



УДК 025.355:077]:004.5]:025.4.03

DOI: 10.36273/2076-9555.2024.6(335).24-33

**Валерій Резніченко,**

кандидат фізико-математичних наук,  
старший науковий співробітник  
Інституту програмних систем НАН України,  
e-mail: reznichenko.valery47@gmail.com,  
ORCID: <https://orcid.org/0000-0002-4451-8931>

**Галина Проскудіна,**

наукова співробітниця  
Інституту програмних систем НАН України,  
e-mail: guproskudina@gmail.com  
ORCID: <https://orcid.org/0000-0001-9094-1565>

**Кузьма Кудім,**

молодший науковий співробітник  
Інституту програмних систем НАН України,  
e-mail: kuzmaka@gmail.com  
ORCID: <https://orcid.org/0000-0001-9483-5495>

### Сучасний підхід до створення та інтеграції бібліотечних електронних інформаційних ресурсів

У статті подано огляд програмного забезпечення та функціональних можливостей Наукової електронної бібліотеки періодичних видань Національної академії наук України. Розглянуто питання інтеграції бібліотечних систем в єдині системи обслуговування користувачів на прикладі сучасних глобальних агрегаторів документів відкритого доступу BASE, OpenAIRE та CORE. Проаналізовано їхні кількісні характеристики, зокрема кількість зібраних описів документів і повних текстів, кількість постачальників даних, наявність інтерфейсу прикладного програмування для отримання даних. Розглянуто склад і види постачальників даних, як-от інституційні репозитарії, відкриті журнали, видавництва, наукові репозитарії препринтів, тематичні електронні бібліотеки, а також системи, які теж є агрегаторами. Досліджено, яку саме інформацію про документи збирають агрегатори, як її представлено в інтерфейсі користувача, яку інформацію збирають про постачальників даних, як її подано в інтерфейсі користувача, в який спосіб відбувається взаємодія агрегатора з постачальниками даних, які протоколи обміну даних підтримуються, як відбувається оновлення зібраних даних. Наведено перелік корисних для науковців сервісів, що їх можуть надавати сучасні агрегатори на базі зібраних корпусів даних, використовуючи методи машинного навчання, бібліометрії, вебометрики, альтиметрії, семантометрії. Автори статті є розробниками низки наукових електронних бібліотек із відкритим доступом і вже зареєстровані як провайдери даних у цих системах. Нині триває робота зі створення агрегатора наукових ресурсів НАН України, тому вивчення світового досвіду в цій сфері є важливим для ефективного розв'язання окресленого завдання.

**Ключові слова:** електронні бібліотеки; розподілені документні інформаційні системи; інтеграція; відкритий доступ; провайдер сервісів; провайдер даних; протокол OAI-PMH

**Постановка проблеми.** Стрімкий розвиток та активне використання сучасних інформаційних технологій зумовили розгортання процесів переведення накопиченої людством інформації в електронну форму для створення принципово нових видів інформаційних ресурсів. Ця форма подання даних дає змогу на якісно іншому рівні організувати процеси виробництва, зберігання й поширення інформації. Забезпечення публічного доступу користувачів до джерел в електронній формі стало одним із найважливіших завдань інформаційного супроводу науки, культури та

освіти. Сьогодні очевидно, що найефективніше досягти цієї мети можна завдяки створенню електронних бібліотек (ЕБ) — розподілених документних інформаційних вебсистем, які дають змогу надійно зберігати й ефективно використовувати різноманітні колекції електронних документів (текст, графіка, аудіо, відео тощо), доступних у зручній для кінцевого користувача формі через глобальні мережі передавання даних. Створення ЕБ — якісно новий рівень вироблення, зберігання, організації та поширення різноманітної інформації.

**Аналіз досліджень і публікацій.** Формування перших ЕБ у США припадає на 1980-ті рр., а у Великій Британії — на початок 1990-х рр. Зазвичай такі роботи розпочинали з виконання невеликих проєктів, які протягом кількох років набували статусу національних і міжнародних програм. Бібліотека Конгресу США реалізує Програму створення національної електронної бібліотеки (National Digital Library Program, NDLP), що передбачає конвертацію історичних колекцій у цифрову форму для надання доступу до них в інтернеті. Від 1994 р. з ініціативи Національного наукового фонду (National Science Foundation, NSF), Агентства передових оборонних і дослідницьких проєктів (Defence Advanced Research Project Agency, DARPA) і Національного управління з повітроплавання та дослідження космічного простору (National Aeronautics and Space Administration, NASA) у Сполучених Штатах розгорнуто дослідницьку програму, присвячену ЕБ, — Ініціативу електронних бібліотек (Digital Libraries Initiative, DLI). На початку 1998 р. ці програми було об'єднано в міжвідомчу програму, у якій, окрім зазначених організацій, взяли участь Національна медична бібліотека, Агентство зі статистики США, Національний гуманітарний фонд, Національний архів США та інші організації.

В Японії спільними зусиллями Агентства з упровадження нових технологій, Національної парламентської бібліотеки, понад 20 бібліотек і культурних центрів, а також низки міністерств формується електронна "Бібліотека ХХІ століття".

Від 1995 р. реалізується національна програма Великої Британії зі створення ЕБ, що має назву eLib.

Такі проєкти зазвичай мають суттєву державну фінансову підтримку. У ЮНЕСКО ухвалено програму "Пам'ять світу" (Memory of the World Programme), яка, зокрема, має сприяти створенню ЕБ. У 1990-х рр. було розгорнуто програму "Створення європейського бібліотечного простору". В її межах ініційовано та реалізовано численні проєкти, пов'язані з автоматизацією бібліотек і створенням ЕБ. В інтернеті наявно чимало інформації щодо дослідницької діяльності, агенцій, організацій, конференцій, науково-дослідних центрів у галузі ЕБ, а також є велика кількість сайтів із мережевими журналами, присвяченими тематиці ЕБ.

**Мета статті** — підсумувати майже двадцятирічний досвід роботи з вивчення, дослідження, створення, підтримки та поширення в Україні сучасних електронних бібліотек, а також

їх інтеграції в єдині портали обслуговування клієнтів на основі аналізу особливостей розроблення найпотужніших на цей час систем.

У першій частині пропонованої статті подано огляд функціональних можливостей Наукової електронної бібліотеки періодичних видань НАН України (NASPLIB), створеної в Інституті програмних систем НАН України у процесі виконання одного із завдань Програми інформатизації НАН України. NASPLIB було встановлено в Національній бібліотеці України імені В. І. Вернадського (<http://dSPACE.nbuv.gov.ua/>), згодом перенесено на сервери Інституту теоретичної фізики ім. М. М. Боголюбова НАН України, наразі підтримується Інститутом програмних систем НАН України.

У другій частині дослідження розглянуто особливості й приклади інтеграції таких систем.

### **Програмне забезпечення**

З метою вибору програмного забезпечення (ПЗ) для формування NASPLIB було проаналізовано майже 20 програм [1] для створення ЕБ, серед яких обрано три найпопулярніші: Greenstone [2], EPrints [3] та DSpace [4]. На підставі подальшого аналізу було обрано ПЗ DSpace, розроблене в Массачусетському технологічному інституті у співпраці з компанією Hewlett-Packard. Це програмне забезпечення використовують як систему для створення електронних репозитаріїв і збереження інформації науково-дослідних та навчальних установ різного типу. Це відкрите ПЗ розроблено з використанням безплатних складників (операційна система, вебсервер, СУБД, засоби індексації та пошуку). Важливою особливістю є те, що DSpace дає змогу репрезентувати ієрархічну структуру інформаційних ресурсів з інтеграцією до неї різноманітних функцій, що виявилось надзвичайно корисним під час проєктування NASPLIB.

Першу версію DSpace було випущено у 2002 р., а нині вийшла вже сьома версія [5]. Новий реліз об'єднав два користувацьких інтерфейси та їхні спільноти JSPUI та XMLUI, які до цього працювали паралельно, всі функції переосмислено та реалізовано в межах єдиної популярної сучасної технології користувацького інтерфейсу Angular. Останній підкріплено цілковито новим REST API, який відкриває всі дані та функції в інтернеті, даючи змогу DSpace інтегруватися або взаємодіяти із зовнішніми системами / сервісами значно ефективніше.

Нині система доступна в таких варіантах: DSpace, DSpace XOAI, DSpace IRIS і DSpace CRIS — це різні розширення DSpace, кожне з яких пропонує унікальні функції та можливості.

DSpace — базова версія, яка дає змогу керувати публікаціями; DSpace XOAI — розширення DSpace, що надає підтримку протоколу OAI-PMH (Open Archives Initiative — Protocol for Metadata Harvesting; <http://www.openarchives.org/OAI/openarchivesprotocol.html>) для обміну метаданими. DSpace IRIS (Institutional Research Information System) та DSpace CRIS (Current Research Information System) — безплатні й відкриті розширення DSpace для управління інформацією та даними про наукові дослідження. Вони поліпшують функціональність DSpace і модель даних цього ПЗ, залишаючись узгодженими з його кодовою базою.

Отже, варіанти DSpace дають змогу збирати дані та управляти ними, а також інформацією про наукові дослідження, зокрема дослідників та групи, відділи й організаційні одиниці, лабораторії та інфраструктуру, проекти та співпрацю, контракти та заходи, публікації та патенти, нагороди тощо.

DSpace має вагомий інформаційний супровід: на сайтах підтримки розміщено багато змістовних інструкцій із вивчення, встановлення, оновлення, резервного копіювання цього ПЗ [5].

#### **Вебсредовище та доступ до NASPLIB**

NASPLIB підтримує вебінтерфейс, що дає змогу будь-якій особі, яка має доступ в інтернет, використовувати створюваний ресурс. Така "загальнодоступність" визначає потребу розв'язання проблем організації доступу безпосередньо до інформаційних ресурсів (IP). Зазвичай більшість ЕБ надають можливість звернення до всіх наявних ресурсів, за винятком повнотекстових документів. Доцільно виокремити такі категорії доступу: відкритий, обмежений, умовно обмежений. NASPLIB є ЕБ відкритого доступу, що дає змогу користувачеві взаємодіяти з повними текстами статей без будь-яких обмежень.

Свого часу світова спільнота підтримала Ініціативу відкритих архівів (Open Archives Initiatives, OAI; <http://www.openarchives.org/>), мета якої — всебічне сприяння у створенні відкритих електронних цифрових масивів. Під *відкритим* доступом розуміють доступ читачів до інформації за допомогою інтернету з правом читати, завантажувати, копіювати, розповсюджувати, друкувати, здійснювати пошук, посилатися на повнотекстові документи, індексувати їх, тобто використовувати із законною метою без фінансових, юридичних і технічних перешкод. Саме такий принцип реалізується в NASPLIB.

Зазначимо, що в певних ситуаціях доступ до текстів статей у NASPLIB може бути закритий

або наданий окремому колу користувачів. *Обмежений та умовно обмежений* доступ передбачають лімітування доступу до IP визначеному колу осіб згідно зі встановленими критеріями, котрих дотримується адміністративна служба ЕБ.

#### **Користувачі NASPLIB**

Є дві категорії користувачів NASPLIB: кінцеві користувачі та служба підтримки функціонування ЕБ.

*Кінцеві користувачі*, які отримують послуги, що надаються ЕБ, поділяються на дві групи. Перша — це *анонімний користувач*, який звертається до ресурсів ЕБ без ідентифікації. Він має право вільного доступу до відкритих IP, а також до тих сервісів, які не потребують ідентифікації особи. Друга — це *зарєєстрований користувач*, який пройшов одноразову процедуру реєстрації самостійно. Цей простий крок передбачає надання основних відомостей про особу (прізвище, ім'я, адреса електронної пошти), а також отримання пароля. Адреса електронної пошти виконує функцію логіна. Відомості про реєстрацію використовують для надання специфічних прав доступу до інформаційних ресурсів і сервісів.

До служби підтримки функціонування ЕБ належать оператори, редактори й адміністратор. *Оператор* — зарєєстрований користувач, якому надано право розміщувати статті в ЕБ. *Редактор* — зарєєстрований користувач, який має право перевіряти коректність документів, котрі вводять в ЕБ. *Адміністратор* — особа, яка відповідає за налаштування, конфігурування й підтримку безперебійної роботи ЕБ. Він має всі права та може використовувати всі функціональні можливості ЕБ, як-от надання користувачам прав на введення документів, повнотекстовий доступ до документів з обмеженим доступом, редагування документів тощо.

#### **Інформаційна структура NASPLIB**

У NASPLIB підтримується така ієрархічна структура зберігання статей: науковий напрям (секція) НАН України; відділення НАН України; журнал відділення; том журналу (випуски за один рік); номер журналу; стаття. Кожний елемент структури має домашню сторінку, на якій розташовано загальну інформацію про цей елемент. Однак, крім суто інформативної функції, така ієрархія передбачає виконання двох важливих завдань: пошук статей та отримання статистичної інформації, що можливо здійснювати як по всій NASPLIB, так і в межах будь-якого рівня цієї ієрархії. Відомо, що IP можуть одночасно зберігатися у формі єдиної сукупності та як сукупність цифрових колекцій.

*Колекція* — це масив IP, об'єднаних у логічне ціле на підставі певних критеріїв. Наприклад, у процесі створення наукової ЕБ колекції можуть об'єднувати журнальні статті, наукові звіти, дисертації, праці конференцій тощо. У NASPLIB також є поняття колекції. У нашому випадку колекція — це номер журналу. Між статтями колекцій можна подавати посилання. Наприклад, якщо одну й ту саму статтю опубліковано в кількох журналах, то фізично її можна зберігати в одній колекції, а в інших — уміщувати посилання на неї.

Усі розміщені статті, що формують архів NASPLIB, можна змінювати, замінювати, видаляти тимчасово або назавжди, закривати до них доступ або надавати доступ певному колу осіб, імпортувати та експортувати. Ці функції перебувають у зоні відповідальності адміністратора NASPLIB. Хоча можливості ЕБ дають змогу зберігати документи в різних форматах (pdf, doc, html, txt тощо), основним було обрано pdf. Водночас у деяких випадках допускається робота з документами в інших форматах.

#### **Введення статей**

Введення статей передбачає розміщення в NASPLIB як повного тексту документа, так і його опису. Опис — це перелік значень описових полів, що допомагають схарактеризувати статтю. В ЕБ такі описи мають назву *метаданих*. Питання вибору складу метаданих для опису IP є принципово важливим. Нині розроблено багато різних наборів метаданих, або *схем метаданих*, для опису тих чи інших видів IP. Найпоширенішою у сфері ЕБ є схема метаданих Дублінського ядра (Dublin Core Metadata Initiative, DCMI; <http://dublincore.org/>), котру використовують для опису статей у NASPLIB.

У Науковій електронній бібліотеці періодичних видань НАН України статті описують такими метаданими: класифікаційний індекс УДК; автори — тільки мовою оригіналу; назва статті — всіма мовами, якими наведено назву в статті; анотація статті — всіма мовами, якими наведено анотацію в статті; бібліографічний опис статті; тематична рубрика журналу, під якою опубліковано статтю; інформація про спонсорську підтримку / висловлення вдячності; мова публікації; ISSN журналу.

Окрім цих метаданих, система автоматично фіксує додаткову інформацію (хто ввів статтю, дата та час введення, результати редакторської перевірки тощо).

Введення, що є багатокроковим процесом (введення опису статті, введення файлів статті,

попередня перевірка введеної інформації, прийняття ліцензійної угоди тощо), може здійснюватися з будь-якого комп'ютера, під'єданого до інтернету. Користувач має змогу призупинити введення статті та продовжити в будь-який зручний час. Право на введення надається за кожним номером журналу окремо; його можуть мати й самостійні користувачі, й користувачі, об'єднані в групи.

#### **Редакторська перевірка статей**

Введену статтю редактор перевіряє на відповідність вимогам оформлення в NASPLIB. У разі виявлення помилок чи неточностей редактор або усуває їх самостійно, або надсилає статтю на доопрацювання із зазначенням причин повернення. Повідомлення про потребу доопрацювання надходить користувачеві електронною поштою. Якщо редактор прийняв статтю, вона зберігається в архіві, після чого стає доступною всім користувачам.

Кожен зареєстрований користувач, зокрема й ті, хто вводить статті, має домашню сторінку, що містить три списки:

- список статей, які перебувають у стані введення;
- список статей, які вже введені, але ще не пройшли редакторську перевірку;
- список статей, які редактор повернув для доопрацювання.

Користувач-редактор у роботі оперує двома списками:

- списком статей, які надійшли на редагування;
- списком статей, які відібрано для редагування.

Наявність двох списків пов'язана з тим, що журнальну статтю можуть опрацьовувати кілька редакторів.

#### **Організація доступу до статей**

У NASPLIB наявно два варіанти знаходження документів: послідовний доступ і пошук.

*Послідовний доступ* використовують, якщо точно відомо, в якому номері журналу міститься потрібна стаття. Алгоритм доступу такий: журнал вибирають зі списку, потім знаходять том, номер журналу, назву статті зі змісту.

Кожна стаття має самостійну сторінку, на якій наведено її опис та посилання на повний текст. Є два види опису: короткий і повний. Зазвичай короткого опису цілком достатньо, щоб мати загальне уявлення про статтю (УДК, автори, назва, бібліографічний опис, анотація) і зрозуміти, чи потрібно відкривати повний текст. Повний опис містить низку додаткових відомостей про

статтю, у тому числі автоматично зафіксовані самою системою (хто розмістив і відредагував статтю, чи були зауваження та які саме тощо).

У разі послідовного доступу важливою є можливість візуалізації шляху ієрархічної структури від домашньої сторінки NASPLIB до тієї, на якій перебуває користувач. Це, по-перше, дає йому змогу орієнтуватися у змісті журналу і, по-друге, повернутися до будь-якого етапу інформаційного пошуку (рис. 1).

Домашня сторінка → Суспільні гуманітарні науки → Відділення історії, філософії та права → Краєзнавство → Краєзнавство, 2010 → Краєзнавство, 2010, № 1–2 → Переглянути статтю

Рис. 1. Приклад візуалізації навігації за ієрархічною структурою

Ще одним різновидом послідовного доступу є можливість перегляду статей за роками, авторами, назвами й тематичними рубриками. Наприклад, можна відкрити домашню сторінку журналу, потім список усіх авторів журналу і, навівши курсор на прізвище, отримати дані про всі статті цього автора, опубліковані в журналі.

Пошук є ключовою функцією NASPLIB із погляду кінцевого користувача. У пошуку важливими є два поняття: простір пошуку та пошуковий вираз.

*Простір пошуку* — це множина статей, за якими здійснюється пошук. Простором пошуку може слугувати вся сукупність статей, а також групи статей, що належать до будь-якої верхівки ієрархічної структури. Це забезпечує можливість пошуку за статтями, науковими напрямками, відділеннями НАН України, найменуваннями журналів, томом і номером. Уточненням простору пошуку є вказівка на те, чи буде здійснюватися пошук за метаданими статті та/або за самим текстом. Пошук за метаданими дає змогу шукати статті за автором, УДК, словами з анотації, мовою публікації тощо (або комбінації метаданих). Повнотекстовий пошук передбачає знаходження статей, тексти яких містять слова чи фрази, задані в пошуковому виразі.

*Пошуковий вираз* — це конструкція пошукової мови, що дає змогу сформулювати умови відбору потрібних статей. Для сучасних ЕБ з погляду пошуку оптимальним є використання механізму індексації та пошуку ПЗ Apache Lucene ([www.lucene.apache.org/](http://www.lucene.apache.org/)), котре застосовують у NASPLIB. Наведемо основні можливості пошукових засобів:

1. Найпростішим варіантом пошукового виразу є перелік слів і фраз, які мають бути наявними у статті, що цікавить користувача (в її метаданих, у повному тексті або будь-якій їх комбінації). Пошукові слова можуть бути усіче-

ними, наприклад, через відкидання префіксів, та/або суфіксів, та/або закінчень слів.

2. Пошук із використанням відстані між словами. Надається можливість пошуку за словами, розташованими на відстані, що не перевищує вказане число.

3. Пошук за співзвучністю слів. Наприклад, якщо в пошуку зазначено слово "семантичний", то будуть знайдені статті, що містять його словоформи, наприклад, "семантичні", "семантичному" тощо, а також близькі за звучанням слова, як-от "генетичний", "статичний", "механічний", "органічний" тощо. Можна додатково вказати ступінь співзвучності слів. Наприклад, за збільшення ступеня співзвучності система видає тільки однокореневі слова, а якщо він становить 100%, то будуть знайдені статті, які містять тільки зазначене слово, тобто відбудеться звичайний пошук.

4. Пошук за коефіцієнтом значущості слів чи фраз. Можна вказати, які зі слів та/або фраз є більш значущими (релевантними). Коефіцієнт значущості слів і фраз впливає на порядок розташування статей у результаті пошуку. Спочатку подано статті з найбільш значущими словами / фразами, а потім — із менш значущими. Додатково для кожного слова / фрази можна позначити ступінь їхньої значущості.

Для зручності використання пошукової мови в NASPLIB вирізняють три рівні: простий, розширений і професійний.

*Простий пошук* — це формулювання пошукових слів і фраз із додатоком використання групових символів, а також засобів пошуку із зазначенням відстані між словами та співзвучності слів.

*Розширений пошук* — це додаткове використання можливостей уточнення слів і фраз метаданими, в яких ці слова / фрази мають відшукуватися, а також логічних зв'язок (І, АБО, НЕ) для формулювання складніших пошукових виразів.

*Професійний пошук* передбачає ознайомлення із синтаксисом пошукової мови та здатність сформулювати інформаційний запит із використанням цього синтаксису.

#### **Адміністрування NASPLIB**

Адміністрування — це функції підтримання системи у працездатному стані. До них належать:

— встановлення, конфігурування та налаштування системи;

— аналіз роботи системи й підтримання ефективності та якості її функціонування;

— забезпечення цілісності та безпеки, у тому числі виконання резервного копіювання й відновлення;

— надання користувачам прав доступу до IP;  
 — введення ЕБ в інші інтеграційні системи інформаційного обслуговування користувачів.

### **Статистика функціонування NASPLIB**

Статистична інформація функціонування ЕБ потрібна різним особам і для різних цілей:

— відповідальному керівництву — для ухвалення організаційно-технічних, економічних, фінансових, юридичних та інших рішень із метою, наприклад, визначення кількісного та якісного складу ЕБ, перспектив подальшого функціонування ЕБ, введення додаткових ресурсів або інтеграції з іншими IP;

— адміністраторам ЕБ — для аналізу функціонування системи;

— користувачам — для аналізу своєї діяльності в ЕБ.

Джерелами збору статистичних даних функціонування NASPLIB є такі:

— механізми збору статистичних даних, вбудовані в ПЗ DSpace;

— під'єднання до системи статистичного аналізу Google Analytics (<https://analytics.google.com/>);

— використання програми, що здійснює збирання статистичних даних щодо функціонування сервера NASPLIB.

Наводять стислі статистичні дані таких категорій:

- кількісні характеристики;
- відвідуваність і відвідувачі;
- використання сторінок;
- пошук;
- перегляди та завантаження.

Наприклад, на момент написання цієї статті на домашній сторінці NASPLIB можна було побачити статистику поточного стану: 484 журнали, 9491 випуск, 183 331 стаття, 128 921 689 переглядів, 129 718 847 завантажень.

### **Інтеграція NASPLIB**

Принципово важливою особливістю сучасних ЕБ є можливість їх інтеграції в єдині системи інформаційного обслуговування користувачів. У найпростішій формі така інтеграція передбачає по дні нижче крок. Кожа з ЕБ створює і підтримує власні IP, а також організує до них доступ. Однак за наявності десятків і навіть сотень таких ЕБ користувачу для пошуку потрібних IP слід звертатися в кожну з них. Щоб спростити інформаційний пошук, створюють єдиний сайт / портал, що інтегрує всю описову інформацію (метадані) наявних ЕБ і надає можливість єдиного пошуку. У разі потреби знайдені документи завантажуються для перегляду та можливого збереження, але вже безпосередньо з

сайтів конкретних власників ЕБ. NASPLIB реалізує цю технологію інтеграції на основі OAI-PMH. Завдяки цьому NASPLIB інтегровано в такі міжнародні реєстри та агрегатори:

— Registry of Open Access Repositories — ROAR (<http://roar.eprints.org/>);

— Directory of Open Access Repositories — OpenDOAR (<http://www.opendoar.org/>);

— OpenAIRE (<https://www.openaire.eu/>);

— Bielefeld Academic Search Engine — BASE (<https://base-search.net/>).

— Connecting REpositories — CORE (<https://core.ac.uk/>).

ROAR — реєстр репозитаріїв відкритого доступу — міжнародна база даних із можливістю пошуку, що індексує створення, розташування й зростання інституційних репозитаріїв відкритого доступу та їхній вміст. ROAR розробили фахівці компанії EPrints в Університеті Саутгемптона, (Велика Британія) 2003. Нині в ROAR зареєстровано понад 5600 інституційних і міжінституційних репозитаріїв.

OpenDOAR — глобальний каталог ЕБ відкритого доступу з гарантованою якістю, які надають вільний доступ до академічних результатів і ресурсів. Кожен запис репозитарію в OpenDOAR ретельно перевірили та опрацювали члени редакційної команди, що гарантує якість і надійність послуг. Сервіс був запущений 2005 р. як результат спільного проєкту Ноттінгемського й Лундського університетів.

Сьогодні ROAR та OpenDOAR є провідними світовими каталогами репозитаріїв відкритого доступу.

OpenAIRE — це мережа постачальників даних відкритого доступу, які підтримують політику відкритого доступу. Агрегатор містить майже 150 млн документів. Попри те, що в минулому проєкт був зосереджений на європейських репозитаріях, нещодавно його розширили, інтегрувавши інституційні та предметні репозитарії з-за меж Європи. Основним напрямом OpenAIRE є допомога Європейській раді з моніторингу дотримання її політики відкритого доступу. Дані ресурсу доступні через API. NASPLIB зареєстровано у цій системі від 2018 р. із рівнем сумісності v2.0 з інструкціями OpenAIRE (<https://guidelines.openaire.eu/>). Наразі актуальним є завдання оновлення системи до рівня сумісності v3.0 або v4.0, оскільки нова інтегрована платформа каталогу EOSC Portal Catalog і Marketplace обслуговуватиме лише зареєстровані джерела даних OpenAIRE, які сумісні з версіями інструкцій OpenAIRE 3.0 та 4.0. Це дасть змогу отримувати додаткові

послуги, наприклад звіти про використання результатів досліджень. OpenAIRE збирає такі дані, об'єднує їх і надає стандартизовані звіти.

*BASE* — глобальна служба збирання метаданих. В її межах акумульовано близько 400 млн описів документів від 11 тис. постачальників даних. *BASE* збирає репозитарії й журнали через OAI-PMH і надає зібраний вміст через API та набір даних. *NASPLIB* під'єднана до цього агрегатора як провайдер даних. У джерелах [6; 7] наведено приклади представлення в агрегаторі *BASE* статей і провайдера даних.

*CORE* — агрегатор Відкритого університету, США, що об'єднує наукові статті у відкритому доступі від тисяч постачальників даних з усього світу, зокрема й інституційні та предметні репозитарії, журнали відкритого та гібридного доступу. *CORE* є найбільшою колекцією наукової літератури відкритого доступу. Наприклад, на момент написання статті агрегатор забезпечував єдину точку доступу до майже 300 млн описів документів та приблизно 10% повних текстів, і колекція постійно зростає. *CORE* надає кілька способів доступу до своїх даних для користувачів і машин, охоплюючи вільний API та повний дамп даних.

Наразі маємо дві електронні бібліотеки, зареєстровані у *CORE*: *NASPLIB* (<https://core.ac.uk/data-providers/944>) та ЕБ Інституту програмних систем НАНУ (<https://core.ac.uk/data-providers/953>), що працює на програмному забезпеченні EPrints. Сторінку представлення повного тексту документа з *NASPLIB* в агрегаторі *CORE* уміщено за адресою: <https://core.ac.uk/reader/132545318>.

#### *Джерела даних CORE*

Станом на квітень 2024 р. *CORE* агрегував контент із 11 162 джерел даних, що охоплюють інституційні репозитарії, академічні видавництва (Elsevier, Springer), журнали відкритого доступу, предметні репозитарії, у тому числі ті, що містять електронні версії (arXiv, ZENODO, PubMed Central) та агрегатори, наприклад, *DOAJ* (<https://doaj.org/>). Під час підрахунку загальної кількості постачальників даних у *CORE* агрегаторів і видавців розглядають як одне джерело даних, попри те, що окремі з них агрегують дані з багатьох джерел. Повний список усіх постачальників даних доступний на вебсайті *CORE*.

Одним із найпотужніших провайдерів даних у *CORE* є БД Crossref — авторитетний показник ідентифікаторів DOI (The Digital Object Identifier), що є обов'язковим міжнародним цифровим ідентифікатором наукової публікації. DOI визначає постійне місце розташування наукової праці

(об'єкта) в інтернеті, її назву та метадані. DOI є обов'язковим складником сучасної системи наукової комунікації, він спрощує процедуру та облік цитування, пошук та локалізацію наукової публікації. Користувач може присвоїти DOI будь-якій публікації, наприклад науковій статті, монографії, главі монографії, дисертації, автореферату, рецензії, підручнику, методичним розробкам / рекомендаціям, звіту, препринту й навіть окремо таблиці, рисунку, схемі, зображенню тощо. DOI призначають публікації один раз і назавжди. Це забезпечує стабільність посилання на публікацію в інтернеті та спрощує пошук потрібної інформації.

#### *Процес збирання документів в агрегаторі CORE*

Процес збирання можна описати як послідовність етапів: на кожному виконується певна дія, і вихідні дані одного етапу передаються на вхід наступному [8]. Вхідними даними для цього процесу є набір постачальників даних, а кінцевим виходом — система, заповнена записами дослідницьких праць. Основні типи завдань, які виконуються в межах системи збирання документів, такі:

1. *Завантаження метаданих.* Метадані, надані постачальником даних за протоколом OAI-PMH, завантажуються та зберігаються у файловій системі (зазвичай XML-файли). Процес завантаження є послідовним, тобто репозитарій надає від 100 до 1000 записів метаданих на запит і маркер продовження. Відтак цей маркер використовують для надання наступної партії записів. Через це процес повного збирання може потребувати багато часу (годин-днів) для великих постачальників даних, однак його було впроваджено для забезпечення стійкості до комунікаційних збоїв.

2. *Витяг метаданих* аналізує, очищає, узгоджує завантажені метадані та зберігає їх у внутрішній структурі даних. Процес узгодження та очищення має враховувати 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).

3. *Завантаження повного тексту.* Використовуючи посилання, отримані з метаданих, *CORE* намагається завантажити та зберегти повні тексти публікацій. Цей процес є нетривіальним. OAI-PMH, як зазначалося, наразі є стандартним протоколом для обміну даними між сховищами. Спочатку його розробили лише для збирання метаданих, однак згодом, через широке засто-



сування та брак альтернатив, використовували як точку входу для збирання й повного тексту. Цей процес відбувається завдяки витягу URL-адрес із записів метаданих, зібраних за допомогою OAI-PMH; витягнуті URL-адреси використовують для визначення місця розташування фактичного ресурсу. Водночас протокол OAI-PMH безпосередньо підтримує лише збирання метаданих, тобто слід реалізувати додаткові функції, щоб використовувати його для збирання вмісту. Розташування посилань на повні тексти в метаданих не стандартизовано, і записи метаданих зазвичай містять кілька посилань. З метаданих незрозуміло, яке з посилань вказує на описане представлення ресурсу, і в багатьох випадках жодне з них не робить цього безпосередньо. Через це всі можливі посилання на ресурс потрібно здобути з метаданих і перевірити, щоб визначити правильний ресурс. Окрім того, OAI-PMH не дає змоги перевірити, чи виявлений ресурс насправді є описаним ресурсом. Подолання цих проблем сприяло прийняттю формату метаданих RIOXX (<https://riox.net/>) або рекомендацій OpenAIRE. Водночас питання однозначного зв'язку записів метаданих та описаного ресурсу залишається актуальним.

4. *Витяг інформації.* Звичайний текст із завантажених документів витягують та опрацьовують для створення напівструктурованого представлення. Цей процес охоплює низку завдань зі здобуття інформації, як, наприклад, витяг посилань.

5. *Збагачення* відбувається через доповнення метаданих і повного тексту, отриманих від постачальників даних, додатковими даними з різних джерел. Деякі збагачення здійснюють безпосередньо через конкретні завдання в процесі збирання документів, наприклад, визначення мови й типу документа (стаття, презентація, дипломна робота, інший). Збагачення зазвичай передбачають застосування моделей машинного навчання та інструментів на основі правил для збирання додаткової інформації про записи, і для певного запису виконуються лише один раз. Решту збагачень, які охоплюють зовнішні набори даних, виконують ззовні, незалежно від процесу збирання, і вбудовують у набір даних. Процес відбувається через збирання різноманітної інформації з великих сторонніх наукових наборів даних. Інформація містить метадані, які необов'язково змінюються, наприклад, ідентифікатор DOI, а також метадані, які зазнають змін, як-от кількість цитувань. Саме через це такі збагачення виконують періодично, тобто всі записи в CORE

проходять через цей процес повторно за певні проміжки часу. Початкове відображення запису здійснюють за допомогою DOI, якщо він доступний. Однак, оскільки більшість записів із репозитаріїв не мають DOI, відбувається процес зіставлення з базою даних Crossref із використанням підмножини полів метаданих, охоплюючи назву, авторів і рік публікації. Після того, як зіставлення виконано, можна узгодити поля, а також зібрати широкий спектр додаткових корисних даних із відповідних зовнішніх БД, збагачуючи запис CORE. Такі дані містять ідентифікатори ORCID, інформацію щодо цитувань, додаткових посилань на вільно доступні повні тексти, галузі дослідження тощо.

5. *Індексування.* Останнім кроком збирання даних є індексування зібраних даних. Отриманий індекс забезпечує роботу служб CORE, у тому числі пошук, API і FastSync.

#### *Інші бази публікацій*

Окрім сервісів агрегації ресурсів відкритого доступу, є низка інших ресурсів для пошуку та завантаження наукової літератури [8]. Їх можна згрупувати у дві категорії: 1) індекси цитування; 2) академічні пошукові системи та наукові графи. Двома основними індексами цитування є Scopus від Elsevier [9] і Web of Science від Clarivate [10], які є послугами передплати преміумкласу. Google Scholar, найвідоміша академічна пошукова система, не надає API для доступу до своїх даних і не дозволяє сканувати свій вебсайт. Semantic Scholar (<https://www.semanticscholar.org/>) є відносно новою академічною пошуковою службою, що має на меті створення "інтелектуальної академічної пошукової системи". Dimensions (<https://www.dimensions.ai/>) — сервіс, орієнтований на аналіз даних, що об'єднує публікації, гранти, політичні документи та показники.

#### *Сервіси інтелектуального аналізу даних*

Особливої уваги варті сервіси, які надаються / плануються власне агрегаторами (або із залученням сторонніх організацій) на основі сучасних досліджень інтелектуального аналізу тексту та даних наукової літератури. До таких сервісів належать: виявлення плагіату нещодавно надісланих публікацій; семантометрики; підбір публікаціям відповідних рецензентів; аналіз науководослідних трендів; рекомендації щодо роботи зі співавторами, проведення заходів тощо; ідентифікація різних версій однієї і тієї самої статті (наприклад, препринти та постпринти); визначення відповідності публікації умовам проведення заходу; визначення важливості, настанов чи типу цитування; узагальнення результатів досліджень;



побудова індексу цитувань; витяг або добування важливих слів чи фраз; категоризація публікацій за областями досліджень; визначення типу публікації; витяг цитат для бібліометрії.

### ***VuFind — відкрите рішення для інтеграції бібліотечних колекцій***

Під час реалізації проєкту "Методи та засоби інтеграції наукових інформаційних ресурсів України у світовий / європейський науковий інформаційний простір і в загальноєвропейську хмару відкритої науки (EOSC)" було здійснено дослідження зі створення системи інтеграції ресурсів із різних відкритих академічних джерел для здобуття зручного й потужного інструменту пошуку та доступу до інформації для користувачів, а також інтеграції зібраних наукових ресурсів України у світовий / європейський науковий інформаційний простір.

Перед системою інтеграції стоять три основні завдання:

1. Збирання й інтеграція метаданих із різних джерел електронних ресурсів.
2. Організація пошуку та видавання відповідних ресурсів.
3. Передавання метаданих із власного агрегатора іншим агрегаторам.

Автори статті проаналізували низку програмних систем, платформи яких придатні для виконання окреслених завдань, і обрали систему VuFind [11]. Її вже розгорнуто та розпочато

експлуатацію (<https://harvester.nas.gov.ua/>). Наразі до VuFind під'єднано 15 електронних бібліотек (провайдерів даних), з яких було зібрано майже 300 тис. метаданих документів, здебільшого статей. Система індексує метадані документів всіх видів академічно релевантних ресурсів, як-от журнали, інституційні репозитарії, цифрові колекції тощо, котрі використовують інтерфейс ОАІ для надання контенту. Проіндексовані дані зберігаються на серверах Інституту програмних систем НАНУ. Апробовано під'єднання та передавання метаданих з VuFind до іншого агрегатора. Вивчено можливості системи з перегляду, пошуку та завантаження статей; підготовлено інструкцію для кінцевого користувача; розроблено схеми метаданих основних видів наукових інформаційних ресурсів НАН України.

**Висновки.** Нині фахівці проводять роботу з оновлення програмного забезпечення зазначеної бібліотеки та налагодження агрегатора Академії наук України, відпрацьовують алгоритми завантаження та оновлення даних, що надходять з різних академічних джерел, регламент роботи, вимоги та інструкції з під'єднання для провайдерів даних.

Вивчення та впровадження світового досвіду з експлуатації таких бібліотек і агрегаторів сприятиме вдосконаленню дослідженої системи та приведенню її у відповідність до сучасних зразків.

### *Список бібліографічних посилань*

1. Захарова О. В., Захарова Е. Г., Резніченко В. А. Каталог програмних засобів створення електронних бібліотек / Ін-т програмних систем НАН України. Київ, 2006. 32 с.
2. Резніченко В. А., Проскудіна Г. Ю., Овдій О. М., Дорошенко А. Ю. Створення цифрових бібліотек на основі Greenstone. *Проблеми програмування*. 2005. № 2. С. 24—41. URL: <http://dspace.nbuv.gov.ua/handle/123456789/2468>.
3. Новицкий А. В., Кудим К. А., Резниченко В. А., Проскудіна Г. Ю. Создание научных архивов с помощью системы EPrints. *Проблеми програмування*. 2007. № 1. С. 46—60. URL: <http://dspace.nbuv.gov.ua/handle/123456789/275>.
4. Кудим К. А., Проскудіна Г. Ю., Резниченко В. А. Создание научных электронных библиотек с помощью системы DSPACE. *Проблеми програмування*. 2007. № 3. С. 49—60. URL: <http://dspace.nbuv.gov.ua/handle/123456789/300>.
5. DSpace Release 7.0 Status. *DSpace*. URL: <https://wiki.lyrasis.org/display/DSPACE/DSpace+Release+7.0+Status#DSpaceRelease7.0StatusWhatfeaturesarecominginalater7.xrelease?7.x>.
6. BASE. URL <https://www.base-search.net/Search/Results?q=coll:ftvarnadskynl&refid=dctableen>.
7. BASE. URL [https://www.base-search.net/about/en/about\\_source.php?id=2328](https://www.base-search.net/about/en/about_source.php?id=2328).
8. Кнот П., Херрманнова Д., Канселлієрі М. [та ін.]. (2023). CORE: Глобальна служба агрегації документів відкритого доступу. *Nature Scientific Data*. Vol. 10. № 366 (2023). doi: <https://doi.org/10.1038/s41597-023-02208-w>.
9. Scopus: Comprehensive, multidisciplinary, trusted abstract and citation database. *Elsevier*. URL: <https://www.elsevier.com/solutions/scopushttps://www.elsevier.com/solutions/scopus>.
10. Web of Science platform. *Clarivate*. URL: <https://clarivate.com/webofsciencegroup/solutions/web-of-science/>.
11. Проскудіна Г. Ю., Кудім К. О., Резніченко В. А. VuFind: відкрите рішення для інтеграції бібліотечних колекцій. *Проблеми програмування*. 2023. № 4. С. 15—26. URL: <https://pp.isoftware.kiev.ua/ojs1/article/view/590>.

**Valeriy Reznichenko,**  
*PhD of Physical and Mathematical Sciences,*  
*Senior Research Fellow,*  
*Institute of Software Systems of the National Academy of Sciences of Ukraine*

**Galyna Proskudina,**  
*Researcher,*  
*Institute of Software Systems of the National Academy of Sciences of Ukraine*

**Kuzma Kudim,**  
*Junior Research Fellow,*  
*Institute of Software Systems of the National Academy of Sciences of Ukraine*

**A modern approach to the creation and integration  
of library digital information resources**

The paper presents an overview of the software and functionalities of the Scientific Digital Library of Periodicals of the National Academy of Sciences of Ukraine. It also considers the issues of integration of library systems into unified user service systems on the example of modern global aggregators of open access documents BASE, OpenAIRE and CORE. Their quantitative characteristics, such as the number of collected document descriptions and full texts, the number of data providers, and the availability of an application programming interface for data retrieval are analyzed. The composition and types of their data providers, such as institutional repositories, open journals, publishing houses, scientific repositories of preprints, thematic digital libraries, and systems that are also aggregators, are considered. We also investigated 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 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 the collected data is updated. The article provides a list of useful services for researchers that can be provided by modern aggregators based on collected data corpora, using machine learning methods, bibliometrics, webometrics, altmetrics, and semantometrics. As developers of a number of open access scientific digital libraries, we are already registered as data providers in these systems. At present, work is underway to create an aggregator of scientific resources of the National Academy of Sciences of Ukraine, so studying the world experience is a necessary step to solve this problem at the proper level.

**Keywords:** digital libraries; distributed documentary information systems; integration; open access; service provider; data provider; OAI-PMH protocol

### References

1. Zaharova O. V., Zaharova E. G. & Reznichenko V. A. (2006). *Katalog programnih zasobiv stvorenniya elektronnih bibliotek*. Kyiv.
2. Reznichenko V. A., Proskudina G. Yu., Ovdij O. M. & Doroshenko A. Yu. (2005). Stvorenniya cifrovih bibliotek na osnovi Greenstone. *Problemi programuvannya*, 2, pp. 24—41. Available at: <http://dspace.nbuv.gov.ua/handle/123456789/2468>.
3. Novickij A. V., Kudim K. A., Reznichenko V. A., Proskudina G. Yu. Sozdanie nauchnyh arhivov s pomoshyu sistemy EPrints. *Problemi programuvannya*, 1, pp. 46—60. Available at: <http://dspace.nbuv.gov.ua/handle/123456789/275>.
4. Kudim K. A., Proskudina G. Yu., Reznichenko V. A. Sozdanie nauchnyh elektronnyh bibliotek s pomoshyu sistemy DSPACE. *Problemi programuvannya*, 3, pp. 49—60. Available at: <http://dspace.nbuv.gov.ua/handle/123456789/300>.
5. *DSpace Release 7.0 Status*. *DSpace*. (n. d.). Available at: <https://wiki.lyrasis.org/display/DSpace/DSpace+Release+7.0+Status#DSpaceRelease7.0StatusWhatfeaturesarecominginalater7.xrelease?7.x>.
6. *BASE*. (n. d.). Available at <https://www.base-search.net/Search/Results?q=coll:ftvarnadskynl&refid=dctableen>.
7. *BASE*. (n. d.). Available at [https://www.base-search.net/about/en/about\\_source.php?id=2328](https://www.base-search.net/about/en/about_source.php?id=2328).
8. Knot P., Herrmannova D., Kanselliyeri M. [ta in.]. (2023). CORE: Globalna sluzhba agregaciyi dokumentiv vidkritogo dostupu. *Nature Scientific Data*, 10, 366 (2023). doi: <https://doi.org/10.1038/s41597-023-02208-w>.
9. Scopus: Comprehensive, multidisciplinary, trusted abstract and citation database. (n. d.). *Elsevier*. Available at: <https://www.elsevier.com/solutions/scopushttps://www.elsevier.com/solutions/scopus>.
10. Web of Science platform. *Clarivate*. (n. d.). Available at: <https://clarivate.com/webofsciencegroup/solutions/web-of-science/>.
11. Proskudina G. Yu., Kudim K. O. & Reznichenko V. A. (2023). VuFind: vidkrite rishennya dlya integraciyi bibliotechnih kolekcij. *Problemi programuvannya*, 4, pp. 15—26. Available at: <https://pp.isofts.kiev.ua/ojs1/article/view/590>.

Надійшла до редакції 27 травня 2024 року