих даних

Категорія колекцій даних	**Використовувані методи аналітичного опрацювання**
POIs	Описова статистика, класифікація, оцінка щільності ядра, Getis-Ord Gi, векторна модель обмежених клітинних автоматів
OSM	Класифікація, векторна модель обмежених клітинних автоматів
RSI	Метод індексу дистанційного зондування, хмарні обчислення (Google Earth Engine), машинне навчання
SVI	Сегментація та розпізнавання зображень, машинне навчання
OBED	Описова статистика, кластеризація, регресія, машинне навчання
CND	Кластеризація, регресія, машинне навчання.
SMD	Описова статистика, кластеризація, регресія, просторовий автокореляційний аналіз, модель класифікації текстів, класифікація, аналіз правил асоціації, машинне навчання, аналіз соціальних мереж
V-GPS	Кластеризація, регресія, машинне навчання, хмарні обчислення
BSCD	Класифікація, кластеризація, регресія
OHBD	Кореляційний аналіз, регресія, машинне навчання, глибоке навчання

Методи на основі класифікації допомагають дослідникам класифікувати та розрізняти окремі дані на основі їхніх атрибутів та виявляти при цьому їхні регулярні моделі та характеристики. Ця техніка аналітичного опрацювання великих даних зазвичай використовується для аналізу даних

інфраструктури «розумних» міст. Зокрема, структурних міських сутностей та об’єктів («POI»), відкритих карт вулиць («OSM») та різнотипових колекцій текстових даних. Наприклад, дані POI можна розділити на різні категорії, щоб кількісно охарактеризувати елементи інфраструктури «розумного» міста та виділити різні типи міського землекористування [18]. Текстові дані можна класифікувати за відповідними правилами, щоб зрозуміти емоційні коливання окремих громадян або соціальних груп.

Методи на основі кластеризації – це методи групування подібних або тісно пов’язаних сутностей, об’єктів та структур даних для досягнення ефекту диференціації «подібні в групах, відмінні між групами» [28]. До методів кластеризації великих даних відносяться:

– вибіркова кластеризація;
– кластеризація з об’єднанням даних;
– кластеризація із зменшенням розмірності;
– паралельна кластеризація.

Кластеризація – це метод навчання без контролю, який використовується лише для отримання різних категорій, незалежно від того, чи є отримані категорії значущими. Ця техніка аналітичного опрацювання здебільшого використовується для розрізнення даних без інтегрованих мета-тегів та має перевагу поділу даних без категорій [29]. Наприклад, для аналітичного опрацювання дані мобільних мереж та дані GPS транспортних засобів можуть бути проаналізовані в «розумних» містах шляхом кластеризації, щоб виявити моделі розподілу населення та моделі переміщення громадян у різний час, забезпечуючи основу для планування міських громадських та транспортних мереж.

Регресія – це контрольований метод навчання, який кількісно визначає зв’язки між двома або більше змінними. Лінійна регресія та логістична регресія є двома найпоширенішими алгоритмами в цій категорії методів аналітичного опрацювання великих даних [28]. Регресія в переважній

більшості використовується для аналізу та прогнозування причинно-наслідкових зв’язків між сутностями з функціями зменшення розмірності, видобування інформації, оцінювання та прогнозування. Регресійні аналітичні моделі даних були створені для визначення факторів впливу на громадян резидентів певних захворювань на основі великих даних про здоров’я мешканців і великих даних про навколишнє середовище в урбаністичних районах проживання [20]. Дослідники використовують регресію для визначення потенційних факторів впливу на ціни на основі великих даних про навколишнє середовище та великих даних про нерухомість.

Аналіз правил асоціації – це метод пошуку кореляції або взаємозалежності між інформаційними сутностями даних. При цьому «апріорні» та «сірі асоціації» є двома найбільш використовуваними алгоритмами аналітичного опрацювання [30]. Аналіз правил асоціації може виводити логічні послідовності на основі виникнення певних подій, аналізувати та визначати ступінь впливу між інформаціними сутностями та виявляти кореляцію між цінними елементами даних. Цей метод в основному використовувався для аналізу взаємозв’язків між даними міської інфраструктури, даними дистанційного зондування та даними соціальних медіа, наприклад, для визначення зв’язку між факторами, що впливають на розширення «розумних» міст та показникми їх життєздатнрсті [16].

Аналіз соціальних мереж – це метод аналітичного опрацюваня даних, який використовує вузли та відношення для визначення та кількісного оцінювання соціальних зв’язків між обширними переліками інформаційних об’єктів на основі теорії мереж. З розвитком соціальних мереж цей метод набуває популярності та використовується для формування нових підходів до аналізу великих даних [31]. У контексті наукових досліджень цей підхід можна використовувати для аналізу напрямків та інтенсивності людських і матеріальних потоків у «розумних» містах. Водночас аналіз соціальних мереж використовується для аналізу відношень та взаємин між людьми,

спільнотами та інформаційними сутностями та фізичними об’єктами на основі даних соціальних мереж та GPS-даних [32].

2.3 Системна платформа для інтеграції даних, засобів візуалізації та аналітичного опрацювання

Основні інформаційні сутності потребують формування узагальненої системної архітектури платформи керування та обробки даних, щоб справлятися з процесами накопичення, підготовки, анотування, архівування та аналітичного опрацювання масивних наборів та колекцій даних. При цьому потрібно ефективно опрацьовувати зростаючі обсяги запитів сучасних науковихд осліджень справлятись з надзвичайною динамікою розвитку методів та інструментів аналітичного опрацювання даних, які можуть з’являтися та зникати з великою швидкістю. Успішне рішення цього завдання має розділити різні рівні задач та допомогти уникнути вузьких місць, спричинених обмеженою кількістю системних архітекторів, програмістів та експертів в галузі комплексного та спеціалізованого аналізу даних, при управлінні та використанні обмежених ресурсів зберігання даних та обчислювальних потужностей. При формуванні пройесів керування даними, важливо відокремлювати метадані від фактичного вмісту інформаційних сутностей даних.

Типові системні рішення в царині наукових досліджень зосереджені лише на конкретних практичних завданнях. Зокрема, інструменти аналізу підтримують функції підготовки та виявлення даних, які допомагають формувати, збагачувати та перетворювати дані наукових досліджень. Зокрема це:

- «TIBCO Spotfire» [33];
- «SBEAMS» [34];
- «myProMS» [35];

– «PeakForest» [36].

Аналітичні платформи зосереджуються на тому, щоб дозволити дослідникам, ануківцям та звичайним користувачам ефективно використовувати та конвеєрувати бажані інструменти для множини різноманітних типів даних. Зокрема це:

– «KNIME» [37];

– «TeraData» [38];

– «GenomeSpace» [39].

Аналітичні інструменти надають інформаційні технології для аналізу, візуалізації та обміну даними в наукових дослідженнях. Зокрема це

– «CLC Genomics Workbench» [40];

– «PanoramaWeb» [41];

– «Google Genomics».

Складні комплексні системи, зокрема, «Seven Bridges» [42], інтегрують базові функції керування інформаційно-технологічними проектами та елементами для організації та спільного використання імпортованих і оброблених наборів та колекцій даних у більш тонкий спосіб із контрольованим розмежуваннм прав та привілеїв доступу. Однак, наскільки нам відомо, на даний час не існує єдиного програмного забезпечення, яке б задовольняло всі потреби основного об’єкта наукових досліджень, а саме:

– адміністративні завдання, зокрема керування всією необхідною для користувачів та дослідників інформацією;

– проекти;

– замовлення;

– комп’ютери;

– інструменти;

– сховища;

– послуги;

– резервування інструментів;

– перевірка проектів;

– стягнення оплати плати за замовлення;

– бронювання засобів та обчислювальних потужностей;

– аналітичні завдання;

– збір метаінформації про інформаційні зразки та елементи;

– згенеровані необроблені дані;

– дані після обробки;

– відстеження всього робочого процесу обробки даних.

У Центрі функціональної геноміки Цюріха було розроблено систему «B-Fabric» [43], яка спрямована на задоволення наведеного переліку вимог. Подамо узагальнену архітектуру систем такого класу (див. рисунок 2.2), не вдаючись у додаткові деталі, які виходять за межі того, що необхідно для розуміння використання ключових принципів реалізації складних систем такого класу для ефективної реалізації засобів візуалізації та аналітичного опрацювання даних.

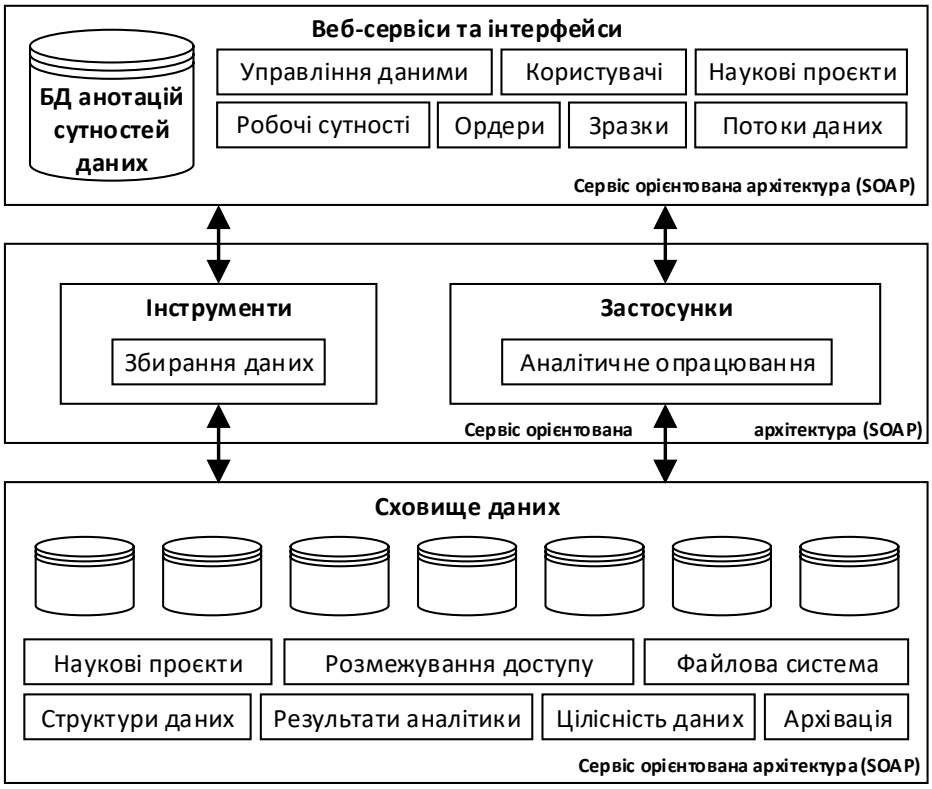


Рисунок 2.2 – Узагальнений ескіз архітектури системи

В поданій архітектурі системи експериментальні дані продукуються мікросервісними інструментами та програмно-алгоритмічними застосунками. Дані зазвичай зберігаються в різнотипових сховищах даних. Накопичені знання про ці дані, зокрема, про їх розташування, відстежуються з використанням комплексної бази метаданих. Інформаційно-технологічні інструменти та програмно-алгоритмічні комплекси використовують веб-сервіси для формування запитів та подачі інформаційних сутностей з бази метаданих. Всі відповідні структуровані дані зв’язуються разом на основі мікросервісної архітектури.

Продукування та генерування даних повністю відокремлено від загальної системи. Дані створюються, наприклад, за допомогою змінних інструментів, які зазвичай мають власні сховища. Виробники даних можуть бути зареєстровані в загальній системі. Експерти в галузі вимірювальних засобів працюють на цьому рівні, щоб установити та налаштувати інструменти за потреби чи на вимогу, а також зареєструвати їх у базі метаданих, при підключенні до загальної системи.

Бімери даних, які є спеціальними програмами, що працюють в оперативній пам’яті функціональних пристроїв, демультиплексують і передають отримані дані в авторизоване сховище відповідно до розмежування прав та привілеїв. Ці сховища даних повинні бути суворо організовані відповідно до загальновизнаних правил. Наприклад, дані зберігаються в папках та файлах, розділених в межах наукових проектів – так званих контейнерах, що представляють одиниці доступу для всіх даних, створених в проекті. На цьому рівні системні інженери, які відповідають за обслуговування різних сховищ, несуть відповідальність за впровадження бімерів даних на основі накопичених знань про відповідні окремі сховища.

База метаданих – це фактично мозок усієї системи [44], який містить усі знання, пов’язані з основними інформаційними та об’єктами. В ній відбувається документування проектів, замовлень на виділення ресурсів та

даних усіх користувачів. Саме тут зберігаються всі інформаційні сутності, оброблені в наукових проектах і фактичних замовленнях, а також відповідна інформація про застосовані процедури та робочі процеси. Завдяки програмно-алгоритмічним застосункам база метаданих також містить усю необхідну інформацію про створені файли та структури даних і, таким чином, забезпечує однорідну загальну картину в дуже різнорідних і автономних сховищах даних. Це дозволяє відповідати на запитання «Хто, коли створив які дані, у якому проекті, за допомогою яких зразків і за допомогою яких протоколів?» Розширена рольова модель реалізує детальний контроль доступу, щоб користувачі могли отримати доступ лише до тої частини світу даних, до якої їм надано відповідні прав та привілеї. Дані бази метаданих доступні через GUI-інтерфейс, а також через веб-сервіси. Концепції структурних сутностей доцільно реалізувати таким загальним чином, що база метаданих була готова зберігати будь-які тип даних майбутніх наукових досліджень.

Зовнішні програми та застосунки використовують метадані для доступу, обробки та візуалізації отриманих результатів досліджень. Необхідна попередня обробка даних, у тому числі обробка даних, зазвичай виконується на зовнішніх інфраструктурах – кластерах, сітках та хмарах [45]. Програмне забезпечення для візуалізації [46], є практичним прикладом реалізації зовнішніх по відношенню до даної системи програмно-алгоритмічних застосунків. Необхідні дані витягуються через веб-сервіси базової інфраструктури. За потреби веб-сервіси повинні зареєструвати результат аналізу в базі метаданих системи.

Важливим для поданої еволюційної архітектури системи є реалізація універсальної концепції застосунків, яка дозволяє впроваджувати динамічні спеціальні робочі процеси будь-якого типу та складності для відстеження всіх етапів процесів обробки даних. Використовуючи попередні та наступні зв’язки, усі програми знають, з якими даними вони можуть працювати [44].

2.4 Висновок до другого розділу

В другому розділі кваліфікаційної роботи досліджено типи інформаційних колекцій та наборів даних, що використовуються в сучасних наукових дослідженнях. Розглянуто поширені методи аналітичного опрацювання колекцій та наборів великих даних у наукових дослідженнях. Проаналізовано системну платформу для інтеграції даних, засобів візуалізації та аналітичного опрацювання.

3 МОДЕЛЮВАННЯ ТА ВИКОРИСТАННЯ ПЛАТФОРМ КЕРУВАННЯ ДАНИМИ ТА ІНСТРУМЕНТІВ ВІЗУАЛІЗАЦІЇ ДЛЯ АНАЛІТИКИ НАУКОВИХ ДОСЛІДЖЕНЬ

Змоделюємо результати проведеного системного аналізу з використанням «Jakarta EE», «JSF», «PrimeFaces», «OmniFaces» для процесів розробки веб-застосунків, «PostgreSQL» для керування даними, «Apache Lucene» для виконання процедур повнотекстового пошуку і «SOAP» для реалізації функціоналу веб-служб. У таблиці 3.1 наведено приклад характеристик використання систем подібного класу [1].

Таблиця 3.1 – Перелік демонстраційних характеристик використання систем

Користувачі	7410	Організації	350	Зразки	289 000
Проекти	4885	Інституції	795	Ресурси даних	2 021 670
Замовлення	24,000	Інструменти	320	Робочі одиниці	228 007
Послуги	1865	Додатки	260	Розмір дискового сховища	>0.4PtB

Продемонструємо наочно, як системна платформа може бути використана для реалізації спеціалізованого підключення програмного забезпечення для візуалізації даних інформаційних систем таким чином, щоб уся спільнота користувачів та дослідників могла отримати від практичні результати та користь. У наведеному прикладі використано середовище «R». Проте описаний інтерфейс можна використовувати в будь-якому іншому середовищі програмування, що підтримує мову опису веб-сервісів «WSDL». Перший приклад прикладної програми підключає інтерактивний інтерфейс користувача вищого рівня до керування даними. Другий приклад більш технічного характеру, щоб продемонструвати, як прикладний програміст

може склеювати дані, метадані та бібліотеки візуалізації засобами командного рядка «R».

3.1 Встановлення та налаштування програмно-алгоритмічних елементів системних платформ керування даними, візуалізації та аналітичного опрацювання

Перш ніж програмно-алгоритмічні засоби можна буде ефективно застосувати для даних, її потрібно коректно налаштувати. Системна платформа повинна дозволяти не лише анотувати метадані, але й надавати механізм анотації для підтримки конфігурування програмно-алгоритмічних елементів. Як показано на рисунку 3.1, системні платформи такого класу дозволяють реєструвати програми з неспецифічними властивостями, наприклад, типами або інформаційними технологіями [1].

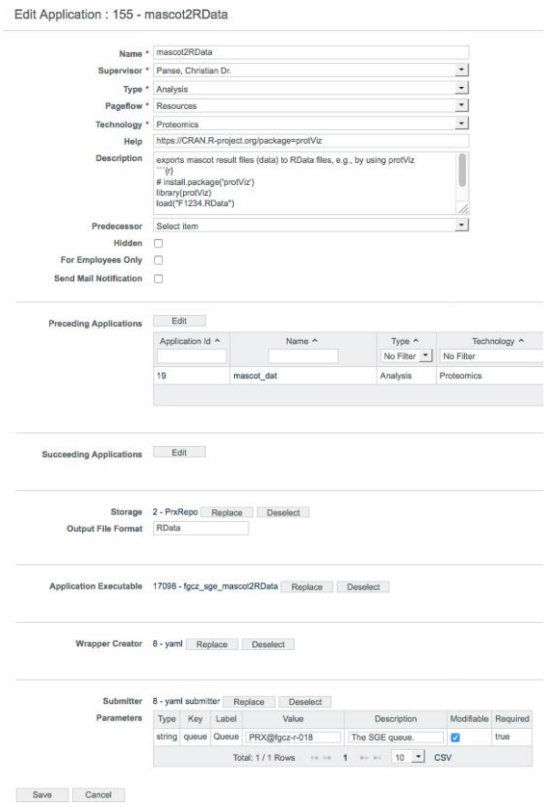


Рисунок 3.1 – Екран конфігурації програми для її інтегрування до системної платформи

Висока гнучкість систем такого класу досягається шляхом прикріплення виконуваних файлів, написаних зовні будь-якою мовою, та функціональних засобів для обгортання таких кодів попередньо зареєстрованими оболонками, що забезпечує необхідний та достатній рівень абстракції для виконуваних файлів програм з необхідними метаданими. Крім того, абстракція субмітера дозволяє запускати виконуваний файл програми в налаштованому середовищі обчислень за вибором дослідника [1].

Кожна програма приймає вхідні дані. У цьому контексті це також програма і надає вихідні дані. Наявність поле «Preceding Applications» визначає вхідні дані програм. Завдяки цьому стають реальними спеціальні робочі процеси. Ідентифікатор програми, показаний на рис. 3.1, вимагає вхідні дані програми пошуку та ідентифікації даних [47], яка має відповідний ідентифікатор програми. Залежно від обчислювальних потреб, завдання можуть виконуватися на локальному обчислювальному сервері або підключеній хмарній інфраструктурі [45]. Механізм створення оболонки реалізує постановку та формування завдань. Наведений на рисунку 3.2 XML-список [1] визначає середовище, у якому виконуються програмні засоби та який саме бінарний файл сценарію буде використано.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<executable id="17098">
 <created>2017-05-25T20:45:46.022+02:00</created>
 <createdby>cpanse</createdby>
 <modified>2017-05-30T10:32:16.394+02:00</modified>
 <modifiedby>cpanse</modifiedby>
 <status>available</status>
 <name>fgcz_sge_mascot2RData</name>
 <size>0</size>
 <context>APPLICATION</context>
 <description>
  Defines the bash script, see also
  http://fgcz-svn.uzh.ch/repos/sgeworker/bin/fgcz_sge_mascot2RData
 </description>
 <program>
  /home/bfabric/sgeworker/bin/fgcz_sge_mascot2RData
 </program>
 <supervisor id="482"/>
 <valid>true</valid>
 <version>0.0.2</version>
</executable>
```

Рисунок 3.2 – Приклад виконуваної XML-конфігурації

Також можна визначити попередньо задані значення аргументів alistofargument sand для передачі у виконуваний файл програми. Форма XML для виконуваної конфігурації виявилася простішою для підтримки та розгортання, особливо якщо програма потребує довгого списку параметрів користувача.

У прикладі з використанням «R shiny» (див. рисунок 3.3) всі завдання, створені окремою програмою, будуть доступні, якщо користувач авторизований в інтерфейсі платформи, наприклад, викликом модуля [1].

```
## https://github.com/fgcz/bfabricShiny/blob/bfabric11/R/bfabric.R
shinyServer(function(input, output, session) {
  bf <- callModule(bfabric, "bfabric8",
                   applicationid = c(155),
                   resoucepattern = 'RData$',
                   resourcemultiple = TRUE)
    resources <- bf$resources()$relativepath
...
}
```

Рисунок 3.3 – Приклад коду для ініціалізація модуля системи

При цьому отримані метадані, зокрема інформація про досліджувані зразки, сховище та шлях до файлу кожного вибраного інформаційного ресурсу, зберігаються в окремому об'єкті. Пізніше цей об'єкт можна використовувати як контейнер для будь-яких програмних цілей.

3.2 Системне використання засобів інтерактивної візуалізації даних

Розглянемо приклади для поданого в попередному розділі узагальненого опису системної платформи. Етап запису необроблених даних на сучасних дослідницьких пристроях, наприклад, мас-спектрометрах, може займати декілька днів в залежності від тривалості циклів досліджень, кількості досліджуваних зразків і процедур контролю якості, котрі запускаються за розкладом системного експерта. Під час первинного запису

дослідницьких даних, вони дані зберігаються на комп’ютері що керує вимірювальним приладом. Після завершення просесів вимірювань дані повинні автоматично копіюватися в сховище даних мікросервісами, які відповідають за передачу даних між вимірювальними приладами та сховищем даних.

Локальна копія даних залишатиметься як резервна копія протягом визначеного системою періоду часу на комп’ютері, підключеному до вимірювального приладу перед видаленням. Сховищем даних зазвичай є мережа зберігання даних (SAN), яка підключається до машин, що продукують набори та колекції даних. Передача даних в дослідницьких системах на даний час зазвичай здійснюються через оптоволоконну мережу. Для зберігання в переважній більшості випадків використовується файлова система «XFS». Безпосередньо після завершення процесів синхронізації даних із SAN запускається перший етап попереднього опрацювання зібраних даних, який перетворює власний необроблений формат окремих інформаційних файлів в загальний машинозчитуваний формат. Наприклад, загальний файл «Mascot» – «*.mgf» або «mzXML». В оптимальному випадку анотація на платформі даних вже буде зроблена до моменту ініціалізації процедур аналізу даних, але її потрібно завершити до того моменту, коли необроблені дані будуть «імпортовані» в системну базу метаданих. Де будуть розміщені відповідні пов’язані анотації необроблених наборів даних.

На наступному етапі, системні експерти або дослідники запускають процедури так званого «пошук у базі даних MS/MS». Наприклад, за допомогою «Matrix Science Mascot Server» [47] або «Comet» [48]. Ці алгоритми інтелектуального пошуку позначають корисні дані відповідними послідовностями. Кількісний показник отримується з площі під кривою (AUC) підрахунку корисних сигналів протягом часу проведення відповідного експерименту. Для подальшої обробки ці дані повинні бути перетворені в універсальний формат обміну, зазвичай сформований на основі XML [49].

Недоліком перетворення даних у XML-формат є дуже великий розмір отриманих файлів. При цьому різко зростає час синтаксичного аналізу. Практично, в подальшому, XML обробляється в більш типовий «R»- список. До нього, якщо необхідно, можна застосувати процедури фільтрації, а отримані дані можна зберегти як стиснутий «Rdata-контейнер» [50]. Подібні бібліотеки та процедури також доступні засобами мов програмування «Python», «Ruby» та «Perl». Після виконання цього етапу дані можна зчитувати безпосередньо в програмах аналітичного опрацювання.

Системний модуль «R-shiny» [51] обробляє всі завдання автентифікації та розміщення даних, використовуючи API-інтерфейси веб-сервісів системної платформи. Постановку потоку даних доцільно виконувати засобами мережевої файлової системи, зокрема, «NFS» чи «RFC7530», або протоколів безпечної оболонки «SSH» чи «RFC4254». Крім того, для мобільних комп'ютерів та портативних комп'ютерів можлива доцільність використання рішення на основі проксі-сервера. На рисунку 3.4 показано приклад використання однієї з системних платформ даних [1].

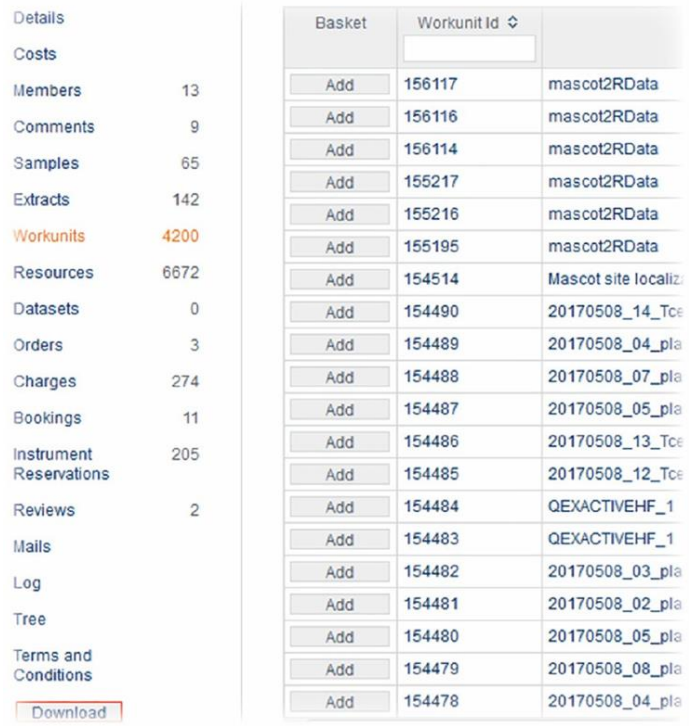


Рисунок 3.4 – Інтеграції наукових даних засобами системних платформ

Прикладне програмне забезпечення [47] виявляє та кількісно характеризує корисні дані та знання під час збору даних. Разом із результатами згаданого раніше алгоритму «пошуку в базі даних MS/MS» та динамічного графічного аналізу, це потужний інструмент для процесу прийняття прийняття рішень вченими та дослідниками.

На рис. 3.4 подано узагальнену статистику дослідницьких проектів, зокрема, кількість зразків і вимірювань, за допомогою запитів в базі даних «emeta» [1]. Окремі ресурси можна зібрати в кошик «Amazon» і передати як вхідні дані до відповідних аналітичних програм.

На рисунку 3.5 відображаються налаштування параметрів для вибраного програмно-алгоритмічного застосунку [1]. Найважливішими є вхідні дані, що містять опис шаблону даних.

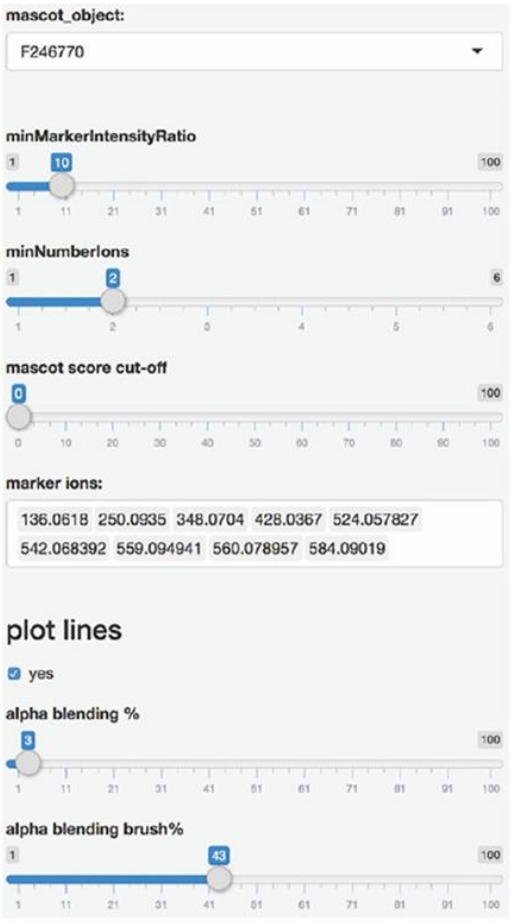


Рисунок 3.5 – Налаштування параметрів для вибраного програмно-алгоритмічного застосунку

Кожен елемент даних (див. рисунок 3.6) представляє обраний корисний сигнал на збільшеній діаграмі від часу утримування одного запуску приладу. Чорні ящики ідентифікують сигнали, де очікуваний фрагмент даних був виявлений за допомогою аналітичного пакету [52].

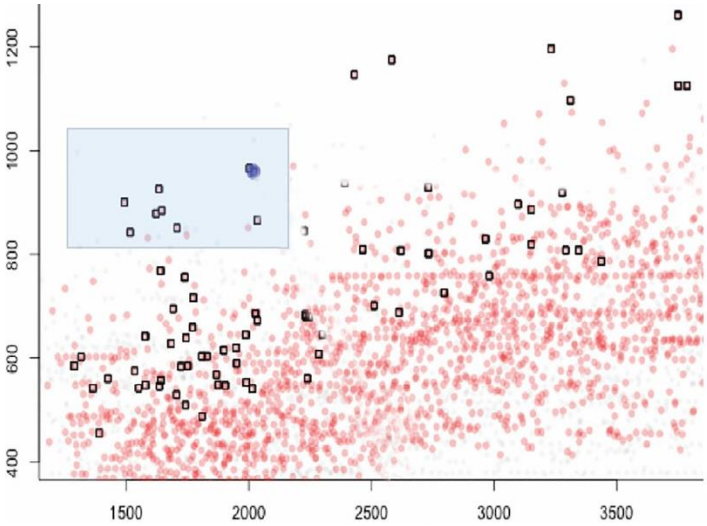


Рисунок 3.6 – Елементи даних та обрані корисні сигнали

Коробкові діаграми (див. рисунок 3.7) відображають розподіл кількості виявлених елементів даних [1]. Лінії, що з’єднують кожен елемент даних на коробковій діаграмі, зображують тенденцію видобування даних.

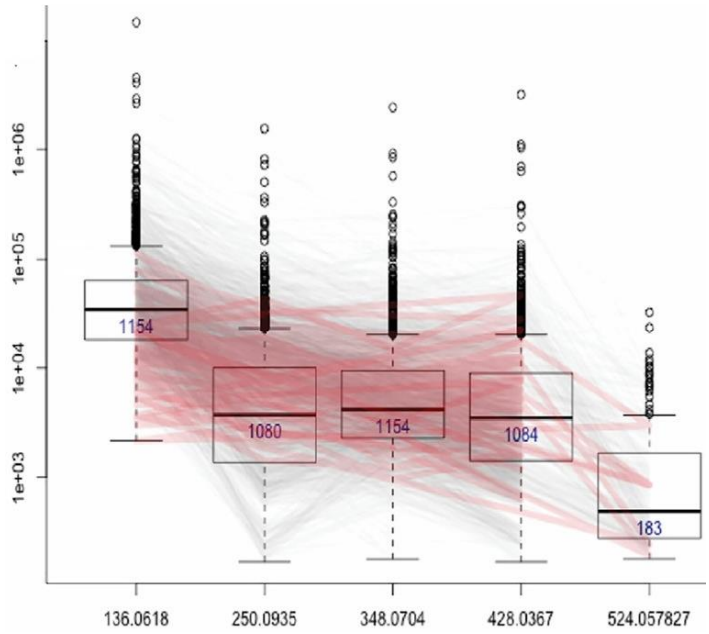


Рисунок 3.7 – Елементи даних та обрані корисні сигнали

На графіку видно, що найвищі піки асоціюються з окремими кластерами видобутих даних. Крім того виявлені дані підсвічуються відповідно до процедур контролю якості. Результат процесу дослідження та використані параметри можна зробити постійними, передавши їх назад у базу метаданих і сховище даних.

3.3 Ініціювання програмно-алгоритмічних засобів візуального дослідження

Розглянемо приклад, який ілюструє взаємодію між програмними елементамисистемної платформи, що виконують контроль якості та програмними аналітичними засобами рівня командного рядка за допомогою середовища «R». Таким чином досвідченіші користувачі можуть використовувати інтерфейс більшості системних платформ даних. Водночас, такий підхід демонструє, як можна ефективно задіяти складніші програмні засоби візуалізації даних. Конфігурація поданого на рисунку 3.8 фрагменту коду вимагає парольної автентифікації [1] веб-служби, яка використовується для доступу до системної платформи.

```
## R --no-save < code_snippet.R
## devtools::install_github("fgcz/bfabricShiny")
## devtools::install_github("fgcz/rawDiag")
stopifnot(R.Version()['major'] >= '4',
    require('rawDiag'),
    require('bfabricShiny'))

## Define B-Fabric input workunit
workunitid <- 165473
## Query metadata from B-Fabric
Q <- bfabricShiny::read(login, webservicepassword,
  endpoint = 'resource',
  query = list('workunitid' = workunitid), as_data_frame=FALSE)
## setting root directory
rawfilenames <- Q$res |>
    sapply(function(x)file.path('/Users/cp/Downloads', x$relativepath))
## Extract mass spectrometry data from BLOBs using proprietary software
## That requires storage access via SSH, NFS, or SAMBA.
RAW <- rawfilenames |>
    parallel::mclapply(rawDiag::read.raw, mc.cores = 12) |>
    base::Reduce(f = rbind)
## Print a summary
RAW |> rawDiag::summary.rawDiag()
## Have fun with visualization https://doi.org/10.1021/acs.jproteome.8b00173
## (a)
RAW |> rawDiag::PlotPrecursorHeatmap(bins = 25)
## (b)
RAW |> rawDiag::PlotPrecursorHeatmap(bins = 25) +
  ggplot2::facet_wrap(~ filename)
## (c)
RAW |> rawDiag::PlotTicBasepeak(method = 'overlay')
## (d)
RAW |> rawDiag::PlotInjectionTime(method = 'overlay')
```

Рисунок 3.8 – Фрагмент коду командного рядка «R» для виводу результатів

У прикладі розглядається менший набір даних, створений у рамках наукового експерименту [4]. Рядки з одинадцятого по тринадцятий виконують запит через WSDL до бази даних системної платформи. Цей запит повертає всі ресурси, що належать даному науковому проєкту. Робоча одиниця – це набір файлів. Для складнішого прикладу сформованого з використанням аналізу виразу можна отримувати додаткові метадані, зокрема, анотацію окремого елемента даних та індивідуальні засоби обробки для кожного вимірювання. Окрім механізму читання, для збереження або видалення результату аналізу можна використовувати методи веб-сервісів «save» та «delete». За допомогою методу збереження файли проміжних результатів можна завантажити на системну платформу. Наступна частина коду об’єднує кореневий каталог із витягнутим відносним шляхом до файлу кожного інфрмаційного ресурсу. Для поданого прикладу програмно-алгоритмічного елемента, щоб прочитати вміст файлів пропрієтарних інструментів і візуалізувати його. Тому доцільно використати R-пакет «rawDiag» [53] – програмний інструмент, який підтримує раціональну оптимізацію методів, надаючи налаштовані оператором діаграми діагностики рівня сканування метадані. Потім відбувається паралельне зчитування даних у форматі мас-спектрометрії. Цей метод може забезпечити механізм отримання даних через віддалений виклик методів (RMI) або за допомогою протоколу захищеної оболонки (SSH). Це виконується для уникнення копіювання всього файлу. На наступному етапі відбувається ґенерація графіку, показаного на рисунку 3.9 [1].

Дизайн візуалізацій ґенерації відповідає концепції, описаній у [54] з використанням [55]. Пакет «rawDiag» [53] можна використовувати окремо або підключати до системи за допомогою модуля «bfabricShiny» [51]. Він використовується як програма, що запускається або у веб-браузері або з командного рядка R [56].

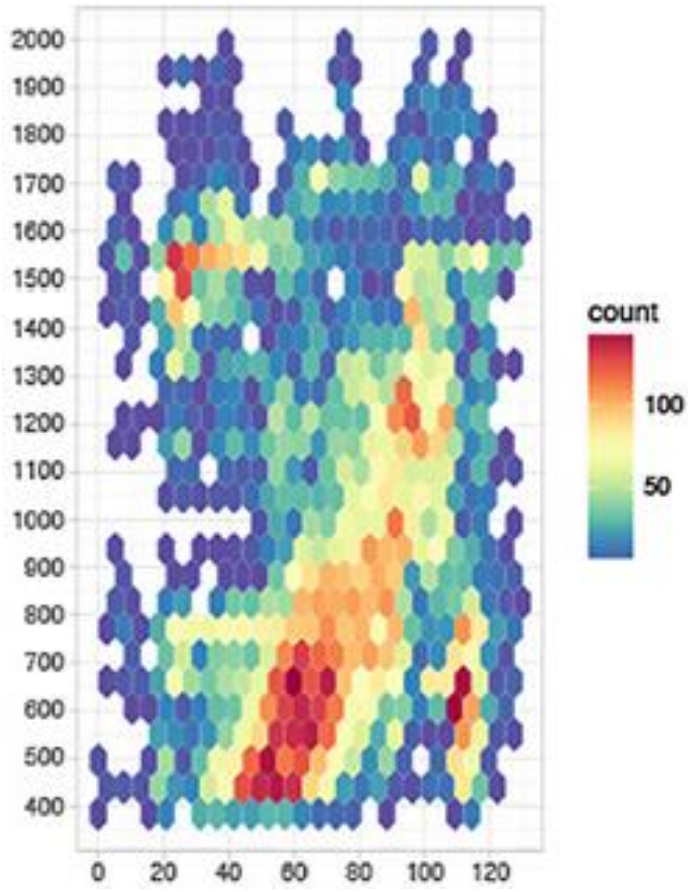


Рисунок 3.9 – Результати ґенерації графіку

Зображений на рисунку 3.10 графік можна використовувати для візуального представлення у вигляді набору ескізів більшого набору даних [57]. Для отримання вихідних даних доцільно використовувати дані, виміряні з використанням в досліджуване середовище пристроїв.

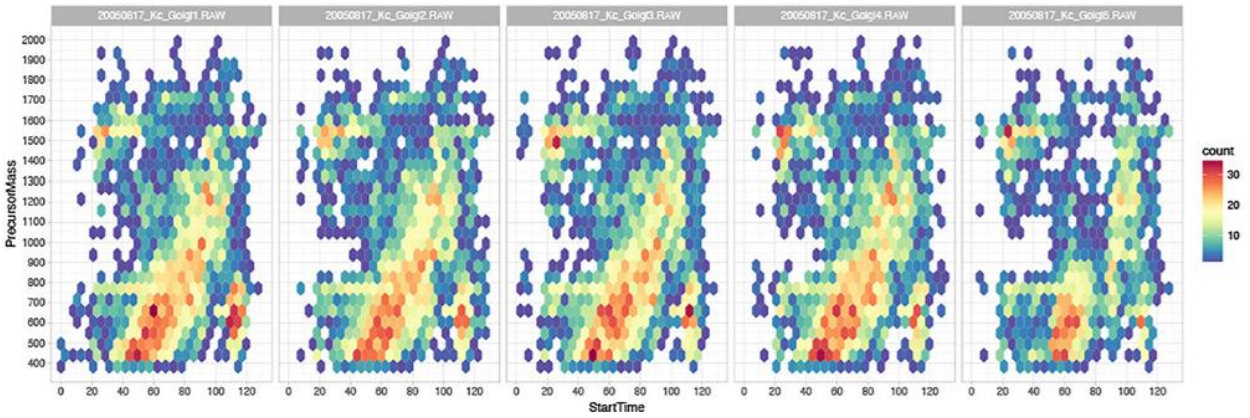


Рисунок 3.10 – Візуальне представлення більшого набору даних у вигляді набору ескізів отриманих за допомогою командного рядка R

3.4 Аналіз результатів наукових розвідок

Розгортання найновіших рішень для візуалізації та дослідження у продуктивному середовищі є складним завданням. Описане в [1] програмно-алгоритмічне рішення, засноване на використанні системної, дозволяє використовувати різні програмні інструменти для процесів візуалізації, обробки та аналізу даних із системою керування даними, що містить масивні сховища даних і метаданих наукових досліджень, зокрема для визначення прикладних експериментальних проектів. Системна платформа аналізує величезну кількість наукових даних протягом тривалого періоду часу. Впродовж останнього періоду часу спостерігається швидка еволюція інформаційних технологій у різних наукових сферах, що, як наслідок, вимагає проєктування та прототипування інтеграційної платформи, готової до будь-якого потенційного перспективного застосування з невеликими модифікаціями.

Здебільшого кожна науково-дослідна установа на даний час вже має свої напрацювання. Що стосується основних об’єктів досліджень, то системи такого класу здебільшого адаптовані до внутрішніх потреб, наприклад, перевірка, облік, замовлення та всі інші необхідні адміністративні процеси. Інша частина науково-дослідних закладів зазвичай покладаються на комерційні лабораторні системи управління інформацією (LIMS). Після того як LIMS буде запущена до використання запрацює, замінити систему такого класу дуже складно, або практично неможливо через величезні фінансові та людські інвестиції. Зазвичай LIMS впроваджуються для конкретного науково-дисциплінарного домену і не є достатньо загальними. З точки зору розробки програмного забезпечення, дуже важливо мати загальну системну достатньо універсальну платформу, що забезпечує інтерфейс для читання та збереження будь-яких типів даних.

3.5 Висновок до третього розділу

В третьому розділі кваліфікаційної роботи описано встановлення та налаштування програмно-алгоритмічних елементів системних платформ керування даними, візуалізації та аналітичного опрацювання. Розглянуто системне використання засобів інтерактивної візуалізації даних. Проаналізовано ініціювання програмно-алгоритмічних засобів візуального дослідження. Проаналізовано результати наукових розвідок.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

4.1 Організація праці при виконанні робіт в обчислювальному центрі

Тема кваліфікаційної роботи освітнього рівня «Магістр» присвячена дослідженню платформ керування даними та інструментів візуалізації для аналітики наукових досліджень. На даний час при проведенні наукових досліджень важко уявити платформи керування даними, інструменти візуалізації та аналітичні інструменти, де б не використовувались ресурси обчислювальних центрів. З того часу, як комп'ютери та обчислювальні центри увійшли в повсякдення людей, проводяться дослідження, як саме вони впливають на здоров'я громадян. Тому актуальним є питання організації праці при виконанні робіт в обчислювальному центрі. При цьому доцільно визначити, які фактори виробничого середовища і трудового процесу впливають на умови праці в обчислювальних центрах та яких заходів безпеки потрібно дотримуватися.

Діяльність більшості працівників сучасних обчислювальних центрів безпосередньо пов'язана з використанням комп'ютерної техніки. Комп'ютер для сучасних громадян є технічною необхідністю, як звичайні побутові пристрої. Адже ми використовуємо їх не задумуючись про шкідливість або нешкідливість, усвідомлюючи лише доступні переваги від їх наявності. Щодо комп'ютерної техніки, то існує обширний обсяг інформації про її безпечність та шкідливість.

Наказом Міністерства соціальної політики України від 14.02.2018 р. № 207 затверджено Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями (в подальшому Вимоги). Терміни у зазначених Вимогах [58] вживаються в таких значеннях:

– екранні пристрої – це електронні засоби для відтворення будь-якої графічної або алфавітно-цифрової інформації, зокрема на основі електронно-променевої трубки, рідкокристалічні, плазмові, проекційні, органічні світлодіодні монітори та інші інноваційні розробки в галузі інформаційних технологій;

– робоче місце, або робоча станція – це сукупність обладнання, що включає екранний пристрій, який може доповнюватися клавіатурою або пристроями введення та програмним забезпеченням, що містить інтерфейси „оператор-дисплей” тощо. Зокрема, периферійні пристрої, електронні носії інформації, смарфони, модеми, друкувальні пристрої, сховища документів, робочі крісла, робочі столи або робочі поверхні „розумних” столів, а також інші необхідні елементи виробничого середовища.

Інші терміни у цих Вимогах вживаються у значеннях, наведених у Законі України «Про охорону праці».

При створенні обчислювальних центрів роботодавці повинні поінформувати працівників під розписку про умови праці та наявність на їх робочих місцях небезпечних та шкідливих виробничих факторів. Зокрема фізичних, хімічних, біологічних, психофізіологічних, які виникають під час роботи з екранними пристроями та які ще не усунуто. Також роботодавці повинні повідомити про можливі наслідки їх впливу на здоров’я працівників відповідно до вимог статті 5 Закону України «Про охорону праці».

Роботодавці повинні для працівників обчислювальних центрів забезпечити навчання і перевірку знань з питань охорони праці та безпечного використання екранних пристроїв до початку роботи з ними, а також у випадках модифікації та організації роботи обладнання. Роботодавці повинні вжити відповідних заходів, щоб забезпечити відповідність робочих місць працівників обчислювальних центрів до цих Вимог.

Під час облаштування робочого місця працівників обчислювальних центрів з екранними пристроями необхідно обирати таке обладнання, яке не

створює зайвого шуму та не виділяє надлишкового тепла [58]. Рівні шуму на робочих місцях осіб, які працюють з екранними пристроями, мають відповідати вимогам Санітарних норм виробничого шуму, ультразвуку та інфразвуку ДСН 3.3.6.037-99, затверджених постановою Головного державного санітарного лікаря України від 01 грудня 1999 року № 37.

Роботодавці повинні за рахунок тривалості робочої зміни організувати внутрішні регламентовані перерви для відпочинку працівників обчислювальних центрів відповідно до Державних санітарних правил і норм роботи з візуальними дисплейними терміналами електронно-обчислювальних машин ДСанПІН 3.3.2.007-98, затверджених постановою Головного державного санітарного лікаря України від 10 грудня 1998 року № 7 (далі - ДСанПІН 3.3.2.007-98).

Роботодавці повинні забезпечити за свій рахунок проведення медичних оглядів працівників обчислювальних центрів відповідно до вимог Порядку проведення медичних оглядів працівників певних категорій, затвердженого наказом Міністерства охорони здоров'я України від 21 травня 2007 року № 246, зареєстрованого в Міністерстві юстиції України 23 липня 2007 року № 846/14113. За результатами цих оглядів роботодавці за потреби повинні забезпечити виконання відповідних оздоровчих заходів.

Роботодавці зобов’язані за необхідності проводити лабораторні дослідження умов праці в обчислювальних центрах з метою виявлення шкідливих і небезпечних факторів виробничого середовища, важкості та напруженості трудових процесів. Зокрема, щодо виявлення ризиків, пов’язаних із погіршенням зору, порушенням фізичного стану, стресом та вживати заходів щодо усунення виявлених ризиків відповідно до статті 13 Закону України «Про охорону праці».

Робочі місця працівників обчислювальних центрів з екранними пристроями мають бути спроектовані так та мати такі розміри, щоб працівники мали простір для безперешкодної зміни робочого положення та

рухів [59]. Для забезпечення безпеки та захисту здоров’я працівників обчислювальних центрів усе випромінювання від екранних пристроїв має бути зведене до гранично допустимого рівня. При цьому вплив на людину факторів довкілля – шуму, вібрації, забруднювачів, температури тощо, який не спричиняє соматичних або психічних розладів, а також змін стану здоров’я, працездатності, поведінки, що виходять за межі пристосувальних реакцій з погляду безпеки та охорони здоров’я працівників.

Організація робочих місць працівників обчислювальних центрів з екранними пристроями має забезпечувати відповідність усіх елементів робочого місця та їх розташування ергономічним, антропологічним, психофізіологічним вимогам, а також характеру виконуваних робіт.

Освітлення робочих місць працівників обчислювальних центрів з екранними пристроями має створювати відповідний контраст між екраном і навколишнім середовищем з урахуванням складності та видів виконуваних робіт і відповідати вимогам ДСанПІН 3.3.2.007-98.

Мікроклімат приміщень обчислювальних центрів з робочими місцями працівників з екранними пристроями має підтримуватись на постійному рівні та відповідати вимогам Санітарних норм мікроклімату виробничих приміщень ДСН 3.3.6.042-99, затверджених постановою Головного державного санітарного лікаря України від 01 грудня 1999 року № 42 (далі – ДСН 3.3.6.042-99).

Робочі столи або робочі поверхні повинні бути достатнього розміру та мати поверхню з низькою відбивною здатністю, допускати гнучкість під час розміщення екрана, клавіатури, документів та відповідного інформаційно-технологічного обладнання. Робочі крісла мають бути стійкими і дозволяти працівникам з обчислювальних центрів з екранними пристроями легко рухатися та займати зручне положення. Сидіння мають регулюватися по висоті, спинка сидіння – як по висоті, так і по нахилу. Слід передбачати підніжку для тих працівників обчислювальних центрів, кому це необхідно.

4.2 Здоровий спосіб життя людини та його вплив на професійну діяльність

Кваліфікаційна робота освітнього рівня «Магістр» присвячена дослідженню платформ керування даними та інструментів візуалізації для аналітики наукових досліджень. Дослідницька діяльність – це пошук нових знань або систематичне розслідування з метою встановлення фактів. Вона пов’язана з обширним переліком видів людської діяльності. Тому доцільно розглянути Здоровий спосіб життя людини та його вплив на професійну діяльність. Адже саме здоровий спосіб життя суттєво впливає на самопочуття дослідників.

Здоров’я людини грунтується на основі сукупності генетичних факторів, способу життя та екологічних умов. Однак певною мірою воно залежить також від свідомого ставлення людини до себе та оточуючого середовища. Здоров’я людини – стан повного соціально-біологічного комфорту коли функція всіх органів і систем людського організму врівноважені з природним і соціальним середовищем, відсутні будь-які хвилювання, хворобливі стани та фізичні дефекти. Критерій здоров’я визначається комплексом показників [60]. Однак за найзагальнішими рисами здоров’я індивідуума можна визначити як природний стан організму, щоактеризується повною зрівноваженістю будь-яких виражених хворобливих змін. Слід пам’ятати, що здоров’я залежить від багатьох факторів які об’єднуються в одне інтегральне поняття – здоровий спосіб життя. Його метою є навчити людину розумно ставитися до свого здоров’я, фізичної та психічної культури, загартовувати свій організм, вміло організовувати працю і відпочинок.

До основних складових здорового способу життя належать спосіб життя, рівень культури, здоров’я в ієрархії потреб, мотивування, зворотні

зв’язки, установа на довге здорове життя, навчання здоровому способу життя та психічний стан.

Спосіб життя. Має велике значення для здоров’я людини і складається з чотирьох категорій:

– економічної – рівень життя;
– соціологічної – якість життя;
– соціально-психологічної – стиль життя;
– соціально-економічної – устрій життя.

Рівень культури. Слід пам’ятати, що людина – суб’єкт і одночасно – головний результат своєї діяльності. Культура з цієї точки зору – це самосвідоме ставлення до самого себе. Однак люди дуже часто нехтують своїм здоров’ям, ведуть неправильний спосіб життя, не дотримуються режиму переїдають, курять. Тому для здоров’я потрібні знання, які увійшли б у повсякденну звичку людини [60].

Здоров’я в ієрархії потреб. Не завжди в житті людини здоров’я займає перше місце порівняно з речами та іншими матеріальними благами. У результаті це призводить до шкоди не лише своєму здоров’ю, а й здоров’ю майбутніх поколінь. Тому, здоров’я повинно займати перше місце в ієрархії потреб людини.

Мотивування. На превеликий жаль, ціну здоров’я більшість людей усвідомлює лише тоді, коли воно значно втрачено. Тільки тоді виникає прагнення вилікувати захворювання, стати здоровим [61].

Зворотні зв’язки – нерозумне і довге випробовування стійкості свого організму нездоровим способом життя (алкоголь, нікотин). Тільки через певний час спрацьовують зворотні зв’язки людини, коли вона кидає шкідливі звички, проте це вже часто запізно.

Установа на довге здорове життя. У повсякденному житті доцільно вміло мобілізувати резерви свого організму на подолання негараздів

життєвого характеру, на зменшення ризику захворювань, що сприяє довголіттю.

Навчання здоровому способу життя. Джерелом навичок з цього питання є передусім приклад батьків, допомагає також і санітарна освіта Важливим фактором, що визначає реакцію людини на екстремальну ситуацію, є її психофізичні якості та загальний стан. Вони проявляються через чутливість людини до виявлення сигналів небезпеки, перед реакцією на ці сигнали. Показники, які зумовлюють можливості людини виявити небезпечну ситуацію та адекватно вщреагувати на неї, залежать від і, індивідуальних особливостей, зокрема від її нервової системи. На поведінку людини у небезпечній ситуації впливає й її психічний та фізичний стан.

Психічний стан. Сучасна людина зустрічається з багатьма факторами ризику, що негативно впливають на стан й нервової та серцево-судинної систем, знижує опірність організму [61]. При цьому виникає стресова реакція організму. Так, наприклад, психічна травма, отримана внаслідок конфлікту, виводить людину з нормального психічного стану, що може призвести до суттєвих змін у виконанні професійних функцій і загального функціонального стану. У перекладі «стрес» означає «напруження», тобто відповідь організму на поставлену перед ним проблему. Адже стрес – це сукупність загальних неспецифічних біохімічних, фізіологічних і психологічних реакцій організму внаслідок дії надзвичайних подразників різної природи і характеру, які викликають порушення функцій органів. Повне звільнення від стресу означає смерть, тому слабкий стрес є нормальним явищем у житті і потрібним для реалізації людської повноцінності. Однак якщо він інтенсивний і довготривалий, то може стати основою розвитку захворювань або зумовити смерть.

Медичні та соціологічні дослідження серед різних категорій населення показують, що люди по-різному реагують на надзвичайні ситуації. Є люди, стресостійкі до побутових негараздів, але дуже стресореактивні до сімейних

проблем та невдач у коханні, інші боляче сприймають невдачі на роботі, ще інші – втрату соціального статусу.

Відомо, що в осіб до тридцяти років життєві потреби значно більші, ніж у людей старшого віку, а відтак стресові стани у них переважають.

Велике значення для розвитку стресового стану має поведінка в екстремальних умовах, наприклад аварія, кримінальна ситуація, стихійне лихо. Неправильна поведінка у таких ситуаціях найчастіше є причиною шкідливих наслідків стресу. Вона зумовлює результат стресу більше, ніж фактори зовнішнього середовища. У цих випадках стрес може виявитись у вигляді паніки, сує́ти, істерики.

Стійкість організму до різноманітних стресових станів є дуже індивідуальною. Деякі люди без усіляких наслідків переносять надзвичайно складні екстремальні ситуації, ніколи не непритомніють, не втрачають сили волі, психологічної рівноваги. Інші вже при незначних екстрімальних ситуаціях втрачають витримку і віру в себе.

4.3 Висновок до четвертого розділу

В четвертому розділі кваліфікаційної роботи освітнього рівня «Магістр» проаналізовано питання організації праці при виконанні робіт в обчислювальному центрі. Описано здоровий спосіб життя людини та його вплив на професійну діяльність.

ВИСНОВКИ

В першому розділі кваліфікаційної роботи освітнього рівня «Магістр»:

– Описано прогресивні методи збирання даних в процесах наукових досліджень.

– Виконано пошук та аналіз наукових публікацій щодо платформ даних в наукових дослідженнях.

– Проаналізовано кількісні показники наукових публікацій щодо платформ даних в наукових дослідженнях.

– Розглянуто критерії оцінювання інформаційно технологічних платформ для зберігання даних в наукових дослідженнях.

В другому розділі кваліфікаційної роботи:

– Описано типи інформаційних колекцій та наборів даних, що використовуються в сучасних наукових дослідженнях.

– Досліджено поширені методи аналітичного опрацювання колекцій та наборів великих даних у наукових дослідженнях.

– Розглянуто системну платформу для інтеграції даних, засобів візуалізації та аналітичного опрацювання.

В третьому розділі кваліфікаційної роботи:

– Описано встановлення та налаштування програмно-алгоритмічних елементів системних платформ керування даними, візуалізації та аналітичного опрацювання.

– Висвітлено системне використання засобів інтерактивної візуалізації даних.

– Розглянуто ініціювання програмно-алгоритмічних засобів візуального дослідження.

– Проаналізовано результати наукових розвідок.

У розділі «Охорона праці та безпека в надзвичайних ситуаціях» проаналізовано організацію праці при виконанні робіт в обчислювальному центрі. Описано здоровий спосіб життя людини та його вплив на професійну діяльність.

ПЕРЕЛІК ДЖЕРЕЛ

1 Barkow-Oesterreicher S, Türker C, Panse C. FCC − an automated rule-based processing tool for life science data. Source Code Biol Med 2013;8:3.

2 Chiva C, Maia TM, Panse C, Stejskal K, Douché T, Matondo M, et al. Quality standards in proteomics research facilities.EMBO Rep 2021;22. https://doi.org/10.15252/embr.202152626.

3 Panse, Christian, Christian Trachsel, and Can Türker. "Bridging data management platforms and visualization tools to enable ad-hoc and smart analytics in life sciences." Journal of Integrative Bioinformatics (2022).

4 Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, et al. A high-quality catalog of the Drosophilamelanogaster proteome. Nat Biotechnol 2007;25:576−83.

5 Keim DA. Information visualization and visual data mining. IEEE Trans Visual Comput Graph 2002;8:1−8.

6 van Wijk JJ. The value of visualization. In: VIS 05. Minneapolis, Minnesota, USA: IEEE Visualization; 2005:79−86 pp.

7 United Nations Statistical Commission (2014) Big data and modernization of statistical systems. Report of the Secretary-General E/CN. 3.2014/11 of the forty-ffth session of UNSC 4–7 March 2014. United Nations, New York.

8 Di Bella E, Leporatti L, Maggino F (2018) Big data and social indicators: actual trends and new perspectives. Soc Indic Res 135:869– 878. https://doi.org/10.1007/s11205-016-1495-y.

9 Allen, Cameron, et al. "A review of scientific advancements in datasets derived from big data for monitoring the Sustainable Development Goals." Sustainability Science 16.5 (2021): 1701-1716.

10 Kong, Lingqiang, Zhifeng Liu, and Jianguo Wu. "A systematic review of big data-based urban sustainability research: State-of-the-science and future directions." Journal of Cleaner Production 273 (2020): 123142.

11 UNECOSOC (2013) Fundamental Principles of Ofcial Statistics. Resolution adopted by the United Nations Economic and Social Council on 24 July 2013. United Nations, New York.

12 Tam S-M, Van Halderen G (2020) The fve V's, seven virtues and ten rules of big data engagement for ofcial statistics. Stat J IAOS 36:423–433. https://doi.org/10.3233/SJI-190595.

13 Florescu D, Karlberg M, Reis F, Del castillo PR, Skaliotis M & Wirthmann A (2014) Will 'big data' transform ofcial statistics. European Conference on the Quality of Ofcial Statistics 2014. Vienna, Austria.

14 Van Den Homberg M, Susha I (2018) Characterizing data ecosystems to support ofcial statistics with open mapping data for reporting on sustainable development goals. ISPRS Int J Geo-Inf 7:456. https://doi.org/10.3390/ijgi7120456.

15 Liu, Xingjian, and Ying Long. "Automated identification and characterization of parcels with OpenStreetMap and points of interest." Environment and Planning B: Planning and Design 43.2 (2016): 341-360.

16 He, Qingsong, et al. "The impact of urban growth patterns on urban vitality in newly built-up areas based on an association rules analysis using geographical 'big data'." Land Use Policy 78 (2018): 726-738.

17 Xing, Hanfa, Yuan Meng, and Yan Shi. "A dynamic human activity-driven model for mixed land use evaluation using social media data." Transactions in GIS 22.5 (2018): 1130-1151.

18 Song, Yimeng, et al. "Dynamic assessments of population exposure to urban greenspace using multi-source big data." Science of the Total Environment 634 (2018): 1315-1325.

19 Goel, Rahul, et al. "Estimating city-level travel patterns using street imagery: A case study of using Google Street View in Britain." PloS one 13.5 (2018): e0196521.

20 Kuo, Ching-Yen, et al. "Application of a time-stratified case-crossover design to explore the effects of air pollution and season on childhood asthma hospitalization in cities of differing urban patterns: Big data analytics of government open data." International Journal of Environmental Research and Public Health 15.4 (2018): 647.

21 Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C., 2011. A tale of one city: using cellular network data for urban planning. Ieee Pervasive Comput. 10 (4), 18e26.

22 Ilieva, R.T., McPhearson, T., 2018. Social-media data for urban sustainability. Nat. Sustain. 1 (10), 553e565.

23 Zhu, Xi, and Diansheng Guo. "Urban event detection with big data of taxi OD trips: A time series decomposition approach." Transactions in GIS 21.3 (2017): 560-574.

24 Tao, Sui, et al. "Exploring Bus Rapid Transit passenger travel behaviour using big data." Applied geography 53 (2014): 90-104.

25 Roof, K., Oleru, N., 2008. Public health: seattle and King County’s push for the built environment. J. Environ. Health 71 (1), 24e27.

26 Kuang, Bing, et al. "How urbanization influence urban land consumption intensity: Evidence from China." Habitat International 100 (2020): 102103.

27 Hassani, Hossein, et al. "A review of data mining applications in crime." Statistical Analysis and Data Mining: The ASA Data Science Journal 9.3 (2016): 139-154.

28 Hassani, Hossein, Xu Huang, and Emmanuel Silva. "Big data and climate change." Big Data and Cognitive Computing 3.1 (2019): 12.

29 Xu, Dongkuan, and Yingjie Tian. "A comprehensive survey of clustering algorithms." Annals of Data Science 2.2 (2015): 165-193.

30 Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases." Proceedings of the 1993 ACM SIGMOD international conference on Management of data. 1993.

31 Chifor, Bogdan-Cosmin, Ion Bica, and Victor-Valeriu Patriciu. "Sensing service architecture for smart cities using social network platforms." Soft Computing 21.16 (2017): 4513-4522.

32 Poorthuis, Ate. "How to draw a neighborhood? The potential of big data, regionalization, and community detection for understanding the heterogeneous nature of urban neighborhoods." Geographical Analysis 50.2 (2018): 182-203.

33 Ahlberg C. Spotfire: an information exploration environment. SIGMOD Rec 1996;25:25−9.

34 Marzolf B, Deutsch EW, Moss P, Campbell D, Johnson MH, Galitski T. SBEAMS-Microarray: database software supportinggenomic expression analyses for systems biology. BMC Bioinf 2006;7:286−91.

35 Poullet P, Carpentier S, Barillot E. myProMS, a web server for management and validation of mass spectrometry-basedproteomic data. Proteomics 2007;7:2553−6.

36 Paulhe N, Canlet C, Damont A, Peyriga L, Durand S, Deborde C, et al. PeakForest: a multi-platform digital infrastructure forinteroperable metabolite spectral data and metadata management. Metabolomics 2022;18:40.

37 Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: the konstanz information miner. In: Studies inClassification, Data Analysis, and Knowledge Organization (GfKL 2007). Springer; 2007.

38 Analyze anything with teradata analytics platform; 2018. Available from: https://www.teradata.com.

39 Qu K, Garamszegi S, Wu F, Thorvaldsdottir H, Liefeld T, Ocana M, et al. Integrative genomic analysis by interoperation ofbioinformatics tools in GenomeSpace. Nat Methods 2016;13:245−7.

40 CLC Genomics workbench; 2018. Available from: https://www.qiagenbioinformatics.com.

41 Sharma V, Eckels J, Taylor GK, Shulman NJ, Stergachis AB, Shannon AJ, et al. Panorama: a targeted proteomics knowledgebase. J Proteome Res 2014;13:4205−10.

42 The seven bridges platform: biomedical data analysis at scale; 2018. Available from: https://www.sevenbridges.com/ platform/.

43 Türker C, Schmid M, Joho D, Akal F, Gürel U. B-Fabric Project Manual; 2018. Available from: http://bfabric.org.

44 Türker C, Akal F, Joho D, Panse C, Barkow-Oesterreicher S, Rehrauer H, et al. B-Fabric: the Swiss army knife for LifeSciences. In: Proceedings of the 13th international conference on extending database technology. EDBT ’10 Lausanne, Switzerland. New York, NY, USA: ACM; 2010:717−20 pp.

45 Aleksiev T, Barkow-Oesterreicher S, Kunszt P, Maffioletti S, Murri R, Panse C. VM-MAD: a cloud/cluster software forservice-oriented academic environments. In: Kunkel JM, Ludwig T, Meuer HW, editors Supercomputing. Berlin, Heidelberg: Springer; 2013:447−61 pp.

46 Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web application framework for R; 2016. R package version 0.13.2. Available from: https://CRAN.R-project.org/package=shiny.

47 Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databasesusing mass spectrometry data. Electrophoresis 1999;20:3551−67.

48 Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics2012;13:22−4.

49 HUPO Proteomics Standards Initiative; 2017. Available from: http://www.psidev.info/.

50 Lang DT, the CRAN Team. XML: tools for parsing and generating XML within R and S-plus; 2017. R package version 3.98-1.7. Available from: https://CRAN.R-project.org/package=XML.

51 Trachsel C, Panse C. bfabricShiny: a shiny module for bridging B-fabric and R using REST; 2022. R package version 0.11.9. Available from https://github.com/fgcz/bfabricShiny.

52 Nanni P, Panse C, Gehrig P, Mueller S, Grossmann J, Schlapbach R. PTM MarkerFinder, a software tool to detect andvalidate spectra from peptides carrying post-translational modifications. Proteomics 2013;13:2251−5.

53 Trachsel C, Panse C, Kockmann T, Wolski WE, Grossmann J, Schlapbach R. rawDiag - an R package supporting rationalLC-MS method optimization for bottom-up proteomics. J Proteome Res 2018;17:2908−14.

54 Wilkinson L. The Grammar of graphics (statistics and computing). Secaucus, NJ, USA: Springer-Verlag, Inc.; 2005.

55 Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2009. Available from: http://ggplot2.org.

56 R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria; 2008. Availablefrom: http://www.R-project.org.

57 Yoghourdjian V, Dwyer T, Klein K, Marriott K, Wybrow M. Graph thumbnails: identifying and comparing multiple graphs ata glance. IEEE Trans Visual Comput Graph 2018;1:3081−95.

58 МІНІСТЕРСТВО СОЦІАЛЬНОЇ ПОЛІТИКИ УКРАЇНИ. Наказ 14.02.2018 № 207 Про затвердження Вимог щодо безпеки та захисту здоров’я працівників під час роботи з екранними пристроями. Доступно онлайн: https://zakon.rada.gov.ua/laws/show/z0508-18#Text.

59 Умови праці працівників, які використовують у роботі персональні комп’ютери. Доступно онлайн: https://zolochiv.net/umovy-pratsi-pratsivnykiv-iaki-vykorystovuiut-u-roboti-personal-ni-komp-iutery/.

60 Здоровий спосіб життя. Доступно онлайн: https://buklib.net/books/27545/.

61 Лекція 2. Теоретичні засади формування здорового способу життя. Доступно онлайн: https://msn.khnu.km.ua/pluginfile.php/281951/mod_resource/content/0/%D0%A2%D0%95%D0%9C%D0%90_02.htm.

ДОДАТКИ

Додаток А

Тези конференції

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ

М А Т Е Р І А Л И

X НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

7–8 грудня 2022 року

ТЕРНОПІЛЬ
2022

УДК 001
М34

ПРОГРАМНИЙ КОМІТЕТ

Голова: Сергій Лупенко– докт. техн. наук, професор.

Співголови: Павло Марущак– докт. техн. наук, професор, проректор з наукової роботи.
Ігор Баран– канд. техн. наук, доцент, декан факультету ФІС.

Науковий секретар: Галина Семенишин– старший викладач.

Члени: докт. фіз.-мат. наук, професор Василь Кривень; докт. техн. наук, професор Ярослав Литвиненко; докт. техн. наук, професор Микола Карпінський; докт. фіз.-мат. наук, професор Михайло Петрик; канд. техн. наук, доцент Галина Осухівська; канд. пед. наук, доцент Жанна Баб'як; канд. техн. наук, доцент Наталія Загородна.

ОРГАНІЗАЦІЙНИЙ КОМІТЕТ

Голова: Юрій Скоренький– канд. фіз.-мат. наук, доцент, завідувач кафедри фізики

Члени: канд. техн. наук, доцент Вячеслав Никитюк; канд. техн. наук, доцент Дмитро Михалик; канд. техн. наук, асистент Марія Стадник; асистент Наталія Шаблій; ст. викладач Ліліана Джиджора.

М34 Матеріали X науково-технічної конфції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя, (Тернопіль, 7–8 грудня 2022 р.). – Тернопіль : Тернопільський національний технічний університет імені Івана Пулюя, 2022. –162 с.

Адреса оргкомітету: ТНТУ ім. І. Пулюя, м. Тернопіль, вул. Руська, 56, 46001,
тел. (0352) 52-41-33, факс (0352) 254983.
E-mail: conffis2022@gmail.com

Редагування, оформлення та верстка: Галина Семенишин

СЕКЦІЇ КОНФЕРЕНЦІЇ, ЯКІ ПРЕДСТВЛЕНІ В ЗБІРНИКУ

- Математичне моделювання;
- Інформаційні системи та технології;
- Комп'ютерні системи та мережі;
- Програмна інженерія та моделювання складних розподілених систем;
- Новітні фізико-технічні та освітні технології.

В збірнику надруковано тези доповідейIX науково-технічної конференції «Інформаційні моделі, системи та технології» (Тернопіль, 7–8 грудня 2022 р.) за такими науковими напрямками: математичне моделювання; інформаційні системи та технології; комп'ютерні системи та мережі; програмна інженерія та моделювання складних розподілених систем; новітні фізико-технічні та освітні технології.

Розрахований на науковців, викладачів та студентів вузів.

За зміст тез та дотримання норм академічної доброчесності відповідальність несе автор.

ЗМІСТ

УДК 004.9

К. Бєлоусов, Т. Маєвський
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)
(Технічний коледж Тернопільського національного технічного університету імені Івана Пулюя)

РОЛЬ ТА ЗНАЧЕННЯ ВЕЛИКИХ ДАНИХ В СУЧАНИХ НАУКОВИХ ДОСЛІДЖЕННЯХ

UDC 004.9

K. Bielousov, T. Maievskyi

THE BIG DATA ROLE AND SIGNIFICANCE IN MODERN SCIENTIFIC RESEARCH

Інноваційний інформаційно-технологічний концепт «Великі дані» трансформує сучасні наукові дослідження. Вони об'єднують великі за обсягом масиви та колекції даних зібрані під час лабораторних експериментів, наукових досліджень, опрацювання різнотипових персональних та медичних записів, в тому числі накопичені з використаним Інтернету речей (IoT-пристроїв). Зокрема, в медичній галузі, експоненційно зростає швидкість отримання інформації геноміки та епігеноміки, транскриптоміки та протеоміки, метаболоміки та фармакогеноміки [1]. Це дає перспективні можливості проведення досліджень та, як наслідок, вагомих досягнень для персоналізованої медицини, прокладаючи при цьому шлях до покращення якості життя та надання медичних послуг.

«Великі дані» (англ. Big Data) характерзуються величезними обсягами забраних, накопичених та опрацьованих даних, створених у приватному та державному секторах людської діяльності. Вони спрямовані на заохочення використання інноваційних інформаційних технологій для аналізу великої кількості доступних наборів даних, видобування та отримання нових знань та інформації. «Великі дані» можуть набувати будь-яку форму складних даних великого обсягу.

Колекції «великих даних», при формуванні обчислювальних систем з метою їх аналітичного опрацювання, зазвичай характеризуються п'ятьма характеристиками: обсяг розмірів загального набору даних, швидкість з якою обчислювальні конвеєри можуть завантажувати та опрацьовувати набори даних, достовірність наборів даних у відображенні справжньої природи стану систем, різноманітність елементів наборів даних один відносно одного в колекції, мінливість – якісна характеристика даних, що представляє постійні зміни в даних та спричиняє відповідний вплив на продуктивність обчислювальної системи [2]. В комплексі всі зазначені якості сприяють системній цінності «великих даних» при застосуванні в будь-якій галузі наукових досліджень. Їх необхідно ретельно враховувати при формуванні процесів роботи з «великими даними».

Тому доцільно критично розглянути та проаналізувати основні використані аналітичні обчислювальні методи, алгоритми та отримані з їх допомогою результати, які сприяли нещодавнім науковим досягненням, отрманим з використанням великих даних. Водночас варто проаналізувати тенденції та перспективні напрямки проведення наукових досліджень в галузі великих за обсягом даних.

Література

1. Khan, Ibrahim Haleem, and Mohd Javaid. «Big data applications in medical field: A literature review. «Journal of Industrial Integration and Management» 6.01 (2021): 53–69.
2. Cremin, Conor John, Sabyasachi Dash, and Xiaofeng Huang. «Big data: Historic advances and emerging trends in biomedical research.» Current Research in Biotechnology (2022).

УДК 004.9

К. Бєлоусов, Т. Маєвський
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)
(Технічний коледж Тернопільського національного технічного університету імені Івана Пулюя, Україна)

ВЕЛИКІ ДАНІ ТА АНАЛІТИЧНЕ ОПРАЦЮВАННЯ В НАУКОВИХ ДОСЛІДЖЕННЯХ

UDC 004.9

K. Bielousov, T. Maievskyi

BIG DATA AND ANALYTICAL PROCESSING IN SCIENTIFIC RESEARCH

На даний час спостерігається безпрецедентне зростання обсягів колекцій даних накопиченихз в результатів наукових досліджень. Установи та організації використовують різнотипові аналітичні інструменти, сформовані на основі методів штучного інтелекту (англ. Artificial intelligence, AI) та методів машинного навчання (англ. Machine Learning, ML) для отримання знань та інформації на основі даних, щоб зменшити витрати, збільшити потоки доходів, розробити персоналізовані сервіси та послуги. Невпинно зростає перелік, які використовують «великі дані» (англ. Big Data, BD) для вдосконалення та оптимізації своїх технологічних та бізнес-процесів, виявлення ринкових споживчих тенденцій.

За оцінками фахівців прогнозується, що капіталізація світового ринку «великих даних» до 2025 року перевищить 70 мільярдів доларів США. Водночас очікується, що Сполучені Штати домінуватимуть в цій галузі, сприяючи відбору та аналітичному опрацюванню понад дев'яносто відсотків колективних «великих даних» Північної Америки до 2025 року. Використання «великих даних» в наукових дослідженнях швидко зростає. На даний час обсяги дослідницьких даних, що генеруються за один день, за оцінками фахівців, співмірна з обсягами, які раніше генерувалися за десятиліття. Нещодавні звіти приватних компаній в галузі аналітичного опрацювання даних, в тому числі Всесвітнього економічного форуму, свідчать, що наразі людство оперує сорока чотирма зетабайтами даних. В перспективі ці обсяги зростатимуть до 463 екзабайтів накопичуваних даних щодня в усьому світі [1]. Великі технологічні гіганти, зокрема Google, Facebook, Amazon і Microsoft, зберігають приблизно 1200 петабайт даних завдяки доступу до цифрових технологій по всьому світу. У сукупності статистичні дані свідчать, що розвиток цифровізації, доступність Інтернету та поширення Інтернету речей (IoT) сприяютимуть зростаючому перевантаженню даними.

Складна природа наборів сучасних досліджницьких наборів та колекцій даних формує складну задачу видобування та отримання значущих висновків та знань з «великих даних». Накопичення наборів даних з більш різноманітними колекціями «великих даних» дозволяє ефективніше ідентифікувати загальні закономірності, притаманні загальній масі наборів даних, за рахунок незначних характеристик, які можуть висвітлити важливі області [2]. Тому аналітичне опрацювання «великих даних» повинно використовувати надійні методи для ідентифікації надійних ознак.

Література

1. How Much Data Is Created Every Day? [27 Staggering Stats], 2022. URL: https://earthweb.com/how-much-data-is-created-every-day/?gclid=Cj0KCQiAkMGcBhCSAR IsAIW6d0B0cDctoIEdGRbWuT3s7MdUY-ppN87LCz9C9Z_RKrGKvUejqzSVDWEaArPS EALw_wcB.
2. Boehm, Kevin M., et al. «Harnessing multimodal data integration to advance precision oncology.» Nature Reviews Cancer 22.2 (2022): 114–126.