Горобець Олена Олександрівна
аспірантка
Національна академія статистики, обліку та аудиту,
м. Київ, Україна

НАУКОВІ ДАНІ ЯК ІННОВАЦІЙНА ТЕНДЕНЦІЯ ЦИФРОВОГО СУСПІЛЬСТВА

Закономірні трансформації суспільства призвели, як відомо, до Четвертої технологічної революції, яка в свою чергу характеризується поширенням інформаційних комп'ютерних технологій та посиленим розвитком споріднених з ними сфер життєдіяльності. Усе це спровокувало появу великих даних – «природнього» середовища наукових даних. Зараз, маючи справу з неконтрольованим, неструктурованим та хаотичним потоком даних, який несе у собі потенційну інформацію про різноаспектні складові життєдіяльності людини, виокремити якісну інформацію стає надважким завданням. Тому, практично усі події в світі спеціально кодуються в дані, що в свою чергу, стало першим кроком до їх гранульованості. А саме: мілкі дані – дані, які стосуються об'єкта дослідження (зазвичай, цим даним притаманна повільна трансформація); глибокі дані – дані, які стосуються життєдіяльності людини (швидкозмінні дані, наприклад, інформація про поїздки, шопінг та ін.), мікро-дані –дані, які які мають надвисоку (у порівнянні з глибокими даними) ступінь деталізації об'єкта; нано-дані – нові дані, які беруть початок в середовищі глибоких даних, а за своїми характеристиками схожі до атомарних явищ [1]. Зауважимо, що кожний процес життєдіяльності людини, або ж наприклад бізнес-процес, характеризується, окрім вище зазначених даних, ще й темними даними, які повторно не використовуються, тим самим провокуючи логічне запитання про доцільність їх зберігання.

Для академічного світу особливий інтерес представляють наукові (дослідницькі) дані, які можуть претендувати на окреме місце в середовищі великих даних (наприклад, їх можна виокремити як категорію у грануляції нано-даних) і належати до статистичної інформації. Варто зауважити, що в цій роботі терміни «дослідницькі дані» й «наукові дані» мають однаковий смисл.

Оксфордський словник визначає наукові дані як зафіксовану інформацію (незалежно від форми та сфери в якій вона може існувати) яка необхідна для підтримки чи підтвердження спостережень або результатів дослідницького проекту [2].

На основі знань про великі дані та статистичні дані, а також враховуючи вищезазначене твердження, доцільно сформувати власне визначення терміну «дослідницькі дані», де дослідницькі дані – це результати різногалузевої наукової діяльності у вигляді висновків наукових досліджень які містяться в традиційних та цифрових джерелах інформації.

Відомо, що результати наукової діяльності оприлюднюються в наукових періодичних виданнях. Тому академічні видавництва є невід'ємними учасниками «життєвого циклу» наукових даних. На сьогодні академічні видавництва є глобальним середовищем. В 2015 р. загальний прибуток семи головних академічних видавничих груп дорівнював 59,328 млрд євро [3].

Видавнича компанія Elsevier, говорячи про наукові дані, підкреслює необхідність обміну даними, акцентуючи увагу на тому, що ці дані повинні бути у відкритому доступі для повторного використання з метою отримання нових наукових даних, що, безпосередньо, робить процес дослідження більш продуктивним [4]. Зі звіту Figshare «Стан відкритих даних 2018», відомо, що відкриті дані стали більш інтегрованими в дослідницьку спільноту – 64 % респондентів показують, що вони зробили свої дані відкритими в 2018 р. Однак, разом з тим, 60 % респондентів ніколи не чули про принципи FAIR – керівництво з покращення можливостей повторного використання академічних даних [5].

Важливість відкритості, упорядкованості та збереження наукових даних для повторного використання в епоху цифровізації полягає в елементарному розумінні того, що сьогодні дані – це невід'ємне енергетичне джерело сучасності, а наукові дані – це його потоки, які живлять усі галузі глобалізованого світу. Більшість прогресивних теорій та практичних досліджень, які можна реалізувати для удосконалення і, відповідно, розвитку технологій у сільському господарстві, харчовій промисловості, металургії, екології, охороні здоров'я та ін. були опубліковані в науковій періодиці. Виходячи з цього, необхідно розуміти глобальну користь своєчасності застосування наукових даних. Результати як сучасних, так і ретроспективних досліджень повинні мати єдине місце зберігання, зрозумілу ієрархію та бути доступними.

Цікавий приклад застосування наукових даних та інноваційних технологій в практичній діяльності, можна знайти в медицині, а саме – застосування системи Watson IBM. Суть роботи системи Watson IBM полягає в тому, що знаючи анамнез пацієнта, система з допомогою алгоритмів аналізує масив даних об'ємом близько 200 млн сторінок, який включає в себе словники, підручники, монографії, веб-сайти, звіти, архіви записів

лікарів, газетні архіви та наукові журнали. Зрозумівши ключові індикатори захворювання, система запускає для пошуку сотні різних алгоритмів, структурує, порівнює та ранжує отриману інформацію з метою отримання діагнозу. У 2013 р. Watson IBM уже допомагав діагностувати захворювання та покращувати курси лікування пацієнтів в Клівлендській клініці. Зараз, на прикладі Watson IBM можна спостерігати за тим, як люди навчаються у машини логіці мислення [6].

Продовжуючи тему використання наукових даних та інноваційних технологій на прикладі медичної галузі, логічно припустити думку про неспівставність штучного інтелекту та людських можливостей. Повертаючись до медицини, наведемо приклад: база наукових медичних журналів MEDLINE містить у собі близько 5600 журналів, в яких із року в рік публікується декілька сотень тисяч наукових статей, які, в свою чергу, мають декілька мільйонів наукових результатів та метаданих. І це лише один приклад, з однією базою даних, в одній галузі.

Таким чином, досліджуючи середовище наукових даних потрібно враховувати їх глобальність. Уже зараз необхідно замислитися над тим, хто буде займатися їх управлінням. З одного боку, це можуть бути спеціалісти сфери програмного забезпечення та IT, адже вони здатні запропонувати методи архівування (наприклад, хмарне зберігання) та структуризації, а також повторного використання таких даних. З другого боку – це можуть бути масштабні видавництва наукових публікацій, враховуючи їх зацікавленість та ініціативність (наприклад, база даних Scopus). З третього боку – статистичні служби, враховуючи наявність інфраструктур та безпосередній досвід роботи з статистичними даними та статистичною інформацією. З четвертого боку – бібліотеки, як перші архіви наукової інформації (успішний сучасний приклад – інституційні репозитарії).

Окрім управління науковими даними, що й провокує до подальшого вивчення цієї теми, залишаються відкритими багато питань, зокрема: яким чином можна організувати фінансування, яку ієрархію варто застосовувати до їх структуризації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. European Commission. Eurostat. CROS. New Techniques and Technologies for Statistics (NTTS-2019). Retrieved from https://coms.events/ntts2019/data/x_abstracts/x_abstract_135.pdf.
2. Research data Oxford. Research data glossary. Retrieved from http://researchdata.ox.ac.uk/home/glossary/

3. Wischenbart R. The Global Ranking of the Publishing Industry 2015. Retrieved from https://www.wischenbart.com/upload/Wischenbart_Ranking 2015_analysis_final.pdf
4. Elsevier. Sharing research data. Retrieved from https://www.elsevier.com/authors/author-resources/research-data
5. Figshare. The State of Open Data Report 2018. Retrieved from https://figshare.com/articles/The_State_of_Open_Data_Report_2018/7195058
6. Форд М. Пришестя роботів. Техніка і загроза майбутнього безробіття // Мартін Форд. – К.: Наш формат, 2016. – С. 134-147.