УДК 004.7 : 021

Тетяна ШЕРЕПА,
мол. наук. співробітник НБУВ

Створення системи індексування інтранет-середовища наукової бібліотеки

Обґрунтовано необхідність архівування наукових публікацій та створення електронних журналів відкритого доступу бібліотеками, які мають взяти на себе функції тематичних та інституційних репозитаріїв. Визначено переваги інтранет-архівів мережевих публікацій перед їхнім інтернет-аналогом. Визначено теоретичні засади технології індексування науково-інформаційних ресурсів інтранет-середовища та методики побудови довідково-пошукового апарату інтранет-архіву.

К л ю ч о в і с л о в а: архівування, електронні колекції, інтранет, пошуковий апарат, індексування, Dublin Core, WWWISIS.

В умовах інтенсивного розвитку глобальних комп'ютерних мереж суспільна потреба забезпечення вільного доступу до джерел інформації та знань обумовила появу ініціатив «самоархівування наукових публікацій» та створення електронних журналів відкритого доступу. Бібліотеки є найбільш активними захисниками Open Access, оскільки відкритий доступ до інформації – це один із центральних принципів їхньої діяльності [1]. У провідних закордонних країнах розгорнута активна робота з реалізації проектів створення архівів мережевих науково-інформаційних ресурсів та організації їх збереження й використання в стінах бібліотек.

Стратегія самоархівування включає в себе розміщення авторами електронних версій власних наукових публікацій у загальнодоступних архівах електронних документів, у вільному доступі в інтернеті. Самоархівування дозволяє підвищити ефективність використання результатів наукових досліджень завдяки вільному доступу до наукових матеріалів. Публікації, які архівуються, повинні розташовуватися переважно в тематичних або інституційних репозитаріях (архівах). Під «архівом» розуміється сайт, який зберігає джерела наукової інформації у відкритому мережевому доступі [2].

Другою стратегією є журнали відкритого доступу – нове покоління журналів, які беруть на себе зобов'язання про відкритий доступ, не обмежують доступ до матеріалів, які вони публікують, та не беруть плату за їх використання [3].

«Державною програмою розвитку діяльності Національної бібліотеки України імені В. І. Вернадського на 2005–2010 рр.» (затвердженою Постановою Кабінету Міністрів України від 25 серпня 2004 р. № 1085) передбачено завдання збору й архівації наукової суспільно значущої інформації та створення Українського науково-інформаційного порталу з розвинутою системою пошуку.

З метою розширення доступу до наукових матеріалів НБУВ формує архів, який містить колекції наукових інтернет-публікацій, книг та інших документів України та про Україну, який доступний локальним користувачам комп'ютерної мережі НБУВ. На сьогодні існує необхідність у створенні швидкої, гнучкої, інтелектуальної пошукової системи на базі індексування колекцій документів, веб-сторінок або файлів інших форматів для задоволення інформаційних потреб користувачів [4]. Програмні засоби системи мають відповідати концепції вільного поширення, забезпечувати інтелектуальний пошук інформації, надавати користувачу типовий веб-орієнтований інтерфейс.

Метою даної статті є визначення теоретичних засад технології індексування науково-інформаційних ресурсів інтранет-середовища та розробки методики побудови довідково-пошукового апарату інтранет-архіву.

Розміщення публікації на веб-сервері автора у

[1] *Suber P.* Removing the Barriers to Research: An Introduction to Open Access for Librarians [Electronic resource]. – Way of access: URL: http://www.earlham.edu/~peters/writing/acrl.htm. – Title from the screen.

[2] *Негуляев Е. А.* Самоархивирование [Электронный ресурс]. – Режим доступа: http://ellib.gpntb.ru/ntb/2004/12/ntb_12_9_2004.htm. – Загл. с экрана.

[3] Будапештская инициатива «Открытый доступ» [Электронный ресурс]. – Режим доступа: http://www.soros.org/openaccess/ru/read.shtml. – Загл. с экрана.

[4] *Копанєва В. О.* Архівування науково-інформаційних ресурсів Інтернет: основні концептуальні положення // Бібліотечний вісник. – 2005. – № 2. – С. 14–19.

вільному доступі не є бажаним для ідеї архівування наукових матеріалів, тому що звичайне веб-середовище не може забезпечити надійної ідентифікації метаданих та організації пошуку за ними, а також не є придатним для довготривалого збереження і гарантії незмінності публікацій. Суттєва перевага відкритих тематичних архівів електронних публікацій полягає в тому, що їх збір та впорядкування здійснюється спеціалістами. Наслідком є забезпечення фільтрації та пошуку даних із вищим рівнем точності, тому що процес індексування таких систем є глибшим за його інтернет-аналог.

З метою уніфікації представлення мережевих ресурсів розроблені єдині принципи їх опису, які базуються на використанні метаданих Дублінського ядра [5]. Основною вимогою до репозитарію є підтримка протоколу OAI PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), який забезпечує можливість збору структурованих метаданих про об'єкти, розміщених у репозитарії, об'єднання з іншими репозитаріями й організацію пошуку в розподілених репозитаріях відкритого доступу [6].

Усі нові електронні документи, які підлягають архівуванню, мають пройти процес індексування. Мета процесу індексування в документальних системах аналогічна меті каталогізації у бібліотеках: надати кожній одиниці зберігання деяку множину ідентифікаторів, які б відображали зміст документа. В традиційних бібліотеках у ролі ідентифікаторів змісту виступають відповідні шифри, які визначають предметну класифікацію і місце зберігання документа. З розвитком автоматичної обробки документів звичайний процес каталогізації трансформується в процес індексування, котрий призначений для надання кожному елементу ідентифікаторів, які також називають індексаційними термінами, ключовими словами, дескрипторами. Усі ці терміни відображають зміст документа і керують пошуком, вибираючи ті документи, терміни яких є найбільш схожими з термінами пошукового запиту.

Зважаючи на великі обсяги інтранет-архівів, проведення ручної класифікації та індексації кожного електронного документа не є можливим, тому що як одиницю обліку фонду інтернет-документів (веб-ресурсів) зручно використовувати веб-сайт чи його фрагмент. Процес комплектування фонду полягає в створенні в бібліотеці копій («дзеркал») веб-сайтів. Оскільки інформація на веб-сайті змінюється з часом, бібліотека повинна створювати «дзеркала» того самого сайта періодично.

Автоматичне індексування базується на текстах вихідних документів, або, принаймні, на фрагментах текстів, таких, як заголовки або реферати. Більшість результатів автоматичного індексування не є досконалими, але мають певні переваги перед ручним індексуванням [7]: ефективність пошуку щодо видачі релевантних документів, одержаних автоматичними методами є несуттєво меншою, ніж при ручному індексуванні цих документів; однак вартість автоматичного індексування та витрачання часу висококваліфікованого персоналу значно скорочується.

Існує декілька безкоштовних інтранет пошукових систем, які забезпечують задоволення інформаційних запитів інтранет-користувачів. Ці системи розроблено для індексування внутрішніх веб-серверів і/або фрагментів цих серверів та створення потрібних пошукових індексів документів, які розміщені на серверах.

Такі пошукові системи можуть бути згруповані за такими категоріями:

1) технічна функціональність: платформа сервера, веб-сервер, відкритість коду, можливість подальшого розвитку системи та ін.;
2) особливості індексування: формати файлів (HTML, PDF тощо), рівень індексування (запис, файл, директорія), розпізнавання стандартних форматів (MARC та ін.), виділення термінів зі спільним коренем, наявність стоп-словника та ін.;
3) особливості пошуку: підтримка булевих операторів, нечіткий пошук, пошук фраз, використання тезаурусів синонімів та ін.;
4) відображення результатів: формати виводу, ранжування результатів, підсвітлювання ключових слів у контексті та ін.;
5) ціна, вимоги ліцензії та реєстрації.

Вибір пошукової системи інтранет має також враховувати ознайомлення інформаційних спеціалістів із доступними продуктами і аспектами їх використання, знання технологій інформаційного пошуку, розуміння та досвід роботи зі стандартними практиками і параметрами індексування, що

[5] Dublin Core Metadata Initiative / [Electronic resource]. – Way of access: URL: http://dublincore.org/. – Title from the screen.

[6] Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) [Electronic resource]. – Way of access: URL: http://www.openarchives.org/pmh/. – Title from the screen.

[7] *Дж. Солтон.* Динамические библиотечно-информационные системы. – М.: Мир, 1979. – 558 с.

забезпечить точний і ефективний пошук даних на базі різноманітних індексів даних інтранету [8].

З урахуванням вищеназваних вимог для апробації довідково-пошукового апарату інтранет-середовища наукової бібліотеки нами було обрано WWWISIS, яка є однією з безкоштовних пошукових систем для бібліографічних та текстових баз даних. WWWISIS підтримує всі функції пакету прикладних програм CDS/ISIS (Computer Documentation System / Integrated System Information Services), що є універсальним інструментарієм для створення автоматизованих систем бібліотек, архівів і музеїв, тобто для обробки структурованих нечислових баз даних [9].

Всесвітньо відома програма CDS/ISIS, яка поширюється UNESCO, добре себе зарекомендувала в діяльності бібліотек, де має місце робота з великою кількістю текстової інформації. Порівняно з іншими програмними системами аналогічного призначення CDS/ISIS має такі переваги:

- порівняно невеликий обсяг системи;
- високу швидкодію системи;
- орієтація на роботу зі стандартними форматами;
- підтримка системою записів змінної довжини;
- широкі можливості в здійсненні пошуку;
- безкоштовне поширення даної системи.

CDS/ISIS із рядом спеціальних утиліт і доробок дозволяє отримати сучасне програмно-технологічне середовище для операційних систем UNIX, WINDOWS, що за функціональними можливостями і сервісу користувачів може бути порівняне з сучасними реляційними СУБД типу ORACLE і INFORMIX.

Пошукова система на базі CDS/ISIS підходить як для колекцій, які містять невелику кількість документів, так і для більших колекцій. Використовуючи програму послідовного перегляду документів колекції та застосовуючи відповідні фільтри, можна індексувати файли текстового формату, HTML чи будь якого іншого формату, що може бути конвертований у HTML-формат. Використання бази формату CDS/ISIS та CGI-модуля WWWISIS дозволяє здійснювати швидкий повнотекстовий пошук за допомогою веб-орієтованого інтерфейсу.

Головною особливістю CDS/ISIS є автоматичне створення й підтримка файлів швидкого доступу («індексних файлів») до кожної бази даних, що забезпечує максимальну швидкість пошуку навіть за великих обсягів даних. Ці файли називаються словником пошукових термінів і вміщують усі терміни, які можуть бути використані під час пошуку в базі даних [10]. Структура інвертованого файлу забезпечує його швидку модифікацію при долученні в колекцію нових документів. Можлива побудова словника стоп-слів з орієнтацією на вилучення другорядних частин мови та загальновживаних слів, вилучення яких не вплине на якість пошуку, більше того може його покращити.

Використання пакета прикладних програм CDS/ISIS як основи пошукової системи інтранет-архіву забезпечує її однорідність із пошуковою системою електронних колекцій бібліотек НБУВ, котра містить наступні інформаційно-ресурсні компоненти: електронний каталог НБУВ, загальнодержавну реферативну базу даних, фонд електронних документів із повними текстами. Головною засадою побудови системи архівування науково-інформаційних ресурсів НБУВ є технологія збору та підготовки тематичних складових інтранет-архіву.

В якості інформаційної бази для структури метаданих архіву обрано стандарт Дублінського ядра метаданих (Dublin Core Metadata), запропонованого Онлайновим комп'ютерним бібліотечним центром OCLC для опису ресурсів інтернету [11]. Формат Dublin Core влючає 15 елементів для опису електронного ресурсу:

- назва (title);
- автор (creator);
- предметна рубрика (subject);
- анотація (description);
- видавець (publisher);
- співавтор (contributor);
- дата (date);
- формат (format);
- тип (type);
- ідентифікатор (identifier);
- джерело (source);
- мова (language);
- відношення (relation);
- покриття (coverage);
- авторські права (rights).

За правилами Dublin Core кожний із 15 елемен-

[8] Indexing and Search Engines for the Web (WISE). Search Engines for Intranets: An overview/ [Electronic resource]. – Way of access: URL: http://www.ncsi.iisc.ernet.in/raja/netlis/wise/search/search.html. – Title from the screen.

[9] UNESCO CDS-ISIS databases [Electronic resource]. – Way of access: URL: http://www.unesco.org/. – Title from the screen.

[10] Ibid.

[11] Dublin Core Metadata Initiative / [Electronic resource]. – Way of access: URL: http://dublincore.org/. – Title from the screen.

тів не є обов'язковим і може повторюватися. Опис із використанням Dublin Core інтернет-ресурсів можна, в першому наближенні, розглядати як бібліографічний опис книги чи аналітичний розпис журналу (газети).

Індексування архіву може бути повнотекстовим або лімітованим деяким фільтром, що обирається створювачем архіву. Зважаючи на великі розміри інтранет-архіву, доцільно зберігати в індексній базі такі частини HTML сторінок:

- TITLE – заголовок сторінки.
- META NAME=«keywords» CONTENT=«.....», що містить ключові слова і словосполучення. Може включати слова, які не зустрічаються в документі, але мають пряме відношення до тематики сайта, що підвищить релевантність пошуку. В середньому дозволяється вказувати до 150–200 символів як ключових слів.
- META NAME=«description» CONTENT=«.....», що містить тематичний опис сайта.
- Заголовки форматування HTML сторінки H1, H2 та ін.

За допомогою ISIS_DLL, прикладного програмного інтерфейсу ISIS для операційних систем Windows та Linux, котрий розроблений та вільно поширюється UNESCO [12], та мови програмування, яка припускає використання ISIS_DLL, можна отримати доступ до попередньо визначеної частини інтранет-архіву і створити записи відповідного формату в індексній базі.

Інформаційні ресурси інтранет-архіву також можливо долучити до системи електронних видань, програмні засоби якої також підтримують формат CDS/ISIS. Система електронних видань є комплексом галузевих серій колекцій документів. Галузеві серії формуються на основі структуризації наявних інформаційних ресурсів бібліотеки шляхом попереднього відбору документів із бібліографічних, реферативних, тематичних і повнотекстових баз даних, їх обробки та впорядкування. Кожна з галузевих серій має розвинений пошуковий апарат, що забезпечує виявлення потрібних документів за елементами їх бібліографічного опису (автор, назва, вихідні дані тощо), а також за текстами документів [13].

На сьогодні система електронних видань НБУВ містить 6 галузевих серій: природничі, технічні, суспільні та гуманітарні, медичні, аграрні науки, бібліотечна справа та науково-інформаційна діяльність. Галузева серія «Бібліотечна справа та науково-інформаційна діяльність» є колекцією документів формату HTML. На етапі створення серія не містила в собі пошукового апарату. Для доопрацювання колекції існувала необхідність у створенні індексної бази формату CDS/ISIS. Така база даних була створена програмно з використанням ISIS_DLL за розглянутою технологією індексування. Лімітування інформації, яка долучалася до індексної бази, виконано за такими дескрипторами: заголовки вебсторінки (TITLE, H1, H2) та дані мета-тегів (META keywords, META description) відповідно до структури полів Dublin Core: назва (title), предметна рубрика (subject), анотація (description).

До сформованої індексної бази застосовано конфігурацію пошукової системи галузевих серій та веб-доступу на базі пакету прикладних програм CDS/ISIS та CGI-модуля WWWISIS. Таким чином, пошук у розглянутій колекції здійснюється як через веб-сервер, так і на компакт-дисках.

Подальший розвиток інтранет-пошукової системи НБУВ доцільно зорієнтувати в напрямах включення до його складу засобів класифікації й опису інформаційних колекцій документів та веб-сайтів як одиниць зберігання інтранет-архіву, досягнення максимальної ресурсоощадності зберігання електронних колекцій, інтелектуалізації пошукового апарату, семантичного аналізу текстів і творення нових знань.

Висновки

1. Інтенсивний розвиток глобальних комп'ютерних мереж зумовив появу ініціатив «самоархівування наукових публікацій» та створення електронних журналів відкритого доступу, найактивнішими захисниками яких є бібліотеки. Публікації, які архівуються, повинні розташовуватися переважно в тематичних або інституційних репозиторіях (архівах). Звичайне веб-середовище не може забезпечити надійної ідентифікації метаданих та організації пошуку за ними, не є придатним для довготривалого збереження і гарантії незмінності публікацій.

2. Основною вимогою до репозитарію є підтримка протоколу OAI PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), який забезпечує можливість збору структурованих метаданих про об'єкти, які розміщені у репозитарії, об'єднання з іншими репозитаріями, а також орга-

[12] ISIS Application Program Interface ISIS_DLL User's Manual Preliminary Version BIREME, Sao Paulo, July 2001 [Electronic resource]. – Way of access: URL: http://www.bireme.br/. – Title from the screen.

[13] *Шерепа Т. А.* Система галузевих серій електронних видань: основні концептуальні положення // Бібліотечний вісник. – 2004. – № 1. – С. 26–29.

нізації пошуку в розподілених репозитаріях відкритого доступу.

3. Суттєвою перевагою інтранет-архівів мережевих публікацій, які створюються в наукових бібліотеках, є здійснення попереднього збору та впорядкування документів спеціалістами. Наслідком є забезпечення фільтрації та пошуку даних із більш високим рівнем точності. Зважаючи на великі обсяги інтранет-архівів, проведення ручної класифікації та індексації кожного електронного документа не є можливим, тому що як одиницю обліку фонду інтернет-документів (веб-ресурсів) зручно використовувати веб-сайт чи його фрагмент. У рамках веб-сайту є необхідність у проведенні автоматичного індексування, надаючи більше значення індексаційним термінам, які є метаданими або заголовками електронних документів.

4. Головною засадою побудови системи архівування науково-інформаційних ресурсів НБУВ є технологія збору та підготовки тематичних складових інтранет-архіву. Конфігурація пошукової системи інтранет-архіву НБУВ на базі пакету прикладних програм CDS/ISIS із CGI-модулем WWWISIS забезпечує її функціонування, швидкодію, а також однорідність із системою електронних колекцій бібліотек НБУВ.

5. Обґрунтованість і достовірність викладеної технології підтверджено в процесі створення пошукової системи галузевої серії системи електронних видань «Бібліотечна справа та науково-інформаційна діяльність». Пошук у колекції може здійснюватись як через веб-сервер, так і на компакт-дисках.