УДК 004.82 https://doi.org/10.15407/pp2025.01.038

Г.Ю. Проскудіна, К.О. Кудім

ОГЛЯД ГЛОБАЛЬНИХ СЛУЖБ АГРЕГАЦІЇ РЕСУРСІВ ВІДКРИТОГО ДОСТУПУ ТА ЇХНІХ ВИМОГ ДО ПОСТАЧАЛЬНИКІВ ДАНИХ

У роботі представлено огляд сучасних глобальних агрегаторів документів відкритого доступу. Проаналізовані їхні кількісні характеристики, такі як кількість зібраних описів документів та повних текстів, кількість постачальників даних, наявність інтерфейсу прикладного програмування для отримання даних. Проаналізовано склад і види їхніх постачальників даних, такі як інституційні репозитарії, відкриті журнали, видавництва, наукові репозитарії препринтів, тематичні електронні бібліотеки, а також системи, які в свою чергу теж є агрегаторами. Досліджено також яку саме інформацію про документи збирають ці агрегатори, як вона представлена в інтерфейсі користувача, а також яка інформація збирається про їхніх постачальників даних та яким чином вона представлена у інтерфейсі користувача. Як саме відбувається взаємодія агрегатора з постачальниками даних, які протоколи обміну даних підтримуються, з якою частотою відбувається оновлення зібраних даних. Також сучасні агрегатори на базі зібраних корпусів даних, використовуючи методи машинного навчання, методи бібліометрії, вебометрики, альтиметрії, семантометрії надають науковцям цілий ряд корисних сервісів. Ми як розробники низки наукових електронних бібліотек з відкритим доступом вже зареєстровані як провайдери даних у деяких з цих систем. Тому знайомі з їхніми вимогами у практичній площині. В цій роботі ми спробували дещо узагальнити ці вимоги.
Ключові слова: відкритий доступ, провайдер сервісів, провайдер даних, протокол OAI-PMH.

G.Yu. Proskudina, K.O. Kudim

OVERVIEW OF GLOBAL OPEN ACCESS RESOURCE AGGREGATION SERVICES AND THEIR REQUIREMENTS FOR DATA PROVIDERES

The paper presents an overview of modern global aggregators of open access documents. Their statistical characteristics are analysed, such as the number of collected document descriptions and full texts, the number of data providers, and the availability of application programming interface to obtain data. The types of data providers, such as institutional repositories, open journals, publishers, scientific repositories of preprints, thematic digital libraries, and systems that are also aggregators, are analysed. We also investigate what kind of information about documents these aggregators collect and how it is presented in the user interface, as well as what information is collected about their data providers and how it is presented in the user interface. How the aggregator interacts with data providers, what data exchange protocols are supported, and how often the collected data is updated. Also, modern aggregators based on collected data corpora, using machine learning methods, bibliometrics, webometrics, altmetrics, semantometrics, provide a range of useful services to researchers. As developers of a certain number of scientific digital libraries, we are already registered as data providers in some of these systems. Therefore, we are familiar with their requirements in the practical sense. In this paper, we have attempted to summarise these requirements.
Key words: open access, service provider, data provider, OAI-PMH protocol.

Вступ

В ході виконання частини проєкту НАНУ «Відкрита наука» були проведені дослідження із створення системи інтеграції (харвестера, агрегатора) ресурсів із різних відкритих академічних джерел з метою отримання зручного і потужного інструменту пошуку та доступу до інформації для користувачів. Також проводяться дослідження з метою інтеграції цих зібраних наукових ресурсів України у світовий/європейський науковий інформаційний простір.

Зазвичай, перед системою інтеграції постають три основні задачі:

1. Збір та інтеграція описової інформації (метаданих) з різних джерел електронних ресурсів;

2. Організація пошуку і видачі відповідних ресурсів;

3. Передача метаданих з власного харвестера іншим харвестерам.

Проаналізувавши низку програмних систем, на платформі яких можна розгорнути агрегатор наукових робіт, ми зупинилися на програмній системі VuFind, що була розроблена в університеті Вілланова, США, про яку ми розповіли в роботі [1].

Наразі ми вже розгорнули таку систему[1] і почали її експлуатацію. До неї підключено близько десяти електронних бібліотек (провайдерів даних), з яких було зібрано близько 300 тис метаданих документів, в основному це – статті. Ми індексуємо метадані документів усіх видів академічно релевантних ресурсів – таких як журнали, інституційні репозитарії, цифрові колекції тощо, які надають інтерфейс OAI та використовують протокол Open Archives Initiative Protocol for Metadata Harvesting[2] (OAI-PMH) для надання свого контенту [2]. Зібрані і проіндексовані дані зберігаються на серверах Інституту програмних систем НАНУ. Апробується підключення та передача метаданих з нашої системи до інших харвестерів. Вивчаються можливості системи VuFind із перегляду, пошуку та завантаження статей, а також відпрацьовуються інструкції для кінцевого користувача та постачальників даних, визначаються схеми метаданих основних видів наукових інформаційних ресурсів НАНУ.

Зараз продовжується робота із налагодження цієї системи інтеграції, відпрацьовуються алгоритми завантаження та оновлення даних, регламент роботи агрегатора, вимоги до провайдерів даних щодо підключення до нашого харвестра.

Тому вивчення світового досвіду з експлуатації таких агрегаторів ресурсів є необхідним кроком у доведенні нашої системи до сучасних зразків, аби зробити максимально доступною для суспільства наукову інформацію, що сприятиме розвитку освітньої, наукової, науково-технічної та інноваційної діяльності.

У першій частині ми даємо основні роз'яснення щодо значення термінів, які використовуються в наших документах.

У другій частині представлено огляд сучасних глобальних агрегаторів ресурсів відкритого доступу. Проаналізовані їхні кількісні характеристики, такі як кількість зібраних описів документів та повних текстів, кількість постачальників даних, наявність API отримання даних. Проаналізовано склад і види постачальників даних, такі як інституційні репозитарії, відкриті журнали, видавництва, наукові репозитарії препринтів, тематичні електронні бібліотеки, а також системи, які в свою чергу теж є агрегаторами. Досліджено також, яку саме інформацію про документи збирають ці агрегатори, як вона представлена в інтерфейсі користувача, а також яка інформація збирається про постачальників даних, а також, як вона представлена у інтерфейсі користувача. Яким чином відбувається взаємодія агрегатора з постачальниками даних, які протоколи обміну даних підтримуються, з якою частотою відбувається оновлення зібраних даних. Також сучасні агрегатори на базі зібраних корпусів даних, використовуючи методи машинного навчання, методи бібліометрії, вебометрики, альтиметрії, семантометрії надають цілий ряд корисних сервісів науковцям. Ми як розробники низки наукових електронних бібліотек з відкритим доступом вже зареєстровані як провайдери даних у деяких із цих систем. Тому знайомі з їхніми вимогами, як провайдерів сервісів до своїх провайдерів даних у практичній площині.

У третьому розділі ми спробували дещо узагальнити ці вимоги.

1. Термінологія

Цей розділ містить роз'яснення щодо значення термінів, які використовуються в нашій роботі.

Відкритий доступ (ВД). Будапештська ініціатива відкритого доступу (Budapest

1 https://harvester.nas.gov.ua/

2 https://www.openarchives.org/OAI/openarchivesprotocol.html

Open Access Initiative, BOAI), яке визначає ВД як «вільну доступність у загальнодоступному інтернеті, що дозволяє будь-яким користувачам читати, завантажувати, копіювати, поширювати, друкувати, шукати або посилатися на повні тексти цих статей, сканувати їх для індексації, передавати їх як дані програмному забезпеченню або використовувати їх для будь-яких інших законних цілей[3]».

Дублінське ядро. Схема метаданих для опису інформаційних ресурсів репозиторіїв провайдерів даних.

Електронна бібліотека (ЕБ). Це розподілена документальна інформаційно-пошукова система, що функціонує на основі повнотекстових репозиторіїв (баз даних) і надає можливість створювати, зберігати та використовувати різноманітні колекції електронних документів інформаційних ресурсів (документів) у глобальній мережі комп'ютерів у зручному для користувачів вигляді.

Запис. Сукупність метаданих, що описують окремий науковий ресурс, розташований у репозитарії. Цей термін зазвичай використовується для позначення опису цифрового об'єкта, такого як текст, зображення, відео тощо. У харвестері термін **запис метаданих** використовується для позначення метаданих наукової публікації, тобто її назву, авторів, анотацію, деталі фінансування проєкту тощо, і термін **повнотекстовий запис** для позначення запису, що містить посилання у метаданих на повний текст інформаційного ресурсу. У харвестері запис є інформаційним ресурсом.

Інформаційна (автоматизована) система. Організаційно-технічна система, в якій реалізується технологія зберігання та обробки інформації з використанням технічних і програмних засобів.

Інформаційний ресурс (ресурс). Електронний документ або їх сукупність у автоматизованих інформаційних системах (ЕБ, архівах, базах даних тощо).

Користувач. Фізична особа, якій надається відкритий доступ до пошуку та перегляду інформаційних ресурсів харвестера, що мають відношення до записів метаданих та похідної від них інформації (наприклад, статистична інформація), та виконання переходу на домашню сторінку знайденого документа у репозитарії провайдера даних.

Персональний електронний кабінет. Індивідуальна персоніфікована веб-сторінка, за допомогою якої користувач здійснює роботу з інформаційними ресурсами, представленими у харвестері.

Провайдер/постачальник даних. Це служба, що підтримує створення і ведення одного чи більше репозиторіїв (бази документів, архівів, електронних бібліотек), здійснює публікацію своїх ресурсів, а також надає доступ до своїх метаданих для їхнього використання в інших системах. Зазвичай, провайдер даних надає вільний доступ до своїх метаданих і, можливо, але не обов'язково, надає вільний доступ до повних текстів своїх документів з ЕБ чи з інших інформаційних ресурсів. Провайдер даних може мати самостійний веб-інтерфейс для організації пошуку, перегляду і доступу до своїх ресурсів, а також інші сервіси, що надаються кінцевим користувачам. Провайдер даних самостійно вирішує питання про відкритість своїх інформаційних ресурсів і доступність до них. Зокрема, провайдер даних може прийняти рішення про інтеграцію усіх або частини своїх інформаційних ресурсів на рівні метаданих у харвестері і для цього організує експорт відповідних метаданих у форматі узгодженого протоколу.

Протокол OAI-PMH. Протокол OAI-PMH збору (харвестінгу) визначає механізм збору записів, що містять метадані інформаційних ресурсів, з репозитаріїв провайдерів даних. Він надає простий спосіб такого представлення метаданих, яке робить їх доступними для систем-агрегаторів інформаційних ресурсів. Зібрані в такий спосіб метадані можуть бути представлені в будь-якому форматі, обраному співтовариством організацій, що вирішили об'єднати свої зусилля для створення інтегрованої федеративної інформаційної системи.

3 https://www.budapestopenaccessinitiative.org

Репозитарій. Електронний архів або сховище для довготривалого, постійного та надійного накопичення, зберігання, управління та розповсюдження інформаційних ресурсів. **Інституційний репозитарій** має відношення до наукових інформаційних ресурсів, отриманих у результаті досліджень певної наукової установи. В **тематичних репозитаріях** інформаційні ресурси фокусуються на певних галузях знань. **Типи репозитаріїв** визначаються типами інформаційних ресурсів (журнальні статті, препринти, тези, книги, монографії, дослідницькі дані, зображення, карти, аудіо, відео і т.ін.). **Централізовані репозитарії** уможливлюють інтегроване зберігання інформаційних ресурсів різних типів і галузей знання у великому обсязі. **Спільнотні репозитарії** створюються спільнотами дослідників, учених або наукових установ з метою обміну даними та співпраці у конкретній галузі.

Цільова сторінка. Сторінка HTML у харвестері, куди зазвичай потрапляють під час доступу до наукової публікації, та з якої можна отримати доступ до пов'язаних ресурсів.

Харвестер. Програмний інструмент для автоматичного збору метаданих наукових періодичних видань НАН України та інформаційних ресурсів установ НАН України, редагування метаданих у харвестері, передачі метаданих у інші харвестери тощо.

2. Чинні служби агрегації відкритого доступу та бази даних публікацій

Наразі існує низка доступних агрегаційних служб відкритого доступу (Табл. 1), прикладами яких є BASE[4], OpenAIRE[5], CORE[6], Unpaywall[7], Paperity[8], SHARE[9].

BASE (Bielfield Academic Search Engine) – глобальна служба збору метаданих. Вона збирає репозитарії та журнали

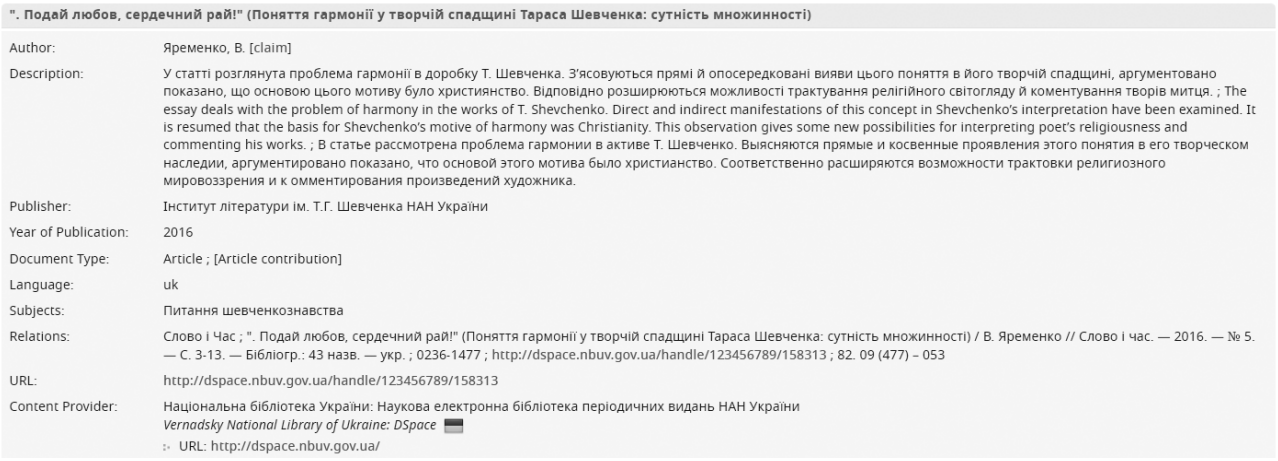
Таблиця 1. Порівняння служб агрегації відкритого доступу

Назва	Записи метаданих [/постачальники даних]	Записи з повним текстом	Записи з повнотекстовим посиланням	API	Набір даних	Синхронізація даних
CORE	298 млн / 11 139	32.8 млн	139 млн	Так	Так	Так■
BASE	362 млн / 11 399	0	~60%	Так	Ні	Ні
OpenAIRE	149 млн	19 млн*	н/в	Ні	Ні	Ні
Unpaywall	50 млн / 50 тис (+ через Crossref і DOAJ)	0	50 млн	Так	Так♦	Так●
Paperity	10.5 млн / 17 158 журналів	10.5 млн	н/в	Ні	Ні	Ні
SHARE	57 млн	0	н/в	Ні	Ні	Ні

Примітка: * Повні тексти, розміщені на OpenAIRE, недоступні для завантаження. ♦ Набір даних, наданий Unpaywall, містить лише посилання, а не повні метадані чи повний текст, як у випадку з CORE. ■ CORE використовує механізм FastSync для синхронізації даних. ● Unpaywall забезпечує синхронізацію даних як частину преміум-сервісу.

4 https://base-search.net/
5 https://www.openaire.eu/
6 https://core.ac.uk/
7 http://unpaywall.org/
8 https://paperity.org/
9 https://share.osf.io/

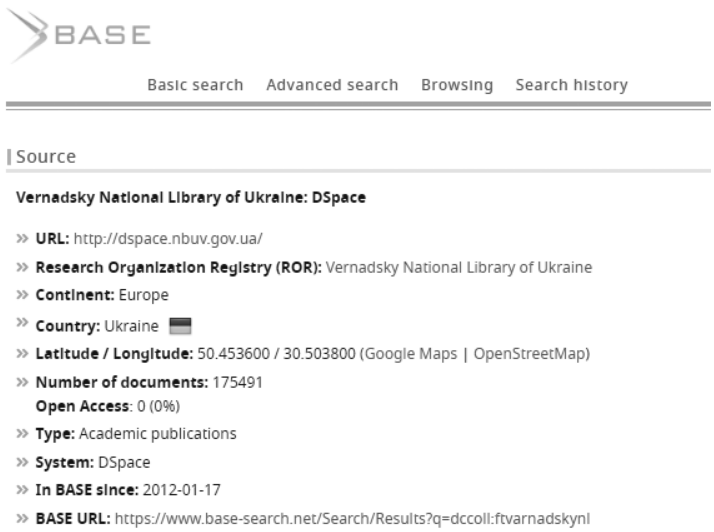
по протоколу OAI-PMH і надає зібраний вміст через API та набір даних. Наукова електронна бібліотека періодичних видань НАН України, https://dspace.nbuv.gov.ua, яку ми підтримуємо, підключена до цього агрегатора у якості провайдера даних. На рис. 1–2 наведені приклади, яким чином тут представлені статті та сам провайдер даних (https://www.base-search.net/about/en/about_source.php?id=2328).



". Подай любов, сердечний рай!" (Поняття гармонії у творчій спадщині Тараса Шевченка: сутність множинності)

Author:	Яременко, В. [claim]
Description:	У статті розглянута проблема гармонії в доробку Т. Шевченка. З'ясовуються прямі й опосередковані вияви цього поняття в його творчій спадщині, аргументовано показано, що основою цього мотиву було християнство. Відповідно розширюються можливості трактування релігійного світогляду й коментування творів митця. ; The essay deals with the problem of harmony in the works of T. Shevchenko. Direct and indirect manifestations of this concept in Shevchenko's interpretation have been examined. It is resumed that the basis for Shevchenko's motive of harmony was Christianity. This observation gives some new possibilities for interpreting poet's religiousness and commenting his works. ; В статье рассмотрена проблема гармонии в активе Т. Шевченко. Выясняются прямые и косвенные проявления этого понятия в его творческом наследии, аргументировано показано, что основой этого мотива было христианство. Соответственно расширяются возможности трактовки религиозного мировоззрения и к омментирования произведений художника.
Publisher:	Інститут літератури ім. Т.Г. Шевченка НАН України
Year of Publication:	2016
Document Type:	Article ; [Article contribution]
Language:	uk
Subjects:	Питання шевченкознавства
Relations:	Слово і Час ; ". Подай любов, сердечний рай!" (Поняття гармонії у творчій спадщині Тараса Шевченка: сутність множинності) / В. Яременко // Слово і час. — 2016. — № 5. — С. 3-13. — Бібліогр.: 43 назв. — укр. ; 0236-1477 ; http://dspace.nbuv.gov.ua/handle/123456789/158313 ; 82. 09 (477) – 053
URL:	http://dspace.nbuv.gov.ua/handle/123456789/158313
Content Provider:	Національна бібліотека України: Наукова електронна бібліотека періодичних видань НАН України *Vernadsky National Library of Ukraine: DSpace* URL: http://dspace.nbuv.gov.ua/

Рис. 1. Приклад представлення опису документу у агрегаторі BASE



BASE

Basic search Advanced search Browsing Search history

Source

Vernadsky National Library of Ukraine: DSpace

- **URL:** http://dspace.nbuv.gov.ua/
- **Research Organization Registry (ROR):** Vernadsky National Library of Ukraine
- **Continent:** Europe
- **Country:** Ukraine
- **Latitude / Longitude:** 50.453600 / 30.503800 (Google Maps | OpenStreetMap)
- **Number of documents:** 175491
  Open Access: 0 (0%)
- **Type:** Academic publications
- **System:** DSpace
- **In BASE since:** 2012-01-17
- **BASE URL:** https://www.base-search.net/Search/Results?q=dccoll:ftvarnadskynl

Рис. 2. Приклад опису провайдера даних у агрегаторі BASE

OpenAIRE – це мережа постачальників даних відкритого доступу, які підтримують політику відкритого доступу. Незважаючи на те, що в минулому проєкт був зосереджений на європейських репозитаріях, нещодавно він розширився за рахунок інституційних і предметних репозитаріїв поза межами Європи. Основним напрямком OpenAIRE є допомога Європейській Раді в моніторингу дотримання її політики відкритого доступу. Дані OpenAIRE доступні через API. Наукова електронна бібліотека періодичних видань зареєстрована у цій системі з 2018 року з рівнем сумісності v2.0 з інструкціями OpenAIRE (рис. 3). На даний час є питання щодо оновлення нашої системи до рівня сумісності v3.0 або v4.0. Оскільки нова інтегрована платформа каталогу EOSC Portal Catalog та Marketplace тепер обслуговуватиме лише зареєстровані джерела даних OpenAIRE, які сумісні з версіями 3.0 та 4.0 інструкцій OpenAIRE. Це дасть нам змогу отримувати додаткові послуги, наприклад, буде можливість отримати звіти про використання досліджень. OpenAIRE збирає дані про використання, а потім об'єднує їх і надає стандартизовані звіти.

Рис. 3. Наукова електронна бібліотека періодичних видань у агрегаторі OpenAIRE

Paperity – перший мультидисциплінар-ний агрегатор журналів і статей відкритого доступу, який був запущений 2014 року. Paperity збирає як метадані, так і повні тексти. Наразі агрегатор містить 10 508 682 повнотекстових статей та 17 158 журналів (рис. 4) в усіх галузях досліджень: від науки, технологій, медицини до соціальних наук, гуманітарних наук і мистецтва. Метою Paperity є надання читачам легкого і необмеженого доступу до тисяч журналів із сотень дисциплін в одному центральному місці; допомога авторам охопити цільову аудиторію, ефективніше поширювати відкриття та максимізувати вплив досліджень; підвищення популярності журналів, збільшення їх читацької аудиторії та заохочення подання нових рукописів. Paperity має наступні особливості:

Доступність. Усі статті, які проіндексовані Paperity, мають відкритий доступ і доступні з повним текстом. Це свідомий вибір. Вчені втомилися від платних платформ і вимагають легкого доступу до потрібної їм літератури, особливо якщо ця література створена ними самими. Наукова комунікація має бути відкритою, адже відкритість – це не про вартість, а про свободу наукових досліджень.

Уніфікація. Сьогодні науковці потребують широкого доступу до літератури з різних галузей, навіть з-поза меж їхньої основної дослідницької сфери. Сучасні дослідження стали міждисциплінарними і найбільш новаторські відкриття відбуваються на перетині різних дисциплін. Академічні служби мають наздоганяти цей процес. Paperity – це шлях до ефективнішої наукової комунікації.

Цілісність та висока якість корпусу. Paperity гарантує, що індексується дійно наукова література. Завдяки унікальній тех-

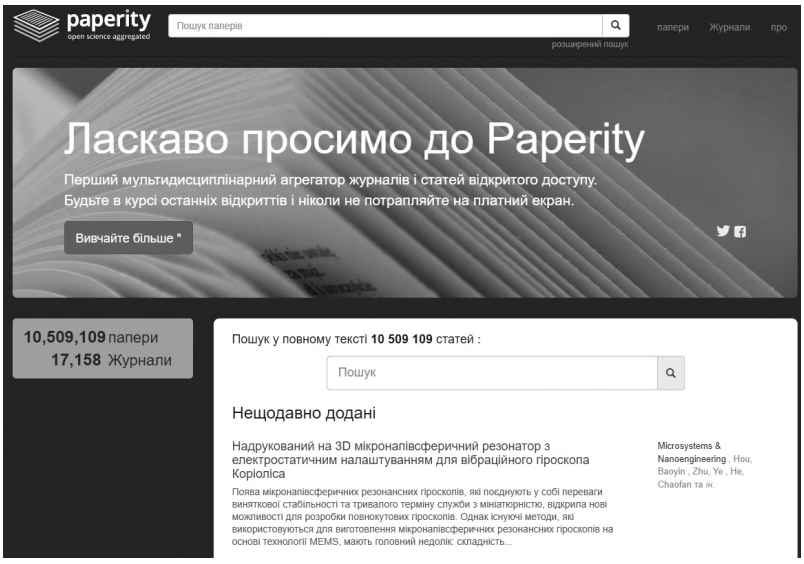


Рис. 4. Домашня сторінка агрегатора Paperity

нології тут збирають детальні метадані про кожну публікацію і пильно стежать за тим, щоб не забруднювати колекцію нерелевантними записами. У Paperity ніколи не знайти студентських завдань, бізнес-презентацій або кулінарних рецептів, класифікованих як наукові роботи.

Unpaywall – це безкоштовна база даних наукових статей відкритого доступу з API та розширенням для браузера. Розроблена некомерційною організацією Impactstory[10], яка опікується проблемами відкритого доступу в науці. Тобто це не агрегатор, а скоріше збір вмісту з Crossref щоразу, коли безкоштовна версія доступна для читання. Обробляє як метадані, так і повний текст, але не розміщує їх. Розкриває виявлені посилання на документи через API. Значну частину даних інструмент отримує з бази даних під назвою oaDOI, яка індексує більше сотні мільйонів документів, яким присвоєно цифрові ідентифікатори об'єктів (DOI).

Unpaywall – це також і плагін для веб-браузера (наприклад, Firefox або Chrome), який визначає потрібний вам документ, а потім перевіряє, чи доступний він безкоштовно будь-де в Інтернеті. І коли ви

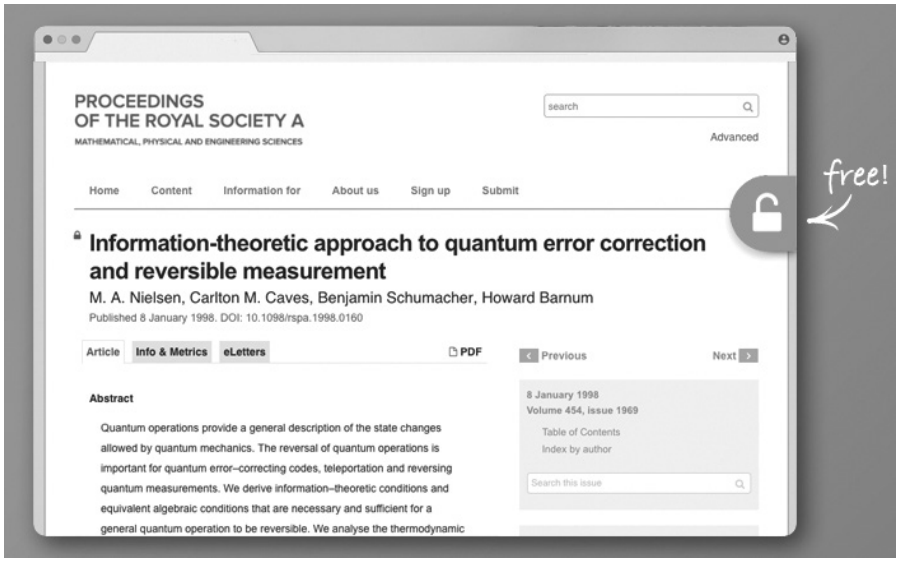


Рис. 5. Робота розширення Unpaywall

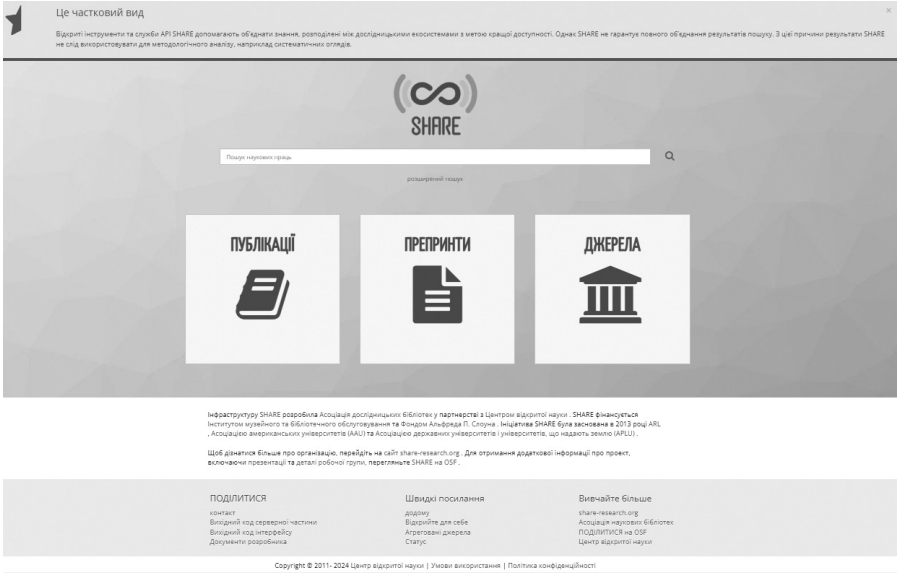


Рис. 6. Дослідницька екосистема спільного доступу SHARE

[10] https://impactstory.org/about

перейдете на сторінку, яка підсумовує чи показує частину статті, з'явиться маленький значок замка (рис. 5), який повідомляє, чи можливо його отримати десь ще безкоштовно. Якщо стаття доступна, сірий значок «замок» Unpaywall стане зеленим і «розблокується». Клацнувши на нього, користувач отримує доступ до PDF-файлів, яке спрощує пошук PDF-файлів із відкритим доступом.

SHARE (SHared Access Research Ecosystem), дослідницька екосистема спільного доступу – це збирач вмісту відкритого доступу з репозиторіїв США, https://aims.fao.org/news/share-making-research-widely-accessible-discoverable-and-reusable. Незважаючи на те, що SHARE збирає як метадані, так і повний текст, він не розміщує останній.

CORE (COnnecting REpositories) об'єднує наукові статті у відкритому доступі від тисяч постачальників даних з усього світу, включно з інституційними та предметними репозитаріями, журналами відкритого та гібридного доступу. CORE є найбільшою колекцією літератури з відкритого доступу – на момент написання цієї статті вона забезпечує єдину точку доступу до наукової літератури, зібраної від понад десяти тисяч постачальників даних з усього світу, і ця колекція постійно зростає. Вона надає кілька способів доступу до своїх даних як для користувачів, так і для машин, включно із безкоштовним API і повним дампом своїх даних. На наш погляд, CORE – найцікавіший агрегатор з точки зору розвитку нашої системи у майбутньому. Наразі ми маємо дві свої бібліотеки, зареєстровані у CORE (рис. 7–9). Це вищезгадана бібліотека періодичних видань НАНУ, що працює на програмному забезпеченні DSpace[11], і електронна бібліотека Інституту програмних систем НАНУ, що працює на програмному забезпеченні EPrints[12].

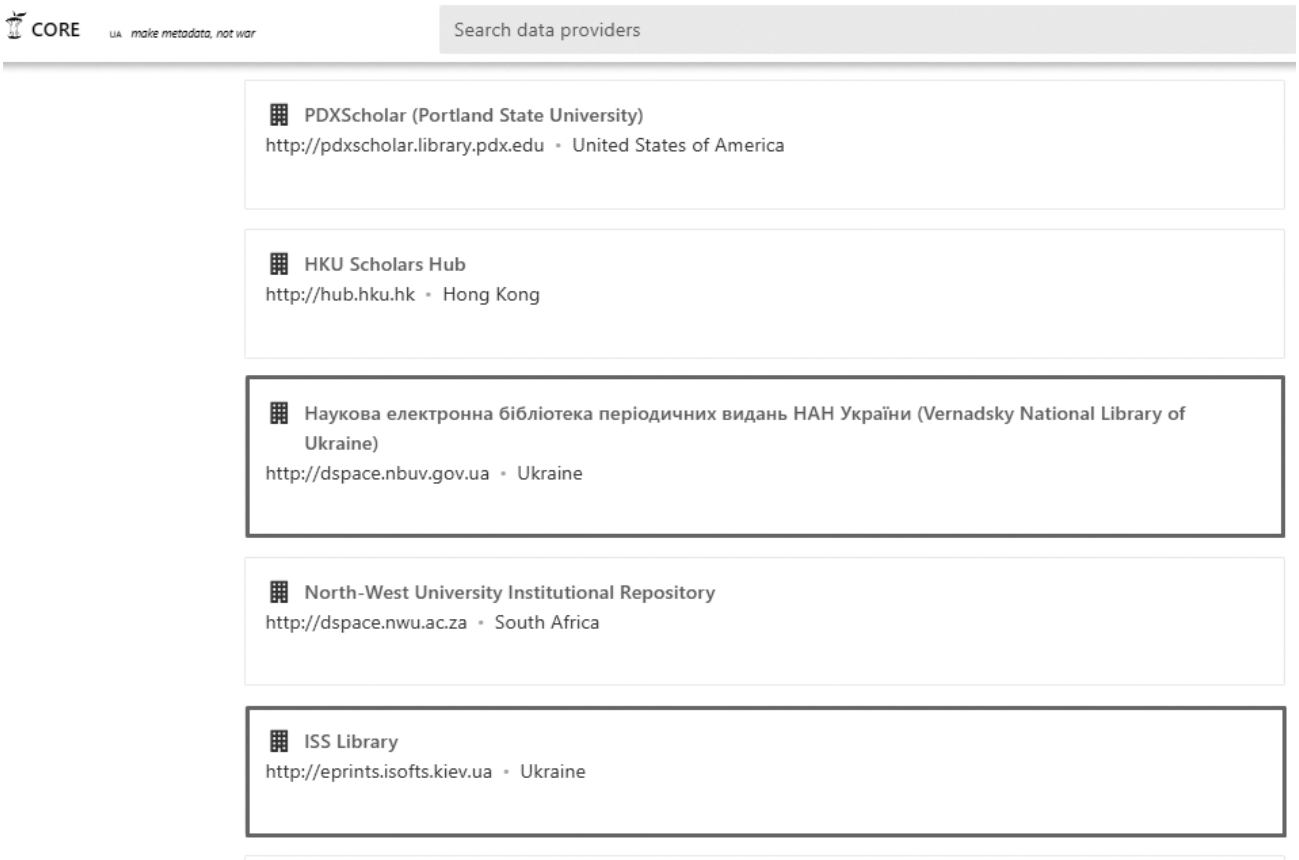


Рис. 7. Наші бібліотеки у списку постачальників даних у агрегаторі CORE

11 https://dspace.lyrasis.org/
12 https://www.eprints.org/uk/

Рис. 8. Наукова електронна бібліотека періодичних видань НАН України у агрегаторі CORE[13]

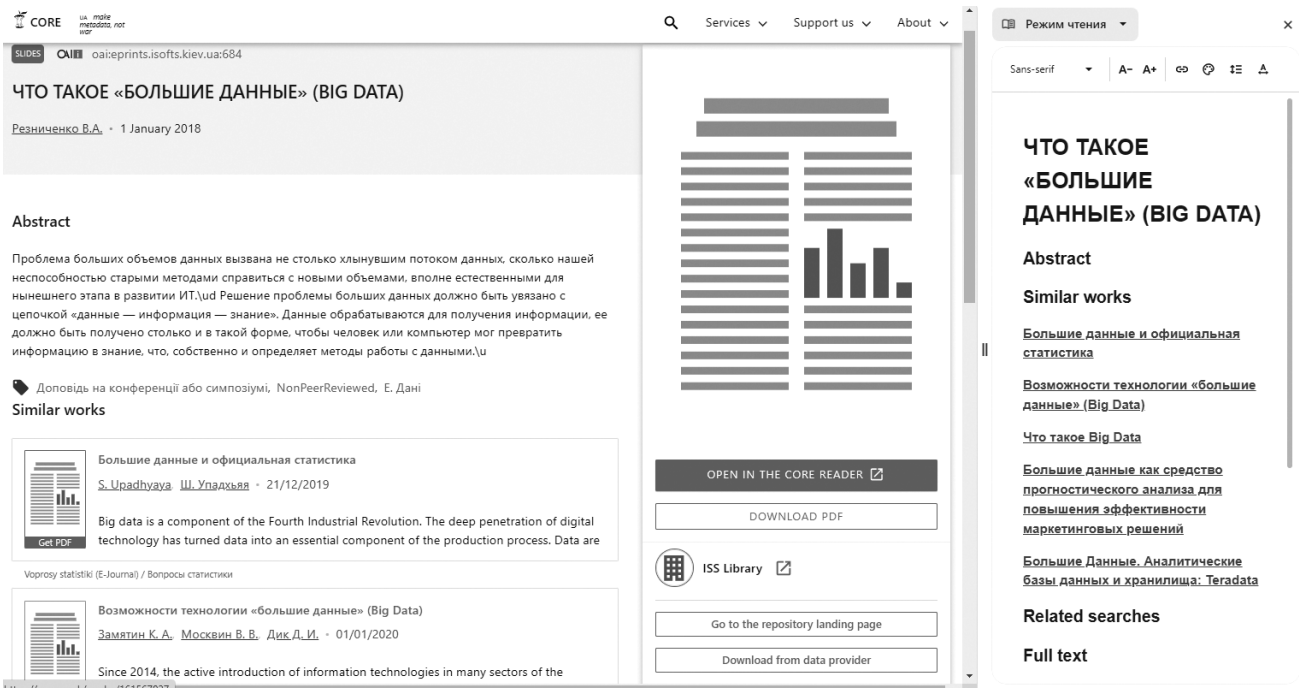


Рис. 9. Сторінка представлення повного тексту документу з ISS Library у агрегаторі CORE

13 https://core.ac.uk/data-providers/944

Джерела даних CORE. Станом на квітень 2024 року CORE агрегував контент із 11 139 джерел даних. Ці джерела даних включають інституційні репозитарії, академічні видавництва (Elsevier, Springer), журнали відкритого доступу, предметні репозиторії, в тому числі ті, що містять електронні версії (arXiv, ZENODO, PubMed Central) та агрегатори (наприклад, DOAJ). Декілька найбільших джерел даних CORE наведено в Таблиці 2. При підрахунку загальної кількості постачальників даних у CORE агрегатори і видавці розглядаються як одне джерело даних, не зважаючи на те, що деякі з них, в свою чергу, агрегують дані з багатьох джерел. Повний список усіх постачальників даних можна знайти на веб-сайті CORE, https://core.ac.uk/data-providers.

Таблиця 2. Найбільші постачальники даних у CORE

Назва, url	Кількість документів / повних текстів
Crossref, https://crossref.org/	118.155.624 / 1.532.012
CiteSeerX, http://citeseerx.ist.psu.edu	6.540.649 / 80.766
Directory of Open Access Journals, https://doaj.org/	5.463.188 / 990.721
Zenodo, https://zenodo.org	3.321.673 / 51,927
Elsevier - Publisher Connector	1.682.191 / 840.998

Примітка. Агрегатори вмісту, такі як DOAJ і Elsevier, представлені як одне джерело даних, незважаючи на те, що вони самі є агрегаторами і збирають дані з багатьох джерел.

Crossref. Як видно з Таблиці 2, серед провайдерів даних CORE, однією з основних баз даних публікацій є Crossref, авторитетний покажчик ідентифікаторів DOI (The Digital Object Identifier).

DOI – це обов'язковий міжнародний цифровий ідентифікатор наукової публікації. DOI визначає постійне місце знаходження наукової роботи (об'єкта) в інтернеті, її назву та метадані[14]. DOI на сьогодні є обов'язковою складовою сучасної системи наукової комунікації. Він полегшує процедуру та облік цитування, пошуку та локалізації наукової публікації. Ви можете присвоїти DOI будь-якій публікації, наприклад: науковій статті, монографії, главі монографії, дисертації, автореферату, рецензії, підручнику, методичним розробкам/рекомендаціям, звіту, препринту і навіть окремо таблиці, рисунку, схемі, зображенню тощо. DOI призначається публікації раз і назавжди. Це забезпечує стабільність посилання на публікацію в Інтернеті та спрощує пошук потрібної інформації.

Отже, основною функцією бази даних публікацій Crossref є збереження метаданих, пов'язаних з кожним DOI. Метадані, які зберігає Crossref, включають записи як ВД, так і не ВД. Crossref не зберігає повний текст публікації, але для багатьох публікацій надає посилання на повні тексти. Хоча Crossref надає API, він не пропонує свої дані для масового завантаження та не надає служби синхронізації даних.

Zenodo – це безкоштовний і відкритий цифровий архів, створений CERN і OpenAIRE, який дає змогу дослідникам ділитися і зберігати свої наукові результати в будь-якому обсязі, форматі та з усіх галузей досліджень. На Zenodo можна знайти різноманітні типи даних:

1. Датасети: набори даних, які дослідники можуть завантажувати та ділитися зі спільнотою.
2. Публікації: наукові статті, дисертації, звіти та інші публікації.
3. Програмне забезпечення: відкритий вихідний код, бібліотеки, інструменти та програми.

[14] https://sciencen.org/o/doi/

4. Презентації: презентації, семінари та доповіді.

Якщо планується використовувати Zenodo, слід ознайомитися з розділом Get started, щоб дізнатися, як створити обліковий запис, завантажити файли та орієнтуватися в інтерфейсі.

Інші бази публікацій. Окрім сервісів агрегації ВД, існує низка інших сервісів для пошуку та завантаження наукової літератури, Таблиця 3.

Решту послуг із Таблиці 3 можна згрупувати приблизно в такі дві категорії: 1) індекси цитування, 2) академічні пошукові

Таблиця 3. Порівняння баз публікацій

Назва	Вільний доступ	Розміщується повний текст	API	Набір даних	Синхронізація даних
CORE	Так	Так	Так	Так	Так
Crossref	Так	Ні	Так	Ні	Ні
Scopus	Ні	Так	Так	Ні	Ні
Web of Science	Ні	Так	Так	Ні	Ні
Google Scholar	н/в*	Ні	Ні	Ні	Ні
Semantic Scholar	Так	Так	Так	Так	Ні
Dimensions	Так	Так	Так	Ні	Ні
1findr	Ні	Так	Ні	Ні	Ні

Примітка. Під «вільним доступом» тут розуміють, чи є доступ до бази даних вільним за допомогою автоматизованих методів (наприклад, через API). *Google Scholar не надає засобів програмного доступу до своїх даних.

системи та наукові графи. Двома основними індексами цитування є Scopus від Elsevier[15] і Web of Science від Clarivate[16], які є послугами передплати преміум-класу. Google Scholar, найвідоміша академічна пошукова система, не надає API для доступу до своїх даних і не дозволяє сканувати свій веб-сайт. Semantic Scholar[17] є відносно новою академічною пошуковою службою, метою якої є створення «інтелектуальної академічної пошукової системи» [4]. Dimensions[18] – це сервіс, орієнтований на аналіз даних. Він об'єднує публікації, гранти, політичні документи та показники. 1findr[19] – це служба індексування рефератів. Він надає посилання на повний текст, але не містить API чи набору даних для завантаження.

Процес збору документів у агрегаторі CORE. Процес збору можна описати як послідовність етапів, на кожному етапі виконується певна дія, і коли вихідні дані одного етапу передаються на вхід наступному [3]. Вхідними даними для цього процесу є набір постачальників даних, а кінцевим виходом є система, заповнена записами дослідницьких робіт. Основні типи завдань, які наразі виконуються в рамках системи збору документів, наступні:

Завантаження метаданих: метадані, надані постачальником даних за протоколом OAI-PMH, завантажуються та зберігаються у файловій системі (зазвичай як XML-файли). Процес завантаження є послідовним, тобто репозитарій зазвичай надає від 100 до 1000 записів метаданих на запит і маркер продовження. Потім цей маркер використовується для надання наступної партії записів. У результаті повне збирання може займати чимало часу (годин-днів) для великих постачальників даних. Такий процес

15 https://www.elsevier.com/solutions/scopus
16 https://clarivate.com/webofsciencegroup/solutions/web-of-science/
17 https://www.semanticscholar.org/
18 https://www.dimensions.ai/
19 https://1findr.1science.com/home

було впроваджено для забезпечення стійкості до низки комунікаційних збоїв.

Витяг метаданих аналізує, очищає та узгоджує завантажені метадані та зберігає їх у внутрішній структурі даних. Процес узгодження та очищення розглядає той факт, що різні постачальники даних описують одну й ту саму інформацію по-різному (синтаксична неоднорідність), а також мають різні інтерпретації для однієї й тієї самої інформації (семантична неоднорідність).

Завантаження повного тексту: Використовуючи посилання, отримані з метаданих, CORE намагається завантажити та зберегти повні тексти публікацій. Цей процес є нетривіальним. Як було сказано вище, OAI-PMH наразі є стандартним протоколом для обміну даними між сховищами. Хоча OAI-PMH спочатку був розроблений лише для збору метаданих, через його широке застосування та відсутність альтернатив він використовувався як точка входу для збору і повного тексту. Збір повного тексту досягається шляхом витягу URL-адрес із записів метаданих, зібраних за допомогою OAI-PMH. А потім витягнуті URL-адреси використовуються для визначення розташування фактичного ресурсу. Однак протокол OAI-PMH безпосередньо підтримує лише збирання метаданих, тобто необхідно реалізувати додаткові функції, щоб використовувати його для збору вмісту. Розташування посилань на повні тексти у метаданих не стандартизовано, і записи метаданих зазвичай містять кілька посилань. З метаданих не зрозуміло, яке з цих посилань вказує на описане представлення ресурсу, і в багатьох випадках жодне з них не робить цього безпосередньо. Тому всі можливі посилання на сам ресурс потрібно витягнути з метаданих і перевірити, щоб визначити правильний ресурс. Крім того, OAI-PMH не сприяє перевірці того, що виявлений ресурс справді є описаним ресурсом. Подолання цих проблем сприяло прийняттю формату метаданих RIOXX[20] або рекомендацій OpenAIRE[21]. Однак питання однозначного зв'язку запису метаданих і описаного ресурсу залишається актуальним.

Витяг інформації. Звичайний текст із завантажених документів витягується та обробляється для створення напів-структурованого представлення. Цей процес включає низку завдань добування інформації, таких, наприклад, як витяг посилань.

Збагачення працює шляхом доповнення метаданих і повного тексту, отриманих від постачальників даних, додатковими даними з різних джерел. Деякі збагачення виконуються безпосередньо конкретними завданнями в процесі виконання збору документів, зокрема, визначення мови і типу документа (стаття, презентація, дипломна робота, інший). Ці збагачення зазвичай передбачають застосування моделей машинного навчання та інструментів на основі правил для збору додаткової інформації про записи. І для певного запису виконуються лише один раз. Решта збагачень, які включають зовнішні набори даних, виконуються ззовні, незалежно від процесу збору, і вбудовуються в набір даних. Це здійснюється шляхом збору різноманітної інформації з великих сторонніх наукових наборів даних. Інформація включає метадані, які не обов'язково змінюються, як-от, ідентифікатор DOI, а також метадані, які зазнають змін, наприклад, кількість цитувань. Саме через це такі збагачення виконуються періодично, тобто всі записи в CORE проходять цей процес повторно через певні проміжки часу. Початкове відображення запису здійснюється за допомогою DOI у разі його доступності. Однак, оскільки більшість записів з репозиторіїв не мають DOI, відбувається процес зіставлення з базою даних Crossref, використовуючи підмножину полів метаданих, включаючи назву, авторів і рік публікації. Після того, як зіставлення виконано, можна узгодити поля, а також зібрати широкий спектр додаткових корисних даних із відповідних зовнішніх баз даних, тим самим збагачуючи запис CORE. Такі дані включають ідентифікатори ORCID, інформацію про цитування, додаткові посилання на вільно

20 https://rioxx.net/

21 https://guidelines.openaire.eu/

доступні повні тексти, інформацію про галузь дослідження тощо.

Індексування. Останнім кроком збирання даних є індексування зібраних даних. Отриманий індекс забезпечує роботу служб CORE, включно із пошуком, API і FastSync.

Сервіси інтелектуального аналізу даних. Особливий інтерес для нас становлять сервіси, які надаються/ плануються самими агрегаторами (або із залученням сторонніх організацій) своїм користувачам на основі сучасних досліджень інтелектуального аналізу тексту та даних наукової літератури. Тут ми просто наводимо їх список: виявлення плагіату нещодавно надісланих публікацій; семантометрики [5]; підбір публікаціям відповідних рецензентів; аналіз науково-дослідних трендів; рекомендації щодо роботи зі співавторами, заходами проведення тощо; ідентифікація різних версій одної і тої самої статті (наприклад, препринти та постпринти); визначення того, чи відповідає публікація умовам проведення заходу; визначення важливості, настанов або типу цитування; узагальнення результатів досліджень; побудова індексу цитувань; витяг або добування важливих слів або фраз; категоризація публікацій за сферами досліджень; визначення типу публікації; витяг цитат для бібліометрії.

3. Вимоги та рекомендації щодо індексування вмісту постачальників даних

Майже всі розглянуті у попередньому розділі служби агрегації документів відкритого доступу мають більш-менш стандартні процедури реєстрації та вимоги до провайдерів даних. У цьому розділі стисло описано основні вимоги та вказівки щодо найкращого налаштування репозитаріїв для індексування у таких агрегаторах.

Підтримка протоколу OAI-PMH. Майже всі сучасні системи агрегації використовують цей протокол для регулярного збору та оновлення інформації, яку вони зберігають від своїх постачальників даних. Обов'язково слід переконатися, що ваш репозиторій видимий через OAI-PMH. Щоб пройти індексацію, постачальники даних мають надати інформацію про свої ресурси через OAI-PMH.

Доволі поширеною проблемою є те, що кінцева точка OAI-PMH постачальника даних неправильно налаштована або не функціонує. Це може статися навіть тоді, коли інші функції репозитарію працюють без проблем. Тому важливо, щоб репозитарії стежили за тим, щоб їхня кінцева точка OAI-PMH залишалася функціональною. Це можна зробити кількома способами.

Валідація через інструменти перевірки OAI-PMH. Постачальники даних, які розглядають можливість реєстрації в агрегаторі можуть використати інструмент перевірки Open Archives OAI-PMH, https://www.openarchives.org/Register/ValidateSite, рис. 10. Інструмент очікує базову URL-адресу OAI-PMH і проводить низку перевірок, щоб оцінити коректність конфігурації кінцевої точки. Якщо валідатор повертає помилки, їх потрібно буде виправити до реєстрації в агрегаторі. Рекомендується виконати цей крок до спроби реєстрації в агрегаторі.

Реєстрація постачальників даних у агрегаторі. Відвідайте сторінку реєстрації постачальника даних у та введіть URL-адресу OAI-PMH свого репозитарію. Якщо URL-адреса OAI-PMH є дійсною, репозитарій буде зареєстровано за умови проходження додаткових перевірок, описаних нижче. Важливо зазначити, що перевірка правильності конфігурації кінцевої точки OAI-PMH не гарантує, що вона правильно відображає всі записи метаданих із системи репозитарію. Це можна перевірити лише після того, як агрегатор спробує отримати дані з репозитарію.

Постачальники даних, які вже зареєстровані в агрегаторі. Зареєстровані постачальники даних можуть отримати доступ до інформаційної панелі агрегатора, як, наприклад, у системі CORE, рис. 11. Панель управління надає доступ до звіту про збір даних, який містить таку інформацію, як останній випадок успішного збору даних, кількість знайдених та проіндексованих метаданих і повнотекстових записів (якщо повні тексти збираються агрегатором), будь-які помилки або попередження, що виникли в процесі збирання даних.

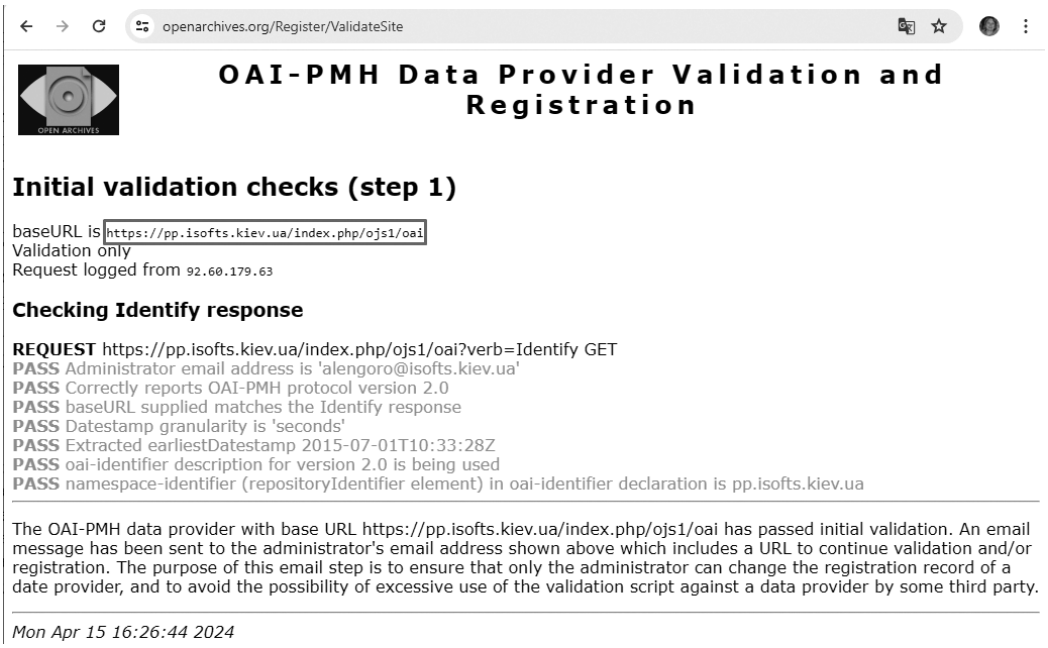


Рис. 10. Зразок перевірки OAI-PMH журналу «Проблеми програмування»

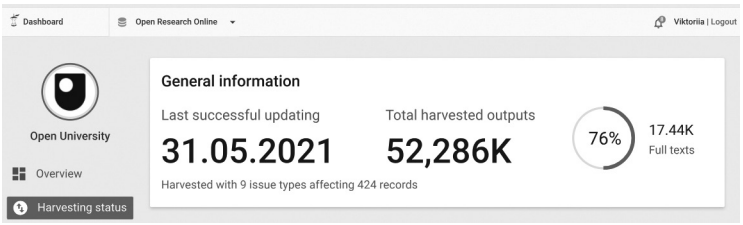


Рис. 11. Показ збору інформації агрегатором CORE репозитарію Відкритого університету

Нерідко кількість проіндексованих записів метаданих у агрегаторах може дещо відрізнятися від кількості, яку постачальник даних може бачити у своїй системі репозитарію, оскільки інформаційна панель надає незалежний зовнішній погляд на репозитарій. Відхилення можуть виникати через низку причин, зокрема, тому, що ваш репозитарій не відкриває всі записи через інтерфейс OAI-PMH, деякі записи відключені або їхні метадані не відповідають мінімальним вимогам до якості. Тому постачальникам даних рекомендується регулярно перевіряти і виправляти будь-які помилки та попередження, що з'являються на інформаційній панелі.

Реєстрація репозитарію у відкритих реєстрах. Агрегатори використовують визнані відкриті реєстри репозитаріїв, як-от OpenDOAR[22], для виявлення нових постачальників даних з відкритим доступом. Тому бажано, щоб ваш репозитарій був попередньо включений до такого міжнародного списку репозитаріїв і щоб базова URL-адреса OAI-PMH підтримувалася в актуальному стані в цьому реєстрі. Наприклад, для реєстрації будь-якого провайдера даних у агрегаторі OpenAIRE, попередня реєстрація у реєстрі OpenDOAR є обов'язковою вимогою. Нерідко зустрічаємо і таку ситуацію, коли одні агрегатори автоматично індексують інші агрегатори, як, наприклад, серед провайдерів даних агрегатора CORE є агрегатор DOAJ (Каталог журналів відкритого доступу), кілька серверів препринтів та наукових праць, які мають ідентифікатори публікацій, зареєстрованих у Crossref.

22 http://v2.sherpa.ac.uk/opendoar/

Конфігурація метаданих. Для того, щоб агрегатор міг успішно збирати дані із репозитарію, як було сказано вище, він повинен мати кінцеву точку, сумісну зі стандартом OAI-PMH. Більшість поширених програм, на яких будуються репозитарії, такі як EPrints, DSpace або Open Journal Systems (OJS) підтримують OAI-PMH. Крім того, для успішного індексування важливо, щоб метадані записів у репозитарії були правильними та у підтримуваному форматі.

Обов'язкова підтримка стандартів метаданих. Агрегатори можуть підтримувати кілька форматів метаданих для опису наукових ресурсів [1]. Як правило, метадані повинні бути представлені в одному з наступних профілів.

i. Dublin Core / Extended Dublin Core (мінімум). Дублінське ядро є однією з найпростіших і найпоширеніших схем метаданих. Розширена версія була формалізована у вигляді термінів метаданих DCMI 2019 року, з приблизно 70 полями даних. Хоча Dublin Core та Extended Dublin Core є достатніми для індексації, наприклад, у агрегаторі BASE, ця схема надає лише обмежені можливості для опису наукових ресурсів порівняно з OpenAIRE Guidelines та Rioxx.

ii. OpenAIRE Guidelines була створена для підтримки стратегії відкритого доступу Європейської Комісії та для задоволення вимог інфраструктури OpenAIRE. Ця нова версія настанов відповідно до розширення цілей ініціативи OpenAIRE та її інфраструктури має ширшу сферу застосування. Впроваджуючи ці настанови, менеджери репозитаріїв дозволять авторам, які розміщують публікації в їхніх репозитаріях, відповідати вимогам Європейської Комісії щодо відкритого доступу.

iii. RIOXX. Агрегатор CORE рекомендує репозитаріям використовувати формат метаданих RIOXX (The Research Outputs Metadata Schema), як найбільш придатний профіль метаданих для опису результатів наукових досліджень. Ця схема метаданих була розроблена для інституційних репозитаріїв з метою обміну метаданими про наукові ресурси, які вони містять. Спочатку розроблена для задоволення вимог до звітності Дослідницьких рад Великої Британії (Research Councils UK, RCUK)[23], RIOXX також виявилася корисною як стандарт для обміну метаданими між репозиторіями та мережевими сервісами, такими як великомасштабні агрегатори метаданих, такі як CORE. RIOXX фокусується на застосуванні узгодженості до полів метаданих, що використовуються для ідентифікаторів наукових результатів, людей і організацій, спонсорів досліджень і проєктів/грантів, і призначений для підтримки послідовного відстеження наукових публікацій з відкритим доступом у різних наукових системах. RIOXX має кілька переваг над іншими схемами. Зокрема, ця схема була розроблена з особливим акцентом на забезпеченні ефективного обміну даними між репозиторіями і машинними агентами, що робить збір даних швидшим і точнішим завдяки уникненню неоднозначності у зв'язках метаданих, що описують ресурси, такі як повні тексти, набори даних, програмне забезпечення, і самі ресурси. Вона надає значні можливості для опису широкого спектру наукових властивостей, корисних для звіту та аналізу досліджень, таких як ідентифікатори грантів, ідентифікатори проєктів, ідентифікатори спонсорів, інформація про ліцензування тощо.

Схема RIOXX сумісна з іншими протоколами машинного доступу до наукових документів, зокрема, Signposting[24] та ResourceSync[25], і логічно інтегрується з ними. Вона забезпечує механізми для розрізнення ресурсів, що знаходяться під управлінням репозитарію, від ресурсів, які знаходяться під зовнішнім управлінням, що дозволяє CORE надавати перевагу посиланням на репозитарій, де це можливо (і доречно), замість зовнішніх посилань на веб-сайт видавця. CORE надає валідатор[26] метаданих для постачальників даних, які підтримують RIOXX.

3 https://www.ukri.org/

4 https://signposting.org/

5 https://www.openarchives.org/rs/1.1/resourcesync

6 https://core.ac.uk/membership-documentation#rioxx-validator

Кодування символів. Весь вміст інтерфейсу OAI-PMH (заголовки, імена авторів, анотації) кодується в UTF-8. Інші кодування або дублювання кодувань спричиняють помилки у відображенні результатів пошуку з вашого джерела.

Розділення кількох записів у полі метаданих. Якщо у полі метаданих вказані кілька записів (наприклад, ім'я автора та його ідентифікатор ORCID), розділяйте їх пробілами, крапкою з комою та пробілами. Таке розділення дозволяє індексувати інформацію окремо і зробити її доступною для пошуку.

Повнотекстова конфігурація. Як вже було сказано вище, низка потужних можливостей деяких агрегаторів заснована на здатності індексувати повнотекстовий контент. До них належать повнотекстовий пошук, рекомендації повнотекстового контенту, виявлення версій і дубльованого контенту та інші. Підтримувані повнотекстові формати – тільки дійсні файли PDF, DOC або DOCX, які містять текст, що витягується. Якщо PDF-файл це відскановане зображення, то використання документа в агрегаторах буде обмежено. Для успішного індексування повного тексту статті на нього має бути посилання одним з наступних способів.

i. Пряме посилання на повний текст. Рекомендується, аби постачальники даних надавали однозначне посилання на повний текст у метаданих. Правильна стратегія посилання залежить від використовуваної схеми метаданих. Там, де використовується Дублінське ядро, рекомендується надавати пряме посилання на повний текст у першому входженні dc:identifier.

<dc:identifier>http://oro.open.ac.uk/37823/1/jcdl2019_v7.pdf</dc:identifier>

Система EPrints за замовчуванням дотримується цієї рекомендації. Такий підхід значно зменшує навантаження, яке агрегатор накладає на репозитарій під час індексації. Якщо це неможливо, то агрегатор може індексувати вміст, якщо в метаданих є чітко визначене посилання, яке вказує на машинодоступний документ (або PDF, або формат DOC/DOCX).

ii. Непряме посилання на повний текст. Агрегатор автоматично визначає, коли постачальник даних не вказує посилання на повний текст безпосередньо, як було описано вище. У таких випадках індексатор відвідає цільову сторінку і збере з неї посилання так само, як це зробив би користувач, намагаючись знайти посилання на документ на сторінці.

Щоб переконатися, що правильний повний текст був знайдений, тобто повний текст, що відповідає запису метаданих, агрегатор запускає процес зіставлення заголовка запису з заголовком, що міститься в PDF-файлі. Цей процес не є на 100% точним, спричиняє додаткове навантаження на сервер і вимагає більше часу на обробку одного документа. Це не працює для документів, які потребують розпізнавання тексту, і тому вони будуть відкинуті.

Висновки

У роботі визначена термінологія та проаналізована низка доступних і популярних агрегаційних служб та баз даних публікацій відкритого доступу BASE, OpenAIRE, CORE та ін. Здійснена спроба узагальнити їхні вимоги та рекомендації щодо збору ресурсів та індексування вмісту постачальників даних. Наші ЕБ, які ми створили і підтримуємо протягом уже майже 20 років, присутні у цих агрегаторах. Тому деякий практичний досвід ми маємо, і це дозволяє нам виконати проєкт НАНУ «Відкрита наука» зі створення агрегатора публікацій, який у подальшому буде підключено до більш потужних європейських або світових агрегаторів наукових публікацій.

Література

1. Проскудіна Г.Ю., Кудім К.О., Резніченко В.А. VUFIND: відкрите рішення для інтеграції бібліотечних колекцій // Проблеми програмування. − 2023. − № 4 – С. 15–26. https://pp.isofts.kiev.ua/ojs1/article/view/590
2. Резніченко В.А, Новицький О.В., Проскудіна Г.Ю. Інтеграція наукових електронних бібліотек на основі протоколу OAI-PMH // Проблеми програмування. −

2007. − № 2 – С. 97–112. http:// dspace.nbuv.gov.ua/hadle/123456789/291
3. Кнот П., Херрманнова Д., Канселлієрі М. та ін. CORE: Глобальна служба агрегації документів відкритого доступу. Nature Scientific Data 10, 366 (2023). https://www.nature.com/articles/s41597-023-02208-w
4. Ammar, W. et al. Construction of the Literature Graph in Semantic Scholar. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers): 84–91 (2018).
5. Кнот П., Херрманнова Д.. К вопросу о семантометрии: новый критерий для оценки вклада научной публикации на основе семантического сходства // Международный форум по информатизации. — 2015. — Т. 40, № 1. — С. 3-8.

References

1. Proskudina G.Yu., Kudim K.A., Reznichenko V.A. 2023. VuFind: an open solution for integrating library collections . Problems in programming. no. 4, pp. 15–26. (in Ukrainian). Available from: https://pp.isofts.kiev.ua/ojs1 / article/view/590 [Accessed 17/10/2024].
2. Reznichenko V.A.,. Novytskyi O.V., Proskudina G.Yu., 2007. OAI-PMH protocol-based integration of scientific digital libraries. Problems in programming, no. 2, pp. 97–112. (in Ukrainian). Available from: http://dspace.nbuv.gov.ua/handle/123456789/291 [Accessed 17/10/2024].
3. Knoth, P., Herrmannova, D., Cancellieri, M. et al. CORE: A Global Aggregation Service for Open Access Papers. Sci Data 10, 366 (2023). https://doi.org/10.1038/s41597-023-02208-w [Accessed 17/10/2024]
4. Ammar, W. et al. Construction of the Literature Graph in Semantic Scholar. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers): 84–91 (2018). [Accessed 17/10/2024].
5. Petr Knoth and Drahomira Herrmannova. owards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. D-Lib Magazine. Volume 20, Number 11/12 (2014) https://www.dlib.org/dlib/november14/knoth/11knoth.html [Accessed 18/10/2024].

Одержано: 15.10.2024
Внутрішня рецензія отримана: 28.10.2024
Зовнішня рецензія отримана: 02.11.2024

Про авторів:

Проскудіна Галина Юріївна,
науковий співробітник.
http://orcid.org/0000-0001-9094-1565

Кудім Кузьма Олексійович,
молодший науковий співробітник.
http://orcid.org/0000-0001-9483-5495

Місце роботи авторів:

Інститут програмних систем НАНУ,
03187, Київ-187, проспект Академіка Глушкова 40, корпус 5.
Phone: +38(050) 368 49 27.
E-mail: kuzmaka@gmail.com
guproskudina@gmail.com