

Розробка харвестера

- [Зміни функціоналу 14.08.2026](#)

Зміни функціоналу

14.08.2026

Розділ 1

Головна сторінка тепер розповідає про масштаб репозиторію

Раніше головна сторінка обмежувалась пошуковим рядком і списком міжнародних агрегаторів. Тепер вона одразу показує, наскільки живий і великий репозиторій — трьома живими числами, двома графіками активності по роках і повністю переробленим списком мов.

Три живі числа замість статичного тексту

Спочатку тут був фіксований текст на кшталт «186 млн публікацій у світі» — загальна, нічим не підтверджена статистика. Замінили на три показники, які рахуються напряду з Solr-індексу при кожному завантаженні сторінки:

- **Проіндексованих ресурсів** — загальна кількість записів.
- **Проіндексованих джерел** — кількість підключених журналів/репозиторіїв.
- **Повнотекстових документів** — скільки записів мають доступний PDF.

Оновлено: 13.08.2026 20:51



ВІДКРИТА НАУКА В НАЦІОНАЛЬНІЙ АКАДЕМІЇ НАУК УКРАЇНИ

Мета впровадження відкритої науки в Національній академії наук України — підвищити представленість результатів досліджень науковців НАН України у глобальному інформаційному середовищі з використанням сучасних технічних та інформаційних засобів у інтересах розвитку науки в Україні та міжнародного наукового співробітництва з оцінкою такої представленості за певними показниками.



Про відкриття науки в НАН України

172

проіндексованих ресурсів

1

проіндексованих джерел

172

повнотекстових документів доступних в системі

СПИСОК МІЖНАРОДНИХ АГРЕГАТОРІВ, ДЕ ІНТЕГРОВАНО ХАРВЕСТЕР ВІДКРИТОЇ НАУКИ НАН УКРАЇНИ



CORE (Connecting REpositories)

CORE (Connecting REpositories) — це неприбутковий сервіс, який підтримується The Open University та членами CORE.



OpenDOAR

OpenDOAR — це надійний глобальний каталог перевірених репозиторіїв відкритого доступу.



OpenAIRE

OpenAIRE AMKE — це неприбуткова організація, мета якої — сприяти відкритій науковій діяльності та покращити виявлення, доступність, обмін і повторне використання результатів досліджень у глобальному масштабі.

[Головна сторінка](#)

[Сторінка репозиторію](#)

[Головна сторінка](#)

[Сторінка репозиторію](#)

[Головна сторінка](#)

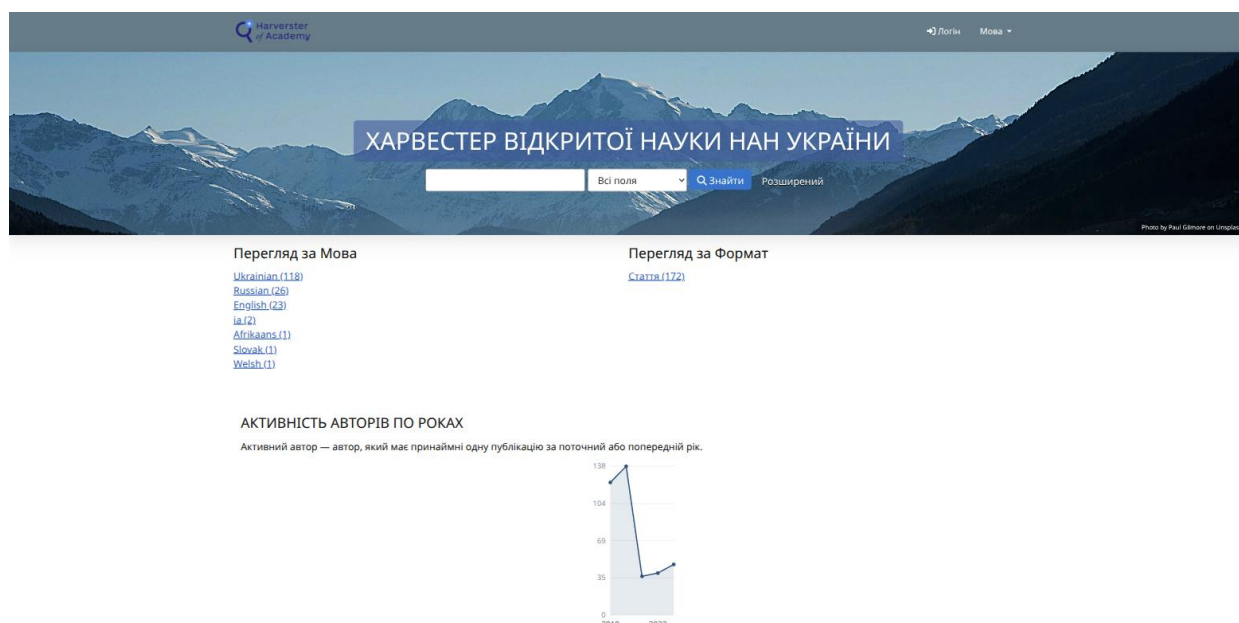
[Сторінка репозиторію](#)

Головна сторінка — секція з живими показниками репозиторію

Графіки активності по роках

Додано два графіки: кількість «активних авторів» (тих, хто публікувався за поточний або попередній рік) і кількість ресурсів по роках. Дані рахуються окремим фоновим скриптом (важкий Solr-запит, тому не на кожен запит сторінки) і кешуються.

Перша версія була стовпчиковою діаграмою — і на реальному масштабі (репозиторій охоплює публікації з 1926 по 2026 рік) підписи років накладалися одне на одне й ставали нечитабельними. Графік переробили на лінійний: підписи років показуються рідше й без накладання, точне значення для будь-якого року з'являється у підказці при наведенні миші, а сам графік завжди розтягується на всю ширину сторінки — без горизонтальної прокрутки, скільки б років не було в даних.



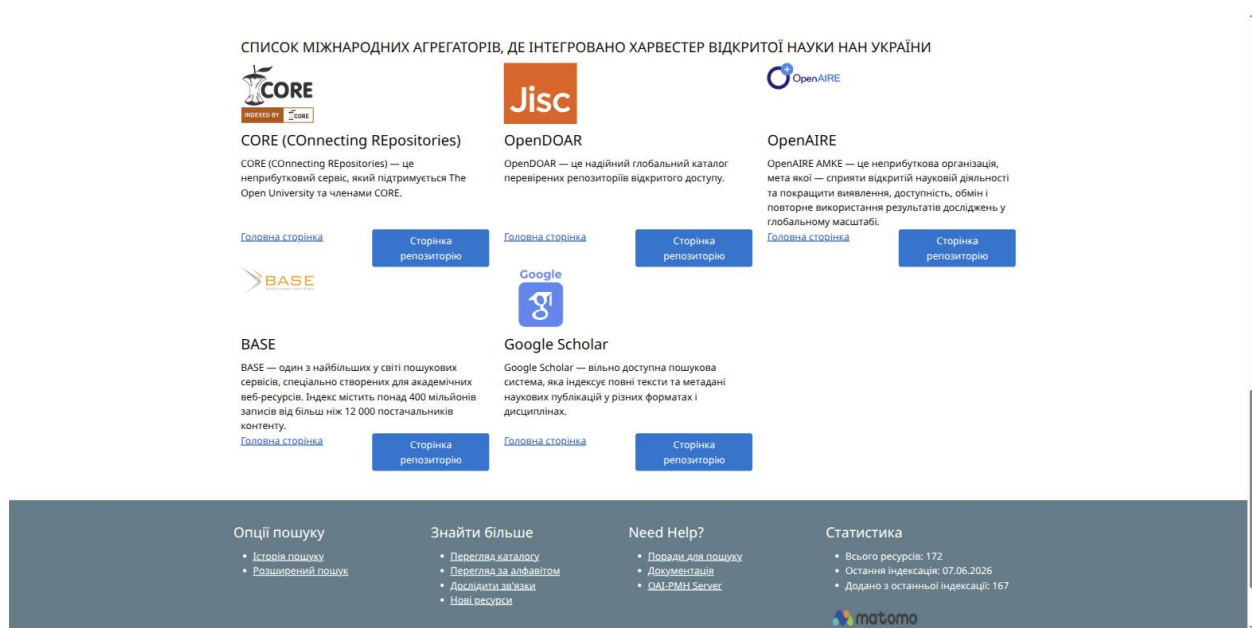
Головна сторінка — фасет «Мова» та графік активності авторів

Список мов: сортування за кількістю, а не за алфавітом

Фасет «Мова» на головній сторінці показував мови за алфавітом і без чисел. Тепер він відсортований за кількістю ресурсів (найпопулярніші мови — вгорі) і показує кількість поруч із кожною мовою, тож одразу видно, наскільки переважає українська.

Новий агрегатор і повернений текст на кнопках

До списку міжнародних агрегаторів додано Google Scholar. Заразом виправлено давню візуальну ваду: напис «Сторінка репозиторію» на кнопках не вміщався та обрізався — причина виявилась у глобальному CSS-правилі, яке фіксувало висоту всіх кнопок сайту під висоту поля пошуку. Кнопки тепер самі підбирають потрібну висоту під текст.



Список агрегаторів (з Google Scholar) та футер сайту

Футер: колір хедера й статистика індексації

Фон футера пофарбовано в той самий колір, що й хедер сайту, розтягнуто на всю ширину екрана (незалежно від ширини основного контенту), і додано короткий блок статистики: загальна кількість ресурсів, дата останньої індексації та скільки ресурсів додано з того часу.

Якість метаданих і повнотекстовий конвеєр

Значна частина роботи за ці два місяці — не в тому, що видно на сторінці, а в тому, наскільки коректні дані під нею: ключові слова, мова запису, автори.

Ключові слова: прибрано дублікати відмінків і сміттєві фрази

Індекс ключових слів для пошукового MCP-сервера (`keyword_index.pkl`) містив багато дублікатів — те саме поняття у формі однини й множини («потік» / «потоків»), або у родовому відмінку («композиція вебсервісу» / «композиція вебсервісів»). Провели кілька проходів очищення морфологічним аналізатором (`py morphology3`):

- Об'єднання форм однини/множини з пріоритетом на користь однини — з обережним винятком для слів, які насправді є окремими словами у множині (напр. «вибори» — «вибір» це не те саме).
- Об'єднання відмінкових варіантів того самого словосполучення.
- Видалення 1124 неусталених англомовних фраз, що випадково потрапили з анотацій статей (напр. «educational activity»), із залишенням визнаних термінів і скорочень (напр. «semantic web», «in vitro», «big data»).

Індекс скоротився з 47 949 до 45 330 ключових слів — без втрати реальних термінів.

Відоме обмеження

Для окремих статей полі `language` у Solr не збігається з мовою, задекларованою в XML-джерелі (наприклад, показує рідкісний код `ia` замість «English»). Причина: мова визначається не з XML, а автоматично за текстом самого PDF — бо OJS у `dc:language` перелічує всі мови анотації статті, а не мову конкретного файлу. Якщо витягнутий з PDF текст пошкоджений (частий випадок для статей з великою кількістю формул), детектор мови може помилитися. Виправлення (запобіжник, що повертається до значення з XML при низькій впевненості) поки не впроваджено.

Ключові слова тепер підбирає власний MCP-сервер

Конвеєр повнотекстового харвестингу (`FullTextExtractor`) під час індексації звертається до окремого сервера підбору ключових слів через протокол MCP (`llm/mcp-keywords-server/`), підключений у `fulltext.ini`) — той самий сервіс, який раніше розробляли й навчали окремо, тепер працює як частина продакшн-конвеєра харвестингу, а не лише як інструмент для локальної розробки.

Розділ 3

Новий спосіб підключити ШІ-агента до харвестера

Додано власний сервер за протоколом MCP (Model Context Protocol) — стандартом, яким користуються AI-асистенти (Claude та інші) для виклику інструментів у зовнішніх системах. На відміну від раніше створеного окремого Python-сервісу для підбору ключових слів, цей новий сервер написаний як звичайний модуль `VuFind` на PHP і працює прямо всередині сайту, використовуючи ті самі внутрішні сервіси пошуку, що й веб-інтерфейс.

Він дає ШІ-агенту два інструменти:

- `search_resources` — пошук ресурсів за текстовим запитом, з фільтрами за форматом і мовою.
- `get_resource` — повний опис одного ресурсу за ідентифікатором (назва, автори, рік, посилання на повний текст).

Підключення: `claude mcp add --transport http harvester https://harvester.nas.gov.ua/Mcp`

Розділ 4

Відстеження завантажень PDF

Клік по PDF-посиланню на сторінці запису тепер надсилає в Matomo дві події: іменовану подію «Download / PDF / назва статті» (для аналітики по конкретних статтях) і стандартний запис у звіт Matomo «Завантаження» — для випадків, коли посилання на файл не закінчується на `.pdf` і автоматичне відстеження Matomo не спрацювало б саме.

Записи: афіліації авторів

На сторінці запису під списком авторів тепер показується рядок «Автори та афіліації» — інституція кожного автора поруч з його іменем, замість лише переліку прізвищ.

The screenshot shows a journal article page with the title "Simulations of hypertrophied heart's hemodynamics". The article is published in "PROBLEMS IN PROGRAMMING" in 2018. The authors are R.D. Grygoryan and T.V. Aksenova, both from the Institute of Software Systems NAS of Ukraine. The page includes a sidebar with "Схожі ресурси" (Similar resources) and a "Додати тег" (Add tag) button.

Цитувати | Відправити е-поштою | Друк | Експортувати запис | Додати у Вибране | Постійне посилання

Simulations of hypertrophied heart's hemodynamics
The paper describes the modeling technology and the main results of the simulation of hemodynamic effects of cardiac hypertrophy (HH), conducted using previously published mathematical model (MM) [9]. The dynamics of hemodynamic abnormalities are not modeled. MM simulates changes in the central hemo...

[Повний опис](#)

Опубліковано в:	PROBLEMS IN PROGRAMMING
Дата:	2018
Випуск:	2-3
Сторінки:	254-263
ISSN:	1727-4907
Автори та афіліації:	R.D. Grygoryan — Institute of Software Systems NAS of Ukraine T.V. Aksenova — Institute of Software Systems NAS of Ukraine
Автори:	Grygoryan, R.D., Aksenova, T.V.
Формат:	Стаття
Мова:	Російська
Опубліковано:	PROBLEMS IN PROGRAMMING 2018
Теми:	mathematical model hemodynamics computer simulation UDC: 519.6+612
Онлайн доступ:	https://po.issofts.kiev.ua/index.php/ojs1/article/view/203
Теги:	Немає тегів. Будь-яке перше, хто поставити тег для цього запису! Додати тег
Назва журналу:	Problems in programming
Завантажити файл:	

Схожі ресурси

- A simulator of mechanisms of acute control of human hemodynamics за авторством: Grygoryan, R.D., та інші
Опубліковано: (2019)
- A simulator of mechanisms of long-term control of human hemodynamics за авторством: Grygoryan, R.D., та інші
Опубліковано: (2019)
- A software technology providing tuning procedures of a quantitative model of human hemodynamics за авторством: Grygoryan, R.D., та інші
Опубліковано: (2021)
- Specialized computer simulator "SimEnPhysiol" за авторством: Grygoryan, R.D., та інші
Опубліковано: (2018)
- Specialized software for simulating the multiple control and modulations of human hemodynamics за авторством: Grygoryan, R.D., та інші
Опубліковано: (2021)

Сторінка окремого запису — афіліації авторів та посилання на PDF

Заразом полагоджено

Блок «Про автора» з короткою довідкою з Вікіпедії (з'являється на сторінках пошуку за автором) перестав працювати, бо Вікіпедія прибрала підтримку старого формату відповіді API, яким користувався код. Перемкнули на сучасний формат — блок знову показує довідку.

Обмін даними з іншими системами: OAI-PMN тепер передає журнал і сторінки

Харвестер сам віддає свої записи іншим системам (OpenAIRE, BASE тощо) через протокол OAI-PMH. У метаданих, які він для цього формує, тепер є не лише сама стаття, а й пов'язаний з нею запис журналу — назва, ISSN, том, випуск, перша й остання сторінки. Раніше зовнішні системи бачили статтю окремо від журналу, в якому вона опублікована; тепер зв'язок явний.

Розділ 7

Нові джерела й інфраструктура

Нове джерело: arheologia.com.ua

Підключено харвестинг журналу arheologia.com.ua через OAI-PMH. Заразом, під час цієї ж роботи, полагоджено два вже наявні джерела, які тихо не працювали: застаріле посилання для UkrainianJournalofPhysics та помилку в регулярному виразі, за яким для www-cpts-com-ua розпізнавався ідентифікатор запису.

Продакшн-сервіс для підбору ключових слів

MCP-сервер підбору ключових слів (розділ 2) отримав власний опис сервісу в продакшн-конфігурації Docker ([docker-compose.prod.yml](#)) — тепер він піднімається як окремий контейнер на продакшн-сервері так само надійно, як і решта сервісів харвестера.

Прибирання старих напрацювань

Каталог [llm/](#) накопичив багато проміжних експериментів під час розробки підбору ключових слів — чорнові скрипти, датасети, ноутбуки в декількох ітераціях каталогів ([1Keyword_extraction/](#), [delete/](#) тощо). Прибрано близько 29 700 рядків такого коду; робоча функціональність лишилась зібраною у двох місцях — [llm/mcp-keywords-server/](#) (продакшн-сервер) і [llm/4final/](#) (пайплайн навчання моделі).