УДК 004.93

ПІДХОДИ ДО ПАРСИНГУ НЕСТАНДАРТНО ОРГАНІЗОВАНИХ ГЕОМЕТРИЧНИХ ДАНИХ

DOI: 10.33842/2313-125X-2026-29-210-219
Морозова М.Ю.,
Maryna.Morozova@infiz.khpi.edu.ua, ORCID: 0009-0004-2795-9315
Сидоренко О.С., канд. техн. наук,
Olena.Sydorenko@khpi.edu.ua, ORCID: 0000-0002-5506-498X
Національний технічний університет «Харківський політехнічний інститут» (м. Харків, Україна)

Статтю присвячено дослідженню проблеми структурного розриву між нестандартно організованими наборами геометричних даних та вимогами програмних інструментів статистичного аналізу. Здійснено огляд наукових джерел, дотичних темі роботи або суміжним напрямам. Описана у роботі методологія і підходи реалізовані на прикладі набору даних Canonical Polyhedra, що містить метричні та топологічні характеристики про 2907 багатогранників для середовища Wolfram Mathematica. Основною особливістю досліджуваного набору є його специфічна структура, а саме представлення просторових об'єктів не у вигляді традиційних таблиць, а у формі абстрактних синтаксичних дерев. У статті детально проаналізовано архітектуру набору: визначено основні вузли даних, а також виділено символьні математичні записи, що потенційно можуть викликати труднощі в парсингу. У наборі даних деякі характеристики геометричних фігур представлені не числовим, а символьним записом. Це забезпечує абсолютну математичну точність даних, проте робить їх непридатними для автоматизованих обчислень без попередньої трансформації. Таким чином, додатково обґрунтовано неспроможність аналізу набору за допомогою базових програмних бібліотек для роботи з даними. Основу методології дослідження складає практична розробка рекурсивного алгоритму парсингу на прикладі конкретного набору даних. Опис алгоритму подано в універсальному загальному вигляді, що робить його доступним для подальшого застосування серед ширшого кола питань. У роботі описано логіку обходження вузлів синтаксичного дерева, ідентифікацію їхніх заголовків і нормалізацію типів даних. Результатом такого підходу є перетворення абстрактного синтаксичного дерева у нормалізовану таблицю, придатну для подальшого статистичного аналізу. Практична значущість дослідження полягає у виділенні універсальних підходів до парсингу складних вкладених ієрархій даних, що часто є джерелом цінної інформації про просторові геометричні фігури для їхнього подальшого вивчення і

застосування у галузях обчислювальної геометрії, машинного навчання тощо.

Ключові слова: набір даних, статистичний аналіз, парсинг, система комп'ютерної математики, багатогранник.

Постановка проблеми. Розуміння структури геометричних даних становить важливий етап у процесі їхнього статистичного аналізу. Проте на етапі підготовки можуть виникати складнощі, пов'язані з організацією інформації. Видавцями наборів геометричних даних здебільшого виступають сучасні математичні комп'ютерні системи, що часто експортують інформацію у вигляді складних ієрархій, зокрема абстрактних синтаксичних дерев (AST), замість двовимірних табличних масивів, які є більш зручними для процесу аналізу.

У таких наборах геометричні фігури представлені через вкладені структури з специфічною внутрішньою номенклатурою та наявними в них символьними типами даних. Стандартні базові модулі мов програмування здатні лише формально зчитати файл у вигляді неструктурованих багатовимірних списків. Це унеможливлює застосування методів описової статистики, які вимагають подання об'єктів у двовимірному масиві ознак, що, в свою чергу, обмежує процес подальшого аналізу.

Актуальність дослідження зумовлена необхідністю подолання розриву між формами організації наборів геометричних даних і жорсткими вимогами до статистичних алгоритмів їхнього аналізу. Таким чином, для вирішення описаних проблем, виникає потреба у розробці спеціалізованих алгоритмів парсингу. Такі алгоритми можуть бути рекурсивними і в загальному розумінні полягають в процесі автоматичного обходження складних структур, вилученні надлишкових системних символів, збереженні цілісності даних і нормалізації числових показників для подальшого статистичного аналізу.

Аналіз останніх досліджень і публікацій. Проблема обробки та підготовки математичних даних до подальшого аналізу залишається актуальною протягом багатьох років. В сучасній науковій літературі продовжує панувати думка, що перетворення складних масивів даних у формати, доступні для машинного зчитування, залишається одним з найбільш ресурсовитратних етапів аналітики [1, 2].

У сучасних працях про дані двовимірна структура масивів (де кожна ознака - це стовпець, об'єкт - рядок) називається основною умовою для ефективного статистичного аналізу [3]. Ця теза також є важливою при використанні програмних бібліотек, за допомогою яких можна досліджувати дані (наприклад, бібліотеки Pandas у мові Python) [4].

Природу цифрового подання математичних об'єктів розглядають дослідники у галузях символьних обчислень і машинного навчання. У відповідних працях описується, що AST-форма є ефективним підходом до опису математичних понять [5]. Це, в свою чергу, дозволяє використовувати

її в комп’ютерних математичних системам, які досить часто випускають набори геометричних даних.

Підходи до експорту геометричних даних описуються у багатьох роботах з аналітики даних або машинного навчання. Зокрема, технічні особливості JSON-формату, у якому бувають представлені ієрархічно вкладені дані, здебільшого називаються однією з причин, що ускладнюють процес автоматичного структурування стандартними програмними модулями читання [6].

Процеси створення алгоритмів для подолання вищеописаного структурного бар’єру наводяться у роботах з методології трансформації даних. Автори активно досліджують можливості перетворення даних деревоподібних структур у табличні [7]. Крім того, частина дослідників концентрується на проблемах наявного у таких наборах символьного шуму, зазначаючи необхідність нормалізації типів даних після їхнього розгортання [2].

Проте, попри ґрунтовну теоретичну базу щодо процесів обробки і аналізу математичних даних, у науковій літературі представлено не так багато праць, присвячених рішенням щодо парсингу специфічних експортних діалектів для їхньої подальшої автоматизованої обробки програмними бібліотечними методами.

Формулювання цілей статті. Основною метою статті є розробка підходів до парсингу геометричних структур даних, поданих у нестандартних та вузькоспеціалізованих формах. Практична частина дослідження реалізована на прикладі набору даних Canonical Polyhedra у форматі JSON від видавця систем комп’ютерної математики Wolfram Research, у якому інформація про геометричні фігури (загалом, 2907 об’єктів) записана у складному для автоматизованого зчитування вигляді [8]. У роботі виконано спроби розробки алгоритму парсингу для виконання поставлених цілей.

Основна частина.

Структурна організація та синтаксичні особливості набору даних. Процес розробки алгоритму парсингу набору геометричних даних Canonical Polyhedra розпочато з аналізу архітектури вхідних даних. Набір містить відомості про 2907 просторових фігур (багатогранників) включно з їхніми метричними і топологічними характеристиками. З точки зору обчислювальної геометрії набір може бути важливим джерелом інформації про канонічні (правильні) багатогранники 4-9 граней, зібрані в єдиній колекції.

Набір організовано у вигляді абстрактного синтаксичного дерева системи Wolfram Mathematica. Внутрішня логіка записів суттєво відрізняється від класичної для описової статистики форми. Основними вузлами цієї структури виступають конструкції:

1. Association. Асоціативний вузол, що відповідає за групування властивостей об’єкта набору даних.

2. Rule. Вказівний вузол, що представляє зв’язок «ключ-значення».

3. List. Вузол для списків (наприклад, координат окремої геометричної фігури з набору даних).

Кожен об’єкт представлений як елемент окремої асоціації, де ключем є його ідентифікатор, а значенням — вкладений набір характеристик. Основною особливістю такого набору є те, що стандартні пари «ключ-значення» не записані як звичайні поля JSON-файлу, а обгорнуті у вузли Rule, де перший елемент списку є назвою параметра, а другий — його вмістом. Більш детально логіку побудови архітектури набору можна розглянути на прикладі першого об’єкта, тетраедра (ідентифікатор об’єкта — 4_1). На рис. 1 показано консольну схему виведення AST-структури обраної фігури.

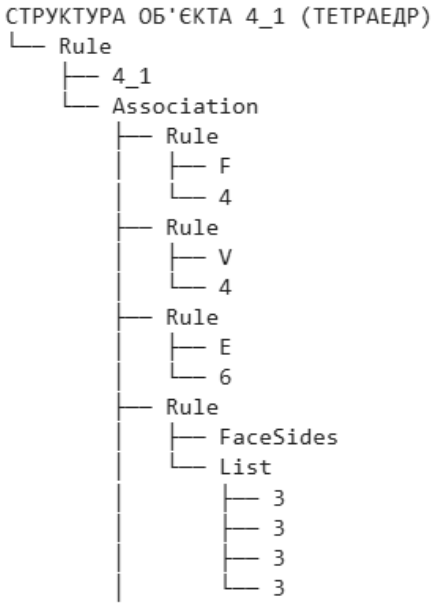

Рис. 1. Схема AST-організації даних про тетраедр
у наборі Canonical Polyhedra

Окрім ієрархічної структури з вкладеннями, складність при автоматизованому зчитуванні набору становлять спеціальні символьні математичні вирази. Замість стандартних числових значень, метричні характеристики об’єктів представлено у вигляді низки допоміжних математичних вузлів. За допомогою них символьно реалізовано обчислення основних операцій: наприклад, вузол Plus відповідає за додавання, Times за множення, Rational за ділення, Power за піднесення до степені, Root за корінь рівняння тощо. Така символьна форма запису забезпечує математичну точність, проте є непридатною для прямого статистичного аналізу, а, отже потребує попередньої обробки, тобто програмного парсингу.

Проблеми зчитування специфічних структур даних стандартними програмними бібліотеками. У сучасних мовах програмування, бібліотеки для роботи з даними (наприклад, Pandas у Python) потребують попередньо

оброблених наборів, організованих за принципом tidy data, де кожному об'єкту відповідає рядок, а кожній ознаці — стовпець [3, 9]. Спроби завантаження наборів з нестандартно організованою структурою даних за допомогою базових методів ведуть до створення фреймів, у яких всі вкладені характеристики залишаються всередині неструктурованих списків.

При базових підходах очікується побачити у файлі об’єкт стандартної структури (наприклад, типовий для аналітики даних JSON-об’єкт вигляду {"key": "value"}) [10, 11]. Замість цього програма стикається із записом вигляду ["Rule", "key", "value"] і не може у повній мірі обробити отриману інформацію або отримати доступ до окремих вузлів системи. Основні вузли, які добре систематизують дані при їхній обробці системами комп’ютерної математики, у цьому випадку є лише елементами списку, а не вказівкою на зв'язок між ознакою та її значенням. Допоміжні символьні вузли, що відповідають за опис математичних операцій, сприймаються як текстові вкладені рядки, що унеможливлює виконання вказаних у них арифметичних дій.

Методологія парсингу даних у нестандартно організованих наборах. Вирішення проблеми структурної невідповідності у сучасній науці стає можливим за допомогою створення спеціальних алгоритмів парсингу [12, 13]. Запропонований підхід базується на принципах рекурсивного обходу абстрактного синтаксичного дерева і подається в загальному вигляді, що робить можливим його адаптацію для інших нестандартно організованих наборів геометричних даних. Алгоритм може бути реалізований протягом наступних етапів:

1. Диференціація та розгортання структурних вузлів. Відбувається зчитування першого елемента кожного вкладеного масиву. Якщо вузол ідентифікується як основний і являє собою контейнер (Association), функція створює порожній словник для збереження у ньому певної інформації. Далі, згідно маркерів зв’язку (Rule), алгоритм рекурсивно занурюється по структурі вглиб, вилучаючи при цьому ім’я ознаки певної геометричної фігури і її вміст.

2. Обхід масивів. На цьому етапі відбувається обробка вузлів списків List, якщо вони наявні у структурі об’єкта. Функція парсингу при цьому застосовується ітеративно до кожного елемента списку. Ігноруючи технічний запис List, функція бере лише наступний після нього корисний вміст (характеристики фігури).

3. Нормалізація даних. Обробка кінцевих вузлів, що містять потрібні числові характеристики, здійснюється за рахунок розробленої функції перетворення типів даних. За допомогою неї виконується розпізнавання символьних математичних записів і подальше приведення їх до вигляду чисел з рухомою комою. Відсутнім значенням, або тим, які неможливо обробити, доцільно присвоювати нульове значення. Програмна реалізація етапів алгоритму рекурсивного парсингу для набору даних Canonical Polyhedra представлена у вигляді блок-схеми на рис. 2.

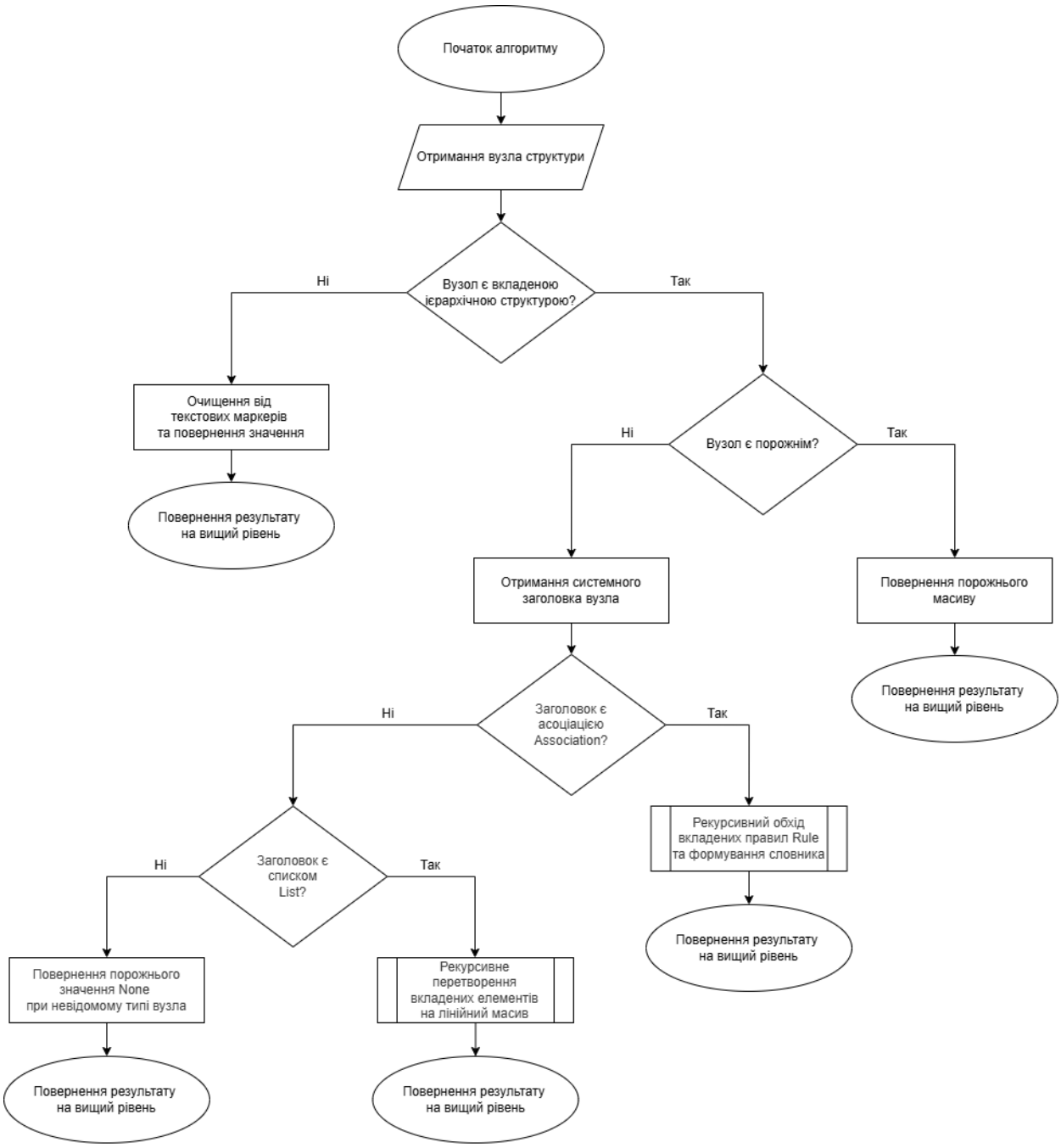

Рис. 2. Блок-схема алгоритму рекурсивного парсингу нестандартно організованих геометричних даних

Завдяки такому підходу, складна структура набору отримує стандартизований вигляд масиву числових ознак. Це, в свою, чергу, дозволяє підготувати специфічно організовані набори геометричних даних для подальшого дослідження за допомогою стандартних програмних бібліотек і методів описової статистики.

Висновки. Основою проведеного дослідження є задача подолання структурного розриву між нестандартним представленнями об’єктів у наборах геометричних даних та вимогами сучасних програмних

інструментів для їхньої обробки. Попри нетипову структуру, такі колекції можуть містити цінну інформацію про наявні у них геометричні фігури. Ключовим результатом статті є опис методології для роботи зі специфічно організованими даними, а також заснована на ньому програмна реалізація алгоритму рекурсивного парсингу на прикладі обраного набору. Перетворення складної ієрархічної структури у таблицю і подальша стандартизація даних відкриває можливість застосування методів описової статистики до усіх наявних у наборі геометричних фігур. Автоматизована обробка символьних записів дозволяє не лише отримати доступ до певних характеристик об'єктів, а і обчислювати на їхній основі потенційно нові, не наявні у наборі. Зокрема, це дає змогу більш детально аналізувати набір, проводити статистичні ідентифікацію і диференціацію окремих груп об'єктів і аналізувати їх на основі наведених метрик і топологій (наприклад, виявляти Платонові, Архімедові, Каталанові та інші види тіл). Таким чином, наведені підходи можуть бути застосованими у галузях обчислювальної геометрії і машинного навчання, де автоматизована підготовка даних є важливим етапом для їхнього подальшого використання.

Література

1. Kandel S., Heer J., Plaisant C., Kennedy J., van Ham F., Riche N. H., Weaver C., Lee B., Brodbeck D., Buono P. Research directions in data wrangling: visualizations and transformations for usable and credible data. *Information Visualization*. 2011. Vol. 10, No. 4. P. 271–288. DOI: https://doi.org/10.1177/1473871611415994.
2. Kotsiantis S. B., Kanellopoulos D., Pintelas P. E. Data preprocessing for supervised learning. *International Journal of Computer Science*. 2006. Vol. 1, No. 1. P. 111–117.
3. Wickham H. Tidy data. Journal of Statistical Software. 2014. Vol. 59, No. 10. P. 1–23. DOI: https://doi.org/10.18637/jss.v059.i10.
4. McKinney W. Data structures for statistical computing in Python. Proceedings of the 9th Python in Science Conference, Austin, TX, USA, June 28 – July 3, 2010. P. 56–61. DOI: https://doi.org/10.25080/Majora-92bf1922-00a.
5. Hosseinpour S., Milani M. M. R. A., Pehlivan H. A step-by-step solution methodology for mathematical expressions. *Symmetry*. 2018. Vol. 10, No. 7. Article 285. DOI: https://doi.org/10.3390/sym10070285.
6. Bourhis P., Reutter J. L., Suárez F., Vrgoč D. JSON: data model, query languages and schema specification. Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2017), Chicago, IL, USA, May 14–19, 2017. P. 123–135. DOI: https://doi.org/10.1145/3034786.3056120.
7. Bahta R., Atay M. Translating JSON data into relational data using schema-oblivious approaches. Proceedings of the 2019 ACM Southeast Conference,

Kennesaw, GA, USA, April 18–20, 2019. P. 233–236. DOI: https://doi.org/10.1145/3299815.3314467.
8. Pegg E. Jr. Canonical Polyhedra. Wolfram Research. URL: https://datarepository.wolframcloud.com/resources/Canonical-Polyhedra/ DOI: https://doi.org/10.24097/wolfram.98862.data.
9. Pezoa F., Reutter J. L., Suárez F., Ugarte M., Vrgoč D. Foundations of JSON Schema. Proceedings of the 25th International Conference on World Wide Web (WWW '16), Montreal, Quebec, Canada, April 11–15, 2016. P. 263–273. DOI: https://doi.org/10.1145/2872427.2883029.
10. Lv T., Yan P., He W. On massive JSON data model and Schema. Journal of Physics: Conference Series. 2019. Vol. 1302, No. 2. Article 022031. DOI: https://doi.org/10.1088/1742-6596/1302/2/022031.
11. Ooms J. The jsonlite package: a practical and consistent mapping between JSON data and R objects. arXiv. 2014. arXiv:1403.2805. URL: https://arxiv.org/abs/1403.2805.
12. Jain S., de Buitleir A., Fallon E. A review of unstructured data analysis and parsing methods. Proceedings of the 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, March 12–14, 2020. P. 164–169. DOI: https://doi.org/10.1109/ESCI48226.2020.9167588.
13. Bacchelli A., Mocci A., Cleve A., Lanza M. Mining structured data in natural language artifacts with island parsing. *Science of Computer Programming*. 2017. Vol. 150. P. 31–55.

APPROACHES TO PARSING NON-STANDARDLY ORGANIZED GEOMETRIC DATA

Maryna Morozova, Olena Sydorenko

The article is devoted to investigating the problem of the structural gap between non-standardly organized geometric datasets and the requirements of software tools for statistical analysis. A review of scientific sources related to the topic of the study and adjacent research areas is conducted. The methodology and approaches described in the paper are implemented using the Canonical Polyhedra dataset as an example, which contains metric and topological characteristics of 2,907 polyhedra for the Wolfram Mathematica environment. The primary feature of the dataset under investigation is its specific structure, namely the representation of spatial objects not in the form of traditional tables but as abstract syntax trees. The article provides a detailed analysis of the dataset architecture, identifying its principal data nodes and highlighting symbolic mathematical expressions that may potentially create difficulties during parsing. In the dataset, some characteristics of geometric figures are represented not

numerically but symbolically. While this ensures absolute mathematical precision, it also makes the data unsuitable for automated computation without prior transformation. Accordingly, the limitations of analyzing such a dataset using standard data-processing libraries are further substantiated. The core of the research methodology is the practical development of a recursive parsing algorithm demonstrated on a specific dataset. The algorithm is described in a universal and generalized form, making it applicable to a broader range of related tasks. The paper outlines the logic of traversing syntax tree nodes, identifying their headers, and normalizing data types. The result of this approach is the transformation of an abstract syntax tree into a normalized table suitable for further statistical analysis. The practical significance of the study lies in identifying universal approaches to parsing complex nested data hierarchies, which often serve as valuable sources of information about spatial geometric figures for further investigation and application in fields such as computational geometry, machine learning, and related disciplines.

Keywords: dataset, statistical analysis, parsing, computer mathematics system, polyhedron.

References

1. Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., Weaver, C., Lee, B., Brodbeck, D., & Buono, P. (2011). Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4), 271–288. DOI: https://doi.org/10.1177/1473871611415994 [In English].
2. Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. *International Journal of Computer Science*, 1(1), 111–117. [In English].
3. Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1–23. DOI: https://doi.org/10.18637/jss.v059.i10 [In English].
4. McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (Austin, TX, USA, June 28 – July 3, 2010) (pp. 56–61). DOI: https://doi.org/10.25080/Majora-92bf1922-00a [In English].
5. Hosseinpour, S., Milani, M. M. R. A., & Pehlivan, H. (2018). A step-by-step solution methodology for mathematical expressions. *Symmetry*, 10(7), Article 285. DOI: https://doi.org/10.3390/sym10070285 [In English].
6. Bourhis, P., Reutter, J. L., Suárez, F., & Vrgoč, D. (2017). JSON: Data model, query languages and schema specification. In Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS 2017) (Chicago, IL, USA, May 14–19, 2017) (pp. 123–135). DOI: https://doi.org/10.1145/3034786.3056120 [In English].

7. Bahta, R., & Atay, M. (2019). Translating JSON data into relational data using schema-oblivious approaches. In Proceedings of the 2019 ACM Southeast Conference (Kennesaw, GA, USA, April 18–20, 2019) (pp. 233–236). DOI: https://doi.org/10.1145/3299815.3314467 [In English].
8. Pegg, E., Jr. Canonical Polyhedra. Wolfram Research. URL: https://datarepository.wolframcloud.com/resources/Canonical-Polyhedra/. DOI: https://doi.org/10.24097/wolfram.98862.data [In English].
9. Pezoa, F., Reutter, J. L., Suárez, F., Ugarte, M., & Vrgoč, D. (2016). Foundations of JSON Schema. In Proceedings of the 25th International Conference on World Wide Web (WWW ’16) (Montreal, Quebec, Canada, April 11–15, 2016) (pp. 263–273). DOI: https://doi.org/10.1145/2872427.2883029 [In English].
10. Lv, T., Yan, P., & He, W. (2019). On massive JSON data model and Schema. *Journal of Physics: Conference Series*, 1302(2), Article 022031. DOI: https://doi.org/10.1088/1742-6596/1302/2/022031 [In English].
11. Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between JSON data and R objects. arXiv:1403.2805. URL: https://arxiv.org/abs/1403.2805 [In English].
12. Jain, S., de Buitleir, A., & Fallon, E. (2020). A review of unstructured data analysis and parsing methods. In Proceedings of the 2020 International Conference on Emerging Smart Computing and Informatics (ESCI) (Pune, India, March 12–14, 2020) (pp. 164–169). DOI: https://doi.org/10.1109/ESCI48226.2020.9167588 [In English].
13. Bacchelli, A., Mocci, A., Cleve, A., & Lanza, M. (2017). Mining structured data in natural language artifacts with island parsing. *Science of Computer Programming*, 150, 31–55. [In English].

Матеріал надійшов до редакції 29.04.2026

Прийнято до друку 13.05.2026 р.