БАЗИ ДАНИХ ЦИТУВАНЬ ТА ПОШУКОВІ ІНСТРУМЕНТИ ДЛЯ НАУКОВЦІВ МАЙБУТНЬОГО

Ми сідаємо за комп'ютер і з легкістю проводимо пошук найновіших наукових публікацій. Незалежно від того, чи працюємо ми з передплаченими базами Web of Science або Scopus, чи застосовуємо безкоштовний Google Scholar, ми можемо швидко відсортувати знайдені документи за кількістю цитувань та побачити, які автори згадували ту, чи іншу публікацію у своїх роботах. Ця найпопулярніша трійка покажчиків наукових цитувань пропонує нам сьогодні пошукові можливості, які ще зовсім недавно були нам недоступні, однак уже з'являються нові інструменти, що готові посперечатися за лідерство в цій галузі.

Сергій Назаровець
канд. наук із соціальних комунікацій,
заст. директора
з наукової роботи Державної науково-технічної бібліотеки України,
м. Київ

Пошук інформації є невід'ємною складовою наукової роботи. Успішний вчений повинен знати найновіші результати досліджень у своїй галузі, дізнатися, хто ще працює над цією проблемою, які журнали публікують відповідні статті, які конференції присвячені цій проблематиці, хто фінансує проведення подібних досліджень тощо. Щодня у світі з'являється дуже багато наукових публікацій, тому практично неможливо перечитати «від А до Я» абсолютно всі наукові роботи у своїй галузі. Щоб марно не витрачати час на перегляд неактуальних публікацій, науковці використовують спеціальні пошукові техніки та інструменти. Так, сучасні популярні пошукові наукові системи та бази даних цитувань дозволяють користувачам проводити пошук рецензованих публікацій, а також дізнаватися про джерела та кількість цитувань знайдених публікацій. Велика кількість цитувань публікації та згадок авторської роботи у працях інших вчених може свідчити про її наукову впливовість, тому сервіси автоматично обраховують й пропонують користувачам різноманітні показники цитованості для вчених, груп, установ та навіть цілих країн. Спочатку такі метрики повинні були допомагати бібліотекарям у відборі відповідних наукових видань для передплати. Проте сьогодні ці показники активно використовуються у багатьох країнах як інструмент для оцінювання наукової продуктивності й прийняття рішень у сфері науки та техніки.

Точність та правильність інтерпретації результатів таких оцінювань залежить насамперед від надійності інструментів та джерел даних, що використовуються для їхнього проведення. Серед таких інструментів фахівці в галузі наукометрії, які займаються кількісним вивченням науки, наукової комунікації та наукової політики, відзначають два комерційні продукти – бібліографічні бази даних Web of Science Core Collection та базу даних цитувань наукової літератури Scopus.

Велика трійка: Web of Science, Scopus і Google Scholar

Бази **Web of Science**[1] беруть свій початок від ініціативи одного з основоположників сучасної наукометрії, американського вченого та підприємця ***Юджина Гарфілда***, який ще у 60-х роках минулого століття розпочав укладати низку інноваційних на той час бібліографічних продуктів, зокрема: Current Contents, Science Citation Index (SCI), Journal Citation Reports, Index Chemicus. Спершу ці індекси цитувань (індекс – у значенні покажчик, список, перелік; не слід плутати з цифровим показником) існували виключно у друкованому вигляді, а згодом, у 1980-х роках, їх перетворили на електронні бази даних. За свою історію бази Web of Science належали багатьом комерційним компаніям і зараз є власністю компанії Clarivate Analytics.

Довгий час покажчики Web of Science були де-факто єдиними надійними постачальниками даних наукових цитувань. Монопольне становище продукту не йшло на користь його користувачам, адже не стимулювало власників Web of Science вкладати кошти у вдосконалення сервісу. Повільним розвитком Web of Science сповна скористався видавничий гігант Elsevier, який у 2004 році створив базу даних рецензованої літератури та інструмент для відстежування наукових цитувань **Scopus**[2].

[1]https://apps.webofknowledge.com/
[2]https://www.scopus.com/

Новий продукт **Elsevier** пропонував користувачам кращі пошукові можливості та охоплював значно більше видань, ніж його основний конкурент. Натомість Web of Science пропонував значно глибші журнальні архіви. Сьогодні різниця між цими двома базами суттєво зменшилася – бази Web of Science покращили пошуковий інтерфейс та розпочали індексувати значно більше видань, натомість Scopus суттєво поглибив свої архіви.

У базах **Web of Science** та **Scopus** представлена обмежена кількість наукових видань. Упорядники цих баз дотримуються критеріїв відбору наукового контенту – видання повинні відповідати певним видавничим вимогам і, крім того, незалежні експерти оцінюють актуальність представлених у них результатів досліджень. Комп'ютерні алгоритми встановлюють цитатні зв'язки між представленими в базах документами, що робить ці інструменти корисними для пошуку впливових наукових досліджень та для проведення наукометричних досліджень.

Попри наявність суворих критеріїв відбору контенту Web of Science та Scopus регулярно припиняють індексацію періодичних видань через неетичні маніпуляції їх редакцій з цитуваннями, або через падіння наукового рівня видання. При цьому вже проіндексовані публікації видання-порушника, у переважній більшості випадків, назавжди залишаються у базі, що негативно впливає на правдивість наукометричних оцінювань. Також Scopus та Web of Science часто критикують через повільне виправлення помилок, недостатнє охоплення соціогуманітарних наук та неангломовних журналів.

Вказані недоліки не перешкоджають Scopus та Web of Science залишатися сьогодні найвживанішими інструментами наукометристів та управлінців. Популярність серед останніх пов'язана з тим, що дані Scopus та Web of Science використовуються при укладанні популярних рейтингів університетів, наприклад, таких, як **Academic Ranking of World Universities** (ARWU) та **Times Higher Education World University Ranking** (THE WUR). І хоча нові результати досліджень доводять, що методологія цих рейтингів є недостатньо прозорою й не враховує місію університету, якість викладання, ефективність та соціальний вплив проведених досліджень[3], студенти, викладачі та управлінці часто покладаються на оцінки цих рейтингів.

Доступ до Web of Science та Scopus потребує інституційної передплати. Відповідно, щоб скористатися цими інструментами, дослідник повинен працювати в установі з відповідним бюджетом на передплату наукових ресурсів. Можливо, подібна поведінка постачальників цитатних даних поглибила б інформаційний розрив між вченими багатих та бідних установ, однак одночасно з появою Scopus, у 2004 році з'явилася, мабуть, найпопулярніша на сьогодні наукова пошукова система у світі **Google Scholar**[4], що запропонувала користувачам безкоштовно проводити пошук академічних документів та джерел цитувань.

Індійський студент ***Анураг Ачарія***, який згодом став одним з головних розробників системи **Google Scholar**, подібно, як і наші студенти, часто потерпав від браку доступу до повних текстів наукових публікацій. Ачарія помітив, що дуже важливо хоча б знати, що певна наукова робота взагалі існує, – тоді можна, наприклад, написати лист до автора з проханням надіслати копію роботи[5]. На відміну від баз даних Web of Science та Scopus, що індексують лише невеличку частину наукових видань, що видаються у світі, Google Scholar намагається зібрати інформацію про всі наукові документи в інтернет. Сьогодні такий підхід може нам здатися очевидним, однак у 2004 році це був величезний технологічний стрибок.

За розміром бази даних та швидкістю індексації система Google Scholar значно переважає інші академічні ресурси (наприклад, Google Scholar індексує майже втричі більше наукових документів ніж Scopus, а середня різниця у затримці індексації між цими базами становить приблизно 2 місяці[6]). Система дозволяє користувачам дізнатися кількість цитувань документів та джерела цих цитувань, автоматично обраховує ряд метрик, пропонує науковцям створити власний профіль в системі[7] та переглянути щорічний рейтинг журналів, укладений згідно з даними Google Scholar[8]. Великою перевагою та особливістю Google Scholar є те, що система індексує повні тексти майже усіх найбільших наукових видавництв. Завдяки цьому Google Scholar може показувати фрагменти публікацій, що відповідають пошуковим запитам, користувачі швидше знаходять потрібні наукові роботи, а видавці отримують нових клієнтів.

Пошукова система Google Scholar – це дуже популярний інструмент саме для виявлення наукової інформації, проте система має ряд серйозних обмежень щодо використання її даних для проведення наукометричних досліджень та оцінки впливовості наукових робіт. Система цілковито залежить від точності роботи комп'ютерної програми, при цьому Google Scholar часто індексує ненаукові документи, кількість цитувань визначається приблизно й система незахищена від неетичних маніпуляцій, зокрема від штучного нарощування кількості цитувань. Проведення наукометричних досліджень на основі даних Google Scholar ускладнюється ще й відсутністю доступу до прикладного програмного інтерфейсу, через що складно експортувати бібліографічні дані системи для їх подальшого аналізу.

Нові індекси цитувань

Сучасний розвиток інформаційно-комп'ютерних технологій та активне використання їх у науковій комунікації відкрили вченим, програмістам та бібліотекарям нові можливості щодо створення наукометричних інструментів. Особливо продуктивними в цьому плані виявилися десяті роки XXI століття – комерційні компанії запропонували нові пошукові системи та бази даних наукових цитувань, й водночас з'явилися «відкриті» індекси цитувань.

[3]Gadd, E. (2020). University rankings need a rethink. Nature, 587(7835), 523–523.
[4]https://scholar.google.com/
[5]За цим принципом працює Open Access Button https://openaccessbutton.org/
[6]Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. Journal of Informetrics, 10(2), 533–551.
[7]https://scholar.google.com.ua/citations
[8]https://scholar.google.com.ua/citations?view_op=top_venues

Наукометричні ресурси критикують через те, що вони корисні для оцінки виключно наукового впливу робіт, проте не підходять для оцінювання технологічного впливу результатів досліджень. Австралійська некомерційна організація Cambia працює над виправленням цього недоліку й у 2000 році створила патентний пошуковий сервіс **The Lens**[9], що дозволяє дізнатися про цитатні зв'язки між патентами та науковими публікаціями.

У 2015 році Інститут штучного інтелекту Аллена створив академічну пошукову систему **Semantic Scholar**[10]. Однією з ключових особливостей цієї пошукової системи є те, що за допомогою технологій машинного навчання вона дозволяє користувачам дізнатися, у якому саме розділі роботи було процитовано конкретну публікацію – у вступі, методах чи результатах. Річ у тому, що багато наукових цитувань походять з розділів публікацій, де представлено огляд літератури з теми, а це не доконечно свідчить, що процитована робота була корисною для конкретного дослідження. Ця проблема вже давно відома в наукометрії, однак лише в цьому столітті ми отримали технології, що можуть допомогти з її вирішенням.

У 2016 році відбувся перезапуск пошукової системи **Microsoft Academic**[11] (перший запуск відбувся у 2012 році), що використовує технології витягнення вебданих, подібно до Google Scholar, а також застосовує технології обробки природної мови. Додаткова цінність цього проєкту для наукометричної спільноти – це набір даних Microsoft Academic Graph (MAG)[12], що містить записи наукових публікацій, цитатні зв'язки між цими публікаціями, а також авторами, установами, журналами, конференціями та галузями дослідження. Запис оновлюється щотижня та доступний для повторного використання за умовами відкритих ліцензій.

Компанія Digital Science у 2018 році запустила платформу **Dimensions**[13], що пропонує користувачам різні варіанти доступу до її функціоналу. Платформа дозволяє індивідуальним користувачам безкоштовно проводити пошук серед понад 111 мільйонів публікацій та наборів даних, переглядати кількість та джерела їх цитувань. Записи публікацій контекстуально пов'язані з даними про фінансування, патентами, клінічними настановами, проте доступ до цієї інформації вимагає вже інституційної передплати. Таким чином, творці платформи Dimensions прагнуть надати своїм користувачам інструмент, що охоплює повний контекст наукових досліджень, та намагаються впровадити гнучкіші моделі передплати.

Згадка публікації у роботах інших вчених не лише може свідчити про її наукову впливовість, оскільки вона може там також піддаватись критиці, а традиційні наукометричні інструменти не здатні розпізнавати подібні «негативні» цитування. У 2018 році американський стартап **Scite**[14] створив новий індекс, що пропонує користувачам «розумні» цитування – за допомогою технології обробки природної мови алгоритм Scite намагається розпізнати контекст цитування й показує біля кожного запису не просто загальну кількість цитувань документу, а групує цитування документу за відповідними категоріями: «Підтримка», «Згадування», або «Незгода».

Ініціатива для відкритих цитувань

Комерціалізація авторитетних інструментів для відстежування наукових цитувань Web of Science та Scopus пов'язана, передусім, з технічною складністю процесу встановлення зв'язків між науковими документами – наукові видання використовують різні стилі цитувань літератури, що передбачають різну послідовність згадування елементів публікації (автор, назва публікації, назва видання, рік, том, номер тощо) та використання різних розділових знаків між цими елементами. Введення у практику наукового видавництва використання унікальних цифрових ідентифікаторів об'єкта **Digital Object Identifier** (DOI)[15] відкрило нові технологічні можливості щодо надійного встановлення цитатних зв'язків між науковими документами й, відповідно, для створення баз наукових цитувань.

Унікальний цифровий ідентифікатор об'єкта DOI, що складається з цифр та літер, пов'язаний з посиланнями на вебсторінку, де знаходяться сам об'єкт або інформація про нього. DOI реєструється для об'єкта одноразово і залишається незмінним, що суттєво спрощує ідентифікацію публікацій та убезпечує від втрати посилань при зміні вебадреси. Присвоювати DOI можуть тільки офіційні реєстраційні агенції, які входять до International DOI Foundation. Сьогодні більшість наукових видавців присвоюють DOI своїм публікаціям за допомогою агенції **Crossref**[16]. Для реєстрації DOI видавцю необхідно передати в агенцію метадані публікації (дані, що описують конкретний об'єкт – автор, назва, рік публікації тощо). Crossref надає доступ до цієї бази метаданих, а також закликає своїх членів додавати до метаданих пристатейні списки літератури, що дозволить відстежувати цитування публікацій[17].

Наявність та відкритість інфраструктури Crossref для роботи з бібліографічними (мета)даними дала новий поштовх для створення **Ініціативи для відкритих цитувань** (**Initiative for Open Citations**, I4OC) – співпраці наукових видавців, дослідників та інших стейкхолдерів щодо сприяння необмеженій доступності даних про цитування в науковій літературі[18]. У 2017 році Ініціативу I4OC підтримало багато провідних наукових видавництв, серед яких: BMJ, Cambridge University Press, Royal Society of Chemistry, SAGE Publishing, Springer Nature, Taylor & Francis, Wiley. У грудні 2020 року найбільший видавець наукової літератури Elsevier також приєднався до I4OC. За попередніми підрахунками станом на січень 2021 року відкрито 84 % наукових цитувань (53,6 мільйона статей з відкритими списками використаної літератури депоновано в Crossref).

[9]https://www.lens.org/
[10]https://www.semanticscholar.org/
[11]https://academic.microsoft.com/
[12]https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/
[13]https://app.dimensions.ai/discover/publication
[14]https://scite.ai/
[15]https://www.doi.org/
[16]https://www.crossref.org/
[17]https://www.crossref.org/services/cited-by/
[18]https://i4oc.org/

Незалежна організація **OpenCitations**[19] опікується розвитком інфраструктури відкритих бібліографічних даних та даних про цитування на основі технологій *Semantic Web* (Linked Data).

Будь-який видавець, що користується послугами Crossref, може з легкістю та безкоштовно підтримати Ініціативу для відкритих цитувань. І саме на використанні цих нових ініціатив базується принцип роботи Open Ukrainian Citation Index.

У 2019 році фахівці Державної науково-технічної бібліотеки України представили **Open Ukrainian Citation Index** (OUCI) – пошукову систему і базу даних наукових цитувань, що використовує відкриті дані Crossref[20]. База даних OUCI містить метадані усіх наукових видань, що отримують DOI від Crossref, обраховує цитування та дозволяє переглянути джерела цитувань (DOI to DOI). Окрім базових фільтрів, таких, як рік, тип, назва видання, видавець, індексація в покажчиках, наявність публікації у відкритому доступі тощо, для українських користувачів система пропонує також спеціальні пошукові можливості.

База даних OUCI покликана спростити пошук наукових публікацій та привернути увагу редакцій до проблеми повноти, якості й відкритості метаданих наукових видань. Творці системи OUCI очікують, що розвиток подібних відкритих наукометричних інструментів зробить інформацію про наукові цитування доступною для всіх, а не лише для вчених з країн, що можуть собі дозволити передплату комерційних ресурсів. Також OUCI можна використовувати як джерело даних для проведення наукометричних досліджень, зокрема в галузях суспільних та гуманітарних наук, де значна частина наукових результатів зосереджена на регіональній тематиці та аудиторії, а відтак вони часто публікуються у національних неангломовних виданнях, що не представлені в комерційних базах даних наукових цитувань.

Зважаючи на активний розвиток відкритих індексів цитувань для ефективного представлення видання у науковому інформаційному просторі, їхні редакції повинні дбати про передачу максимально якісних метаданих реєстраційним агенціям DOI. Також майбутнє відкритих наукометричних інструментів залежить й від підтримки академічною спільнотою **Ініціативи для відкритих анотацій** (**Initiative for Open Abstracts**, I4OA), що закликає всіх наукових видавців відкрити анотації своїх публікацій та зробити їх доступними[21]. Сьогодні багато анотацій представлені в різних комерційних бібліографічних базах даних, проте часто доступ до цих ресурсів вимагає передплати, а записи не придатні для обробки комп'ютерними програмами. Відкритий доступ до анотацій допоможе науковим видавцям максимізувати видимість та розширити читацьку аудиторію своїх журналів та книг. Відкриті анотації допоможуть вченим швидко знаходити та цитувати важливі публікації, сприятимуть включенню цих наукових робіт у систематичні огляди, спростять та розширять можливості щодо використання технологій інтелектуального аналізу тексту, обробки природної мови та штучного інтелекту в наукометрії.

Відкрите майбутнє наукометрії

Сьогодні «велика трійка» наукометричних інструментів зберігає свої провідні позиції на ринку постачальників інформації про цитування. Завдяки великому охопленню інтернет-ресурсів Google Scholar надзвичайно популярна серед пошукувачів наукової інформації. Однак дані про цитування Google Scholar надто неточні й не підходять для оцінки наукової ефективності. Попри те, що наукометричні показники не гарантують об'єктивної й справедливої оцінки наукових результатів[22], у багатьох країнах світу для оцінювання роботи вчених, наукових груп та установ використовують дані Scopus та Web of Science. Вибірковий підхід до індексації контенту та алгоритми встановлення цитатних зв'язків цих комерційних баз пропонує нині найякісніші цитатні дані для проведення наукометричних аналізів.

Проте чи варто сплачувати такі кошти за доступ до даних про наукові цитування? Чи власники цих баз зацікавлені у науковому прогресі, чи вони більше піклуються про те, аби адміністрації університетів продовжували оцінювати результати своєї роботи за допомогою їх інструментів?

Безумовно, ще десять років тому не існувало жодної гідної альтернативи цим комерційним покажчикам, тому їх висока вартість передплати, напевно, була виправдана. Однак впродовж останніх років з'явилося багато схожих інструментів, що пропонують розпізнавання контексту цитувань, краще охоплення, цікавіші пошукові функції та відкриті метадані. Тому відповідь на запитання щодо вартості передплати сьогодні вже далеко неоднозначна.

Компанії Clarivate Analytics та Elsevier продовжують витрачати значні фінансові ресурси на розвиток своїх брендів та рекламу своєї продукції. Цей активний маркетинг спонукає наших науковців, управлінців та бібліотекарів до парадоксальної поведінки – вони слідкують за точністю інформації про свої установи у продуктах цих комерційних компаній, безкоштовно виявляють помилки та покращують якість даних, а потім купують доступ до цих же даних. Натомість розповсюдження наукових метаданих під відкритими ліцензіями не вимагало б від наших наукових стейкхолдерів більших зусиль, однак дозволило б їх повторне використання як в комерційних, так і в некомерційних аналітичних інструментах.

Навіть якщо всі дані наукових цитувань стануть незабаром вільнодоступними, потрібно буде ще докласти значних зусиль щодо написання комп'ютерних алгоритмів для їх обробки та створення зручних користувацьких інтерфейсів. Зараз складно прогнозувати, чи збережуть в майбутньому наявні комерційні інструменти своє панівне становище, чи комусь із нових гравців таки вдасться стати основним постачальником даних для проведення наукометричних досліджень. Однак підтримка відкритих ініціатив гарантовано допоможе в розбудові нової справедливішої академічної інфраструктури, де кожен, хто зацікавлений у поширені наукових знань, зможе скористатися безперешкодним доступом до цитувань. ■

[19]https://opencitations.net/
[20]https://ouci.dntb.gov.ua/
[21]https://i4oa.org/
[22]https://sfdora.org/read/