УДК 681.3

В.А Резніченко, Г.Ю. Проскудіна, О.М.Овдій, А.Ю.Дорошенко

СТВОРЕННЯ ЦИФРОВИХ БІБЛІОТЕК ПЕРІОДИЧНИХ ВИДАНЬ НА ОСНОВІ GREENSTONE

Описуэться процес створення цифрової бібліотеки наукових періодичних видань на основі програмного забезпечення Greenstone. Представлена інформаційна модель наукового періодичного видання; визначені поняття цифрова бібліотека та колекція; наведено стислий огляд програмного забезпечення Greenstone; а також можливості цієї системи по створенню колекцій цифрової бібліотеки на прикладі наукового журналу.

1 ВСТУП

Останніми роками інтенсивно росте кількість і розмаїтість науково-технічних інформаційних ресурсів, представлених в електронному вигляді. Поряд із класичними базами даних, в основному бібліографічними і реферативними, що домінували в інформаційному обслуговуванні науки до середини 1990-х років, наукові установи і служби науково-технічної інформації стали створювати найрізноманітніші колекції наукових документів і даних, розрахованих як на загальне, так і на локальне використання.

Це повнотекстові колекції опублікованих і неопублікованих науково-технічних документів, електронні карти, електронні енциклопедії і довідники, наукові форуми і дискусії, комп'ютерні моделі різних наукових об'єктів, масиви даних, отриманих у результаті експериментів і спостережень та ін. Найбільш загальним терміном, що охоплює, у залежності від трактування, усі або багато видів таких колекцій, став термін "електронні (цифрові) бібліотеки" (ЦБ).

За ініціативи Національного наукового фонду США (NSF), Агенства перспективных досліджень Міністерства оборони (DARPA) і національного агенства по космічним дослідженням (NASA) программа “Digital Libraries Initiative” (DLI) [1] стимулювала розробки в ряді країн, зокрема в Росії [2], інформаційних систем нового класу, ключову роль в яких грають сформовані у 90-роки технології управління інформаційними ресурсами та комунікаційні засоби.

Більшість установ Академії наук мають свої Web-сайти, де крім інформації про установу та основні напрямки наукової діяльності, представлена також інформація про видання, публікації, виконані НДР, зміст наукових журналів з анотаціями, БД, електроні видання і т.д.

Проведений аналіз показав, що існуючі електроні інформаційні ресурси НАН України несистематизовані і розрізнені, як логічно, так і фізично. Інформація в цілому не представлена належним чином для доступу в мережі Інтернет. І хоча в ряді організацій Академії проводиться робота по публікації даних в мережі Інтернет, все ще існує ряд недоліків. Інформація переважно статична, погано структурована, має різнорідні інтерфейси та формати, та не завжди має пошукові засоби. В Україні відсутня єдина загальнонаціональна система дистанційного пошуку наукової інформації та доступу до неї.

Тому існує потреба в наданні науковим колективам Академії наук (особливо гуманітарного напрямку) методики створення ЦБ повнотекстових документів. Головною задачею створення інформаційних систем такого типу є їх інформаційне наповнення, ефективний пошук та одержання існуючих ресурсів, а також наступна інтеграція в єдину інформаційну систему НАН України.

У цій статті для створення ЦБ наукових періодичних видань пропонується використати програмне забезпечення Greenstone. Система являє собою ефективне Open Source-рішення для побудови цифрових бібліотек [3-5]. Вона містить пошук з попереднім

індексуванням по документам усіх популярних форматів, і передусім doc і pdf, які можуть бути представлені у заархівованому вигляді. Система створює каталог документів, конвертує їх у html-формат, а потім забезпечує віддалений доступ до бібліотеки за допомогою броузера.

Стаття містить опис процесу створення ЦБ наукових періодичних видань на основі програмного забезпечення (ПЗ) Greenstone: опис інформаційної моделі наукового періодичного видання (розділ 2); визначення поняття ЦБ (розділ 3); стислий огляд програмного забезпечення Greenstone (розділ 4); можливості цієї системи по створенню колекцій ЦБ на прикладі наукового журналу (розділ 5), перспективи розвитку системи Greenstone (розділ 6).

2 ОПИС ПРЕДМЕТНОЇ ОБЛАСТІ

Предметом нашого аналізу є наукове періодичне видання. В загальному випадку таке видання представляється інформаційними ресурсами наступних трьох рівнів: періодичне видання в цілому, випуск періодичного видання та окрема публікація у випуску. Далі у цьому розділі надається короткий опис цих складових наукового періодичного видання.

2.1. Періодичне видання

Періодичне видання має наступні основні характеристики чи атрибути:

1. *Назва*. Ім'я, що дається періодичному виданню. Може мати кілька значень на різних мовах.
2. *Тематичні розділи і підрозділи.* Повний перелік тем змісту, який виражається за допомогою ключових фраз, ключових слів чи класифікаційних кодів, що описують теми публікації. На практиці значення вибираються з контрольованого словника чи формальної схеми класифікації, наприклад, УДК. Тематичні розділи можуть бути простим лінійним списком або ієрархічно структуровані.
3. *Опис.*Коротке повідомлення про зміст періодичного видання. Опис може бути представлено у вигляді: реферату, змісту, посилання на графічне представлення чи простого текстового викладу змісту.
4. *Тип.*Характер або жанр змісту періодичного видання
5. *Видавець*. Організація, відповідальна за введення періодичного видання в обіг.
6. *Головний редактор*.
7. *Редакційна колегія.*
8. *Дата*. Містить дату заснування періодичного видання.
9. *Періодичність*.
10. *Ідентифікатори* ISBN, ISSN, DOI. Можливо кілька значень.
11. *Мова.* Містить мову інтелектуального змісту періодичного видання.
12. *Адреса.* Містить адресу редакції періодичного видання.
13. *Телефон.*
14. *E-mail.*
15. *URL*. Може мати кілька значень.

Всі ці характеристики розглянуті на прикдаді трьох видань Академії у галузі комп'ютерних наук :"Проблемы программирования", "Математичні машини і системи"та "Кибернетика и системный анализ". Позначемо її як Ресурс 1, 2 і 3 відповідно. (Див. Додаток Табл.1).

2.2 Випуск

Описує конкретний випуск періодичного видання і має наступні атрибути:

1. *Тема випуску*. Можливо кілька значень.
2. *Присвята*.. Наприклад, “До 80-річчя з дня народження В.М. Глушкова”
3. *Дата і номер*.
4. *Зміст*. Може мати кілька значень на різних мовах.

2.3 Публікація

Описує конкретну публікацію (статтю) у випуску журналу. Має наступні атрибути:

1. *Назва*. Кілька значень на різних мовах.
2. *Автор*. Може мати кілька значень.
3. *Тип публікації*. Наприклад, стаття, повідомлення, праця конференції, доповідь на конференції або семінарі.
4. *Бібліографічний опис*.
5. *Мова*.
6. *Повний код УДК*. (Міжнародна система тематичної класифікації публікацій, UDC - Universal Decimal Classification) [6]
7. *Спеціальні ідентифікатори*. Наприклад, ідентифікатори цієї публікації в системі ACM Classification System [7].
8. *Дата і номер*.
9. *Реферат*. Кілька значень, можливо, на різних мовах.
10. *Ключові слова, що характеризують зміст публікації*. Кілька значень на різних мовах.
11. *Номера сторінок*.
12. *Повний текст*. Кілька значень. Може бути представлений файлом або URL.

Приклад опису публікації.

Назва: “Онтології у контексті інтеграції інформації: представленя, методи та інструменти побудови”.
Автор: Овдій О.М., Проскудіна Г.Ю.
Тип публікації: Стаття.
Бібліографічний опис: Овдій О.М., Проскудіна Г.Ю. Онтології у контексті інтеграції інформації: представленя, методи та інструменти побудови // Проблемы программирования. — 2004. — №2-3 — С.353-365.
Мова: українська
Повний код УДК: 681.03
Назва періодичного видання: Проблемы программирования
Дата і номер: 2004, №2-3
Реферат: Розглядається використання онтологій для підтримки задач інтеграції у семантично гетерогенних інформаційних системах. Представлені основні поняття та визначення онтологій, цілі та приклади їх побудови.
Ключові слова, що характеризують зміст публікації: онтологія, об’єднання онтологій, інженерія онтологій, відображення онтологій.

Далі розглянемо, як можна створити інформаційну систему, що відноситься до класу *цифрових бібліотек*, побудувавши колекцію інформаційних ресурсів, наукове видання - випуск – публікація, що мають наведену вище модель.

3 ЩО ТАКЕ ЦИФРОВІ БІБЛІОТЕКИ?

У цей час визначення цифрової бібліотеки не має чіткого формулювання й скоріше являє собою набір побажань по використанню нових технологій в галузі обробки, зберігання й пошуку документів. Під цифровою бібліотекою, найчастіше, розуміється розподілена інформаційна система, що дозволяє надійно накопичувати, зберігати й ефективно використовувати різноманітні колекції електронних документів, доступні в зручному для користувачів вигляді через глобальні мережі передачі даних.

Поняття цифрових бібліотек іде своїми корінням в бібліотечне поширення "знань для всіх" [1]. ЦБ – "це система, яка забезпечує співтовариству користувачів доступ зрозумілим для них чином до великих репозиторіїв мультимедійної інформації та знань, що організовані за відсутністю будь-яких відомостей про їх використання" [8]. Цифрові бібліотеки ламають бар'єри фізичних границь і прагнуть дати доступ до інформації в різних галузях і співтовариствах. Хоча термін "Цифрова Бібліотека" був популяризований на початку 1990-их, він сягає до проектів, що працюють над з'єднанням розподілених систем, автоматизованим зберіганням і добуванням інформації, бібліотечним мережам і зусиллям по оперативному спільному використанню ресурсів. Існує багато визначень терміна "цифрова бібліотека", але всіх їх поєднує - інтеграція технології й політики. Для сучасних систем цифрових бібліотек ця інтеграція забезпечує структуру керування й забезпечення механізмів доступу до інформаційних ресурсів. Технології, що підтримують створення й підтримку цифрової бібліотеки, з'явилися в минуле десятиліття, й призвели до збільшення обчислювальної швидкості й продуктивності, навіть на скромних обчислювальних платформах. Таким чином, майже будь-яка організація або персона, може розглядати встановлення й подання цифрової бібліотеки. Потужність обробки середнього комп'ютера дозволяє одночасно обслуговувати декількох користувачів, дозволяє шифрування (кодування) і дешифрування (декодування) обмежених матеріалів і підтримку складних процесів ідентифікації користувачів і дотримання прав доступу. Збільшення швидкодії мережевого доступу дозволяє надати зміст цифрової бібліотеки всесвітній аудиторії. Скорочення вартості носіїв даних знімає бар'єри до поміщення навіть великих колекцій в оперативний доступ. Звичайні доступні інструменти для створення й подання інформації в різних формах роблять зміст широко доступним без дорогих спеціальних інструментів.

Хоча галузь цифрових бібліотек розвивається в науку, що складається зі знання, теорій, визначень і моделей, необхідно адекватно оцінити успіх цифрової бібліотеки в межах специфічного контексту. Оцінка цифрової бібліотеки вимагає ясного розуміння мети, який вона призначена служити.

Цифрові бібліотеки можна розглядати як організовані, спеціалізовані колекції інформації. Вони сконцентровані на окремому предметі чи темі, і хороші цифрові бібліотеки добре роз’яснюють принципи управління тим, що вони містять. Вони створюються для того, щоб інформація стала доступною, чітко визначеною і будуть включати опис того, як вона організована.

Програмне забезпечення ЦБ Greenstone - комплексна система для побудови та поширення колекцій цифрових бібліотек. Вона забезпечує спосіб організації та публікації інформації в Інтернеті (або на CD-ROM). Отже система Greenstone може вирішити задачу збереження та добування в електронному вигляді періодичних видань (ПВ) і задовольнити потреби наукових працівників в одержані інформації про періодичні видання, випуски періодичних видань або публікації.

4 КОРОТКИЙ ОГЛЯД GREENSTONE

Програмне забезпечення Greenstone використовується для створення і поширення в цифровому форматі бібліотечних колекцій.

ПЗ Greenstone розроблено на факультеті комп'ютерних наук університету Вайкато в Новій Зеландії в рамках проекту по створенню цифрових бібліотек. Разрботка проводилася

при сприянні ЮНЕСКО і неурядової організації Human info. В даний час Greenstone постійно допрацьовується. Програма вільно доступна на сайті *http://greenstone.org* і відповідає умовам GNU (General Public License).На сайті розташована дистрибутивна версія системи, яку можна вільно одержати, документація, FAQ (Frequently Asked Questions), а також надається технічна підтримка.

Існують дві версії системи – локальна і мережева. У локальній версії формування цифрової бібліотеки і доступ до неї формується в локальній мережі комп'ютерів. У мережевій версії усі функціональні можливості по створенню і використанню бібліотек надаються з використанням технології клієнт–сервер. Система працює на платформах Windows (95/98/NT/XP/2000) і Unix з використанням стандартних Web-серверів.

В даний час Greenstone широко використовується в багатьох організаціях таких країн, як США, Канада, Німеччина, Великобританія, Нова Зеландія та інш. На згаданому вище сайті є посилання на більш, ніж 20 колекцій цифрових бібліотек Greenstone. На сайті http://www.nzdl.org можна оглянути більш 50 колекцій цифрових бібліотек, створених при сприянні розроблювачів системи. Показові колекції включають статті з газет, технічні документи, художні книги, наукові журнали, фольклор, аудіо та відео інформацію.

4.1 Функції і можливості Greenstone

ПЗ Greenstone надає можливість користувачам [9]:

- створювати *колекції* електронних документів;
- детально визначати документи в залежності від метаданих;
- зберігати десятки Gb тексту та пов’язаних з ним зображень;
- здійснювати повнотекстовий пошук та пошук і перегляд документів по полям метаданих;
- документи, що додаються до колекції, а також їх метадані можуть мати різні формати;
- здійснювати обробку документів на будь-якій мові та підтримувати багатомовний інтерфейс користувача;
- організовувати та опубліковувати інформацію в Інтернеті або на компакт-дисках;
- використовувати стандартні та нестандартні метадані для опису змісту документів.

Далі, розглядаючи ПЗ Greenstone, зупинимось на деяких, на наш погляд, головних, моментах.

4.2 Колекції

Типова цифрова бібліотека, що створена за допомогою Greenstone, містить у собі безліч колекцій, організованих поодинці, хоча вони мають багато подібностей одна з іншою. Легко підтримувані, ці колекції можуть бути доповнені і перебудовані автоматично.

Колекції - сукупність документів різних форматів, що зібрані разом на підставі обумовлених користувачем критеріїв і до яких застосовується єдині механізми збереження, індексації, пошуку, перегляду і представлення.

Колекції можуть складатися із сотень тисяч і навіть мільйонів документів. Колекції можуть включати документи різної природи: текстові документи (статті, журнали, газети, звіти) а також аудіо і відео-документи. В колекції можна створювати підколекції, і в деяких випадках, колекції можна логічно об’єднувати.

Кожен текстовий документ може бути ієрархічно структурований у вигляді вкладених розділів (sections) (розділи, підрозділи, підпідрозділи і т.д.) Ієрархічна структура розділів відбиває змістовну структуру документа. Кожний з розділів, у свою чергу, складається з одного чи більш абзаців (paragraphs). Таким чином, змістовна структуризація звичайних документів на частини, глави, розділи і т.д. представляється в документах Greenstone у вигляді ієрархічної структури розділів Greenstone. Структура документа може використовуватися при формуванні пошукових індексів. Якщо вихідні документи не мають структури, то в колекції Greenstone вони можуть бути представлені у вигляді послідовності сторінок, що дозволяє переглядати документи по сторінках.

Greenstone надає наступні можливості маніпулювання з колекцією:

- Створити нову коллекцію з структурою, що співпадає з структурою існуючої колекції;
- Створити нову коллекцію з новою структурою;
- Додати нові документи в існуючу колекцію;
- Відредагувати структуру (конфігурацію) існуючої колекції;
- Повністю вилучити колекцію;
- Переписати колекцію на компакт-диск.

Вхідні інформаційні ресурси для побудови колекції можуть розташовуватися: на локальному комп'ютері, в локальній мережі та глобальній мережі з використанням протоколів HTTP та FTP.

Вхідні документи можуть мати різні формати. Для підтримки імпорту документів різних форматів розроблені так звані плагіни (спеціальні утиліти імпорту документів відповідних форматів). Плагіни написані мовою Perl. Усі вхідні документи, внесені в систему Greenstone, конвертуються у формат архіву Greenstone (Greenstone Archive Format). Це формат XML. Система Greenstone кожному документу автоматично привласнює унікальний ідентифікатор OID (Object Identifier).

У Greenstone структура кожної колекції визначається в процесі її створення. Це включає визначення формату використовуваних документів, їхнє виведення на екран, джерело метаданих, які предметні покажчики повинні бути включені, які повнотекстові індекси варто надати і як повинні відображатися результати пошуку. Після того, як колекція створена, в неї легко додати нові документи за умови, що вони того ж формату, що й існуючі документи, і мають схожі метадані. Кожна колекція містить файл конфігурації [5], у якому встановлюються параметри побудови і використання колекції.

На Рис.1 представлена домашня сторінка колекції журналу, яка містить загальну інформацію про колекцію (див. розділ 2.1).

Більшість колекцій можна відкрити шляхом *пошуку* і *перегляду*.

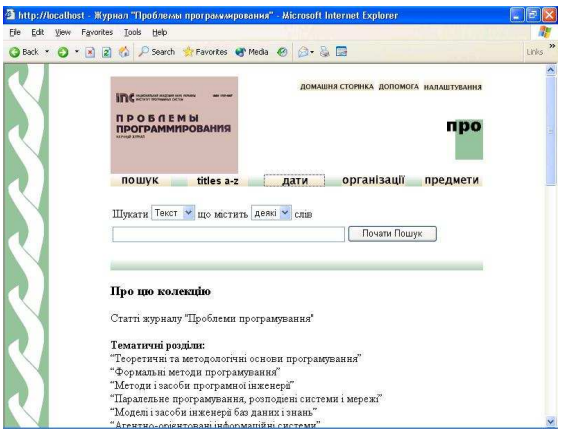

Рис. 1 Домашня сторінка колекції

4.3 Пошук

Користувач Greenstone може здійснювати повнотекстовий *пошук* по документах ("full text search"). Діапазон пошуку визначають індекси, які будуються на різних частинах документів. Повнотекстові індекси – це такі індекси, що дозволяють проводити пошук будь-яких слів по всьому тексту документа. За допомогою індексів можна шукати за окремим словом, набором слів або фраз. Результати пошуку представляються впорядкованими, за їх важливістю для користувача.

Колекції можуть мати індекси повних документів, індекси параграфів, індекси назв або авторів, по кожному з яких можна здійснювати пошук визначених слів або фраз. У багатьох колекціях такі дані, як автор, назва, дата, ключові слова, має кожен документ. Ця інформація називається метаданими. Багато колекцій мають повнотекстові індекси

визначених метаданих. Наприклад, у багатьох колекціях можна здійснювати пошук по індексу назв документів. Результати можуть бути упорядковані або відсортовані по елементам метаданих. Greenstone надає можливість робити пошук по декількох колекціях відразу з наступним об'єднанням результатів пошуку.

Рис. 2-3 показано екрани здійснення пошуку в колекції відповідно - формування запиту і результат.

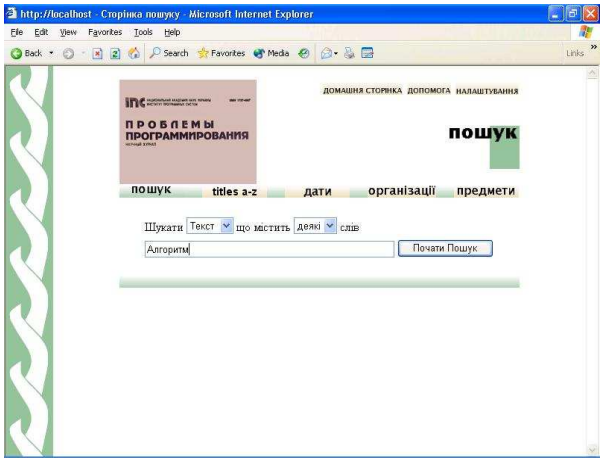

Рис. 2 Запит на пошук у колекції

Рис. 3 Результат виконання команди пошуку

4.4 Перегляд

Перегляд колекції містить у собі визначений перелік метаданих, що використовує користувач: перелік авторів, назв, дат, ієрархічні класифікаційні структури і т.д. Метадані є основою і початковим пунктом для здійснення перегляду. Різні колекції пропонують різні можливості для перегляду. Інтерфейси перегляду і пошуку створюються в процесі побудови колекції відповідно до інформації конфігурації колекції.

Для створення *структур перегляду* метаданих, використовується система *класифікаторів*. За допомогою їх можна створити індекси для перегляду такі як: алфавітні покажчики, дані і різноманітні ієрархічні структури. В Greenstone можна створювати нові структури для перегляду.

У Greenstone розроблено набір стандартних класифікаторів [5]. Усі класифікатори генерують ієрархічну структуру, що використовується для відображення індексу перегляду. На самому нижньому рівні цієї структури звичайно розташовуються документи, але можуть визначатися і розділи документів. Класифікатори можуть мати встановлену або довільну кількість рівнів ієрархії.

На Рис.4-10 демонструється процес перегляду колекції журналу.

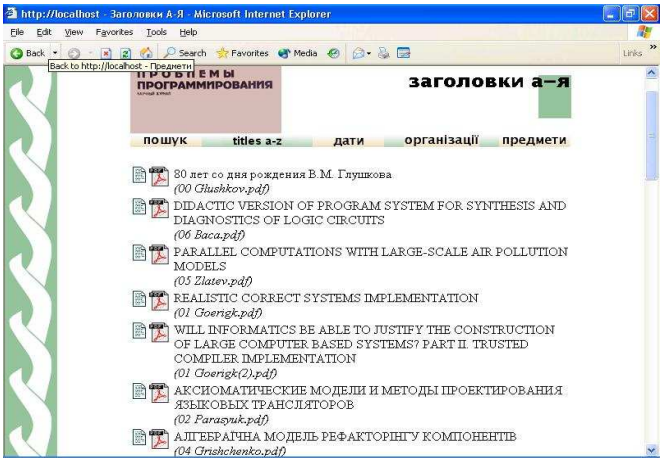

Рис. 4 Перегляд колекції по заголовкам

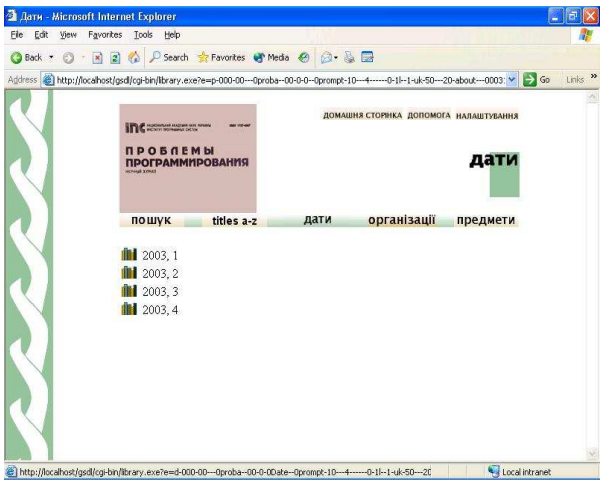

Рис. 5 Перегляд колекції по випускам

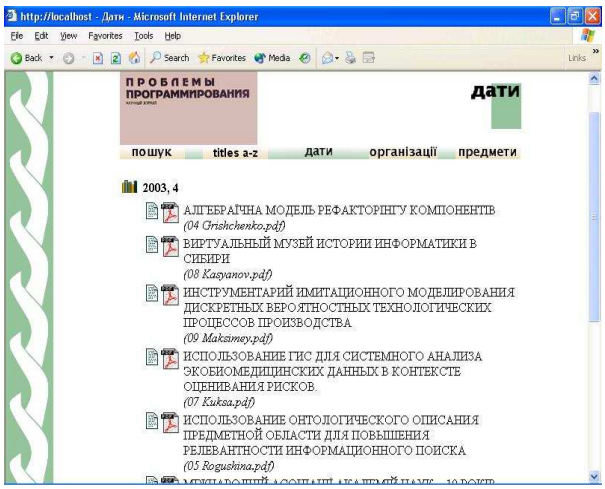

Рис. 6 Перегляд одного випуску

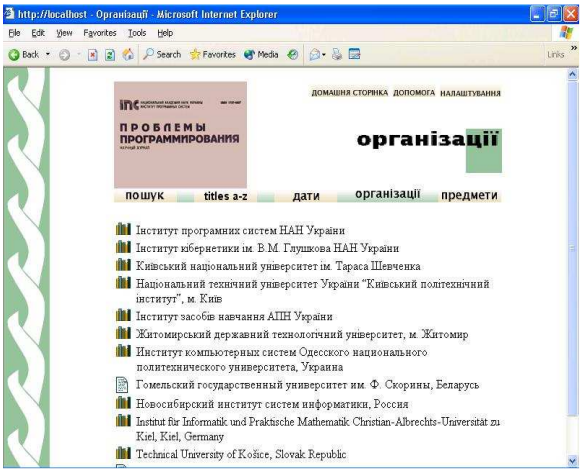

Рис. 7 Перегляд колекції по організаціям

Рис. 8 Перегляд колекції по організаціям

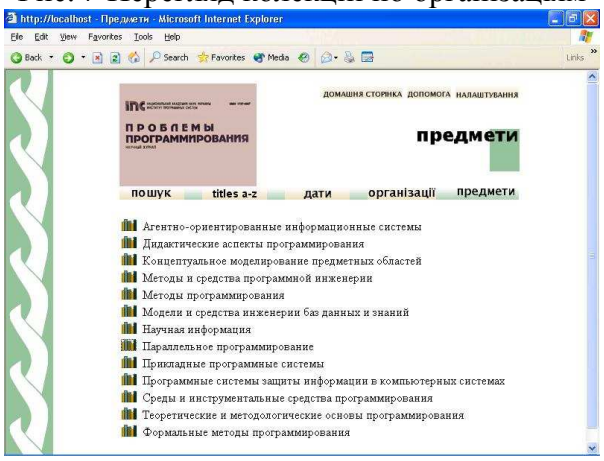

Рис. 9 Перегляд по предметній класифікації

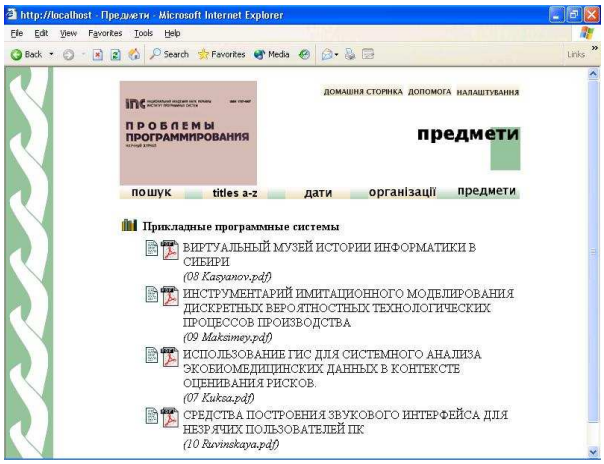

Рис. 10 Перегляд по предметній класифікації

4.5 Багатомовність

У системі використовується набір символів UNICODE. У зв'язку з цим і документи і зовнішній інтерфейс можуть бути представлені на різних мовах. У цьому розумінні система Greenstone є багатомовною. Це означає, що можна створювати багатомовний інтерфейс і в діалозі установки поточної конфігурації вибирати ту чи іншу мову.

Наприкінці огляду слід додати, що можливості, які забезпечуються Greenstone, і інтерфейс з якого бібліотечні користувачі звертаються до них, легко настроювати на різних рівнях. Користувачі можуть задавати формати документів (наприклад, HTML, Word, PDF,

Postscript, PowerPoint, Exel) або формат зображень (напрклад, TIFF, GIF, PNG, JPEG), що включаються до колекції. Крім того можна задавати набір доступних метаданих (наприклад, MARK, архіви OAI, BibTex, бази данних CDS/ISIS), а також які будуть забезпечуватися повнотекстові індекси (наприклад, всього тексту, можливо поділені мовою або іншими ознаками, і вибраними метаданими, наприклад заголовками або резюме) і структури перегляду (наприклад, список авторів, заголовків, дат, ієрархія предметної класифікації). Більш досвідчені користувачі можуть керувати представленням елементів колекції на екрані. Все це здійснюється за допомогою бібліотечного інтерфейсу Greenstone.

5 СТВОРЕННЯ КОЛЕКЦІЙ НАУКОВИХ ПЕРІОДИЧНИХ ВИДАНЬ ЗА ДОПОМОГОЮ GLI

Створення колекцій проводиться на основі інтерфейсу бібліотекаря GLI (Greenstone's Librarian Interface), інструмента для збору й обробки документів з наступним створенням колекцій цифрових бібліотек, що працюють під керуванням Greenstone. Він забезпечує доступ до функціональних можливостей програмного забезпечення бібліотеки Greenstone графічним шляхом. GLI дозволяє користувачам додавати документи і метадані до колекції, створювати нові колекцій та настроювати їх на зручний перегляд. Програмне забезпечення GLI поставляється та інсталюється разом із програмою Greenstone.

5.1 Роль і структура метаданих

Організація цифрової бібліотеки здебільш спирається на метадані — структуровану інформацію про ресурси (документи), що містить бібліотека. Метадані це щось на зразок традиційних карткових каталогів, "цеглини і цемент" бібліотек (не залежно від того, комп'ютерізовані вони чи ні) [10]. Якщо бібліотека "структурована", то нею можна осмислено керувати без обов'язкового розуміння її змісту. Наприклад, для колекцій документів, бібліографічна інформація про кожний документ була б метаданими для колекції. Метадані документа містять інформацію описового характеру, таку як дані про автора, заголовок, дату, ключові слова і т.д. Метадані можуть асоціюватися з документом у цілому чи з окремими розділами документа. Поняття "метаданні" не абсолютне, а відносне: воно тільки дійсно значимо в контексті і ясно дає зрозуміти, чим власне є дані.

Використання метаданих в якості будівельного матеріалу - дійсно визначальна характеристика цифрових бібліотек: це – те, що відрізняє її від інших колекцій інтерактивної інформації. Метадані дозволяють розташувати в бібліотеці новий матеріал і закріпити за існуючими структурами таким чином, що він відразу ж стає повноправним членом бібліотеки.

Метадані є основою для організації індексування документів, побудови класифікаторів і також можуть використовуватися при описі форматів представлення результатів пошуку або перегляду документів.

У Greenstone з кожною колекцією пов'язується один або кілька наборів метаданих. Існує цілий ряд стандартних наборів, наприклад Дублінське ядро [11]. GLI надає можливість визначати нові набори — як правило, додаючи кілька додаткових елементів до існуючого набору. Ще один важливий набір - набір здобутих метаданих, що містить інформацію, автоматично здобуту безпосередньо з документів. Наприклад, для HTML-файлів це тег заголовок, тег META, чи вбудовані метадані в DOC-файли, автор і заголовок.

Система зберігає набори метаданих, використовуючи різні простори імен (namespaces). Наприклад, документи можуть мати два атрибути Заголовок з набору метаданих Дублінське ядро (dc.Title) і з набору здобутих метаданих (ex.Title). Вони не обов'язково повинні мати те ж саме значення. Перелік описових елементів як для документа в цілому, так і його розділів не фіксований. Документ і його розділи можуть містити свої власні описові елементи (тобто їх склад може змінюватися від документа до документа або

від одного розділу документа до іншого). Здобуті метадані розташовуються безпосередньо в документах, а набори метаданих в окремих файлах в форматі XML. Елементи метаданих мають такий вигляд:

<Metadata name="Title">First and only chapter</Metadata>

Для того, щоб прискорити ручний ввід метаданих, GLI дозволяє зв'язати метадані як з папками документів так і з окремими документами. Це означає, що користувачі можуть використовувати перевагу групувань документів, щоб записати спільні для групи документів метадані за одну операцію. У GLI користувачі можуть організовувати ієрархію документів, перетягаючи елементи колекції і створюючи нові під-ієрархії, що можуть прискорити спільне призначення метаданих. Значення метаданих, що приписуються папці, успадковуються усіма файлами, вкладеними усередину цієї папки. Якщо користувач згодом вибирає файл і змінює успадковане значення метаданих, з'явиться попередження, що така дія скасує успадковане значення.

Метадані в Greenstone можуть бути простим текстовим рядком (наприклад, назва, автор, видавець). Або вони можуть бути ієрархічно структуровані. Крім того, вони багатозначні, тобто кожний елемент може мати більш ніж одне значення. Це використовується, наприклад, коли публікація має кілька авторів. GLI зберігає уже введені значення метаданих, і там, де потрібно їх ще раз використати, вони просто вибираються зі списків, усуваючи необхідність їх повторного введення.

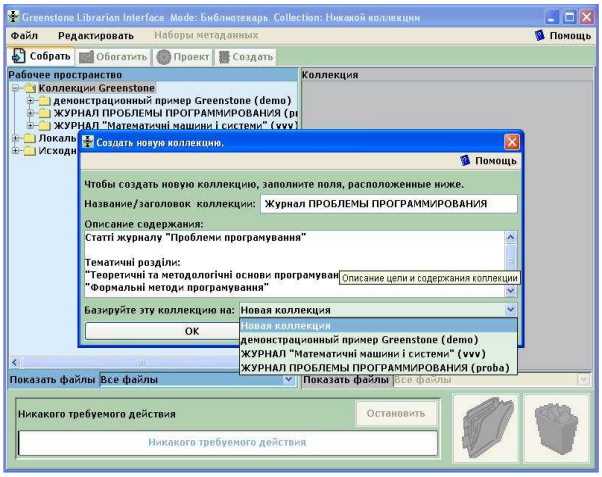

Рис.11. Запуск нової колекції

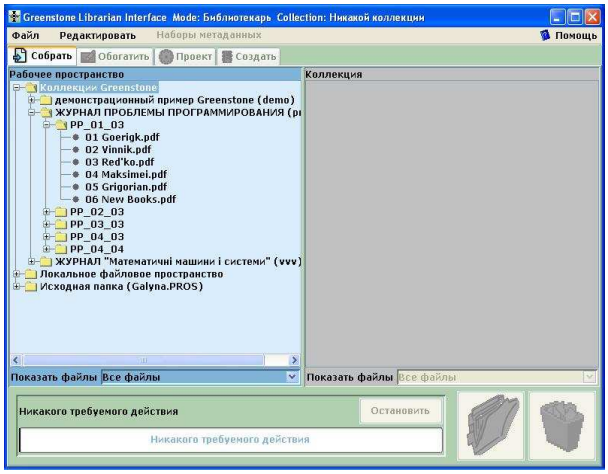

Рис. 12. Дослідження локального файлового простору

5.2 Робота з GLI

У GLI користувачі збирають набори документів, імпортують або прописують метадані і формують їх у колекції Greenstone. Це - інтерактивна незалежна від платформи Java-програма. Він тісно співпрацює із сервером, особливо в процесі проектування та побудови колекцій Greenstone. GLI включає різні Open Source-пакети для таких завдань як перегляд файлів, візуалізація HTML, відзеркалювання (mirroring) Web і ефективне табличне сортування.

GLI підтримує п'ять основних дій, що можуть чергуватися, але мають свій логічний порядок.

1. Внесення документів у колекцію, незалежно від того, хочемо ми заповнити нові колекції чи модернізувати існуючі. Будь-які документи, що імпортовані з існуючих колекцій, прибувають з приєднаними метаданими.
2. Збагачення документів, шляхом додавання до них метаданих. Документи можуть групуватися в папки.
3. Проектування колекції, визначаючи її зовнішній вигляд і засоби доступу, що будуть підтримуватися: це - повно-текстові індекси пошуку, структури перегляду, формат пунктів виведених документів на Web-сторінках, що генерує Greenstone і т.д.
4. Побудова колекції, з використанням Greenstone. Ця робота виконується системою, а користувачі забезпечуються індикатором виконання
5. Передача новоствореної колекцію бібліотечному серверу Greenstone для попереднього перегляду. Колекція автоматично інсталюється в персональну цифрову бібліотеку користувача , і відкривається Web- сторінка, показуючи домашню сторінку колекції (Рис. 1).

На Рис.11-20 представлено роботу GLI на прикладі побудови колекції журналу "Проблеми програмування". Це – екрани в різних точках протягом взаємодії користувача та інтерфейсу, виконуючи один прохід послідовно по усіх кроках, вказаних вище. Однак більш реалістична модель використання це - перехід назад і вперед по різних кроках по мірі виконання задачі.

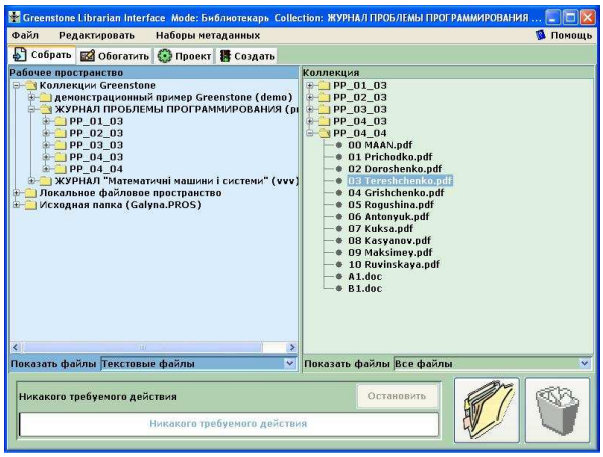

Рис. 13 Фільтрація файлових дерев

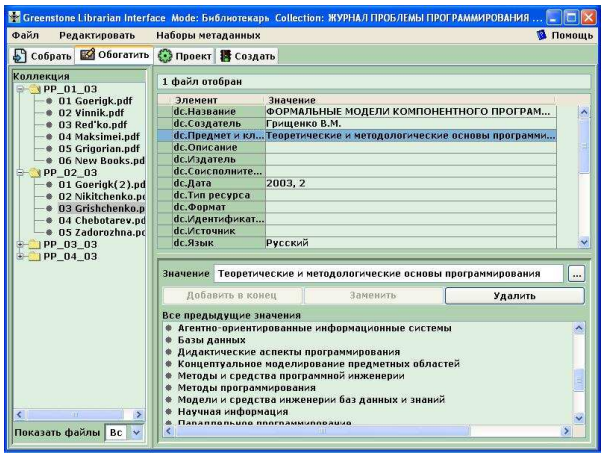

Рис. 14 Призначення метаданих документам колекції

5.2.1 Збір інформації

Спочатку, користувачі або відкривають існуючу колекцію або починають нову. На Рис. 11 ми бачимо процес запуску нової колекції. Тут користувач заповнює загальну інформацію про колекцію — назву і короткий опис змісту. В наведеному прикладі ми пов’язуємо з колекцією періодичне видання, тому на цьому етапі ми будемо використовувати деякі метадані рівня періодичного видання. Назва - коротка фраза, що використовується для ідентифікації колекції в ЦБ: наша колекція має назву *Журнал ПРОБЛЕМИ ПРОГРАМУВАННЯ*. Опис - твердження про принципи призначення колекції, і з'являється під заголовком *Про цю колекцію* на домашній сторінці колекції.

Далі користувач вирішує будувати колекцію на основі існуючої (Рис. 11), чи проектувати нову. У цьому прикладі ми будемо проектувати нову колекцію, і тепер для неї потрібно вибрати один або декілька наборів метаданих. Ми вибираємо Дублінське Ядро [3].
На цьому етапі надається можливість переглянути локальний файловий простір і існуючі колекції та накопичувати вибрані документи в новій колекції. Панель Зібрати (Собрать) розділена на дві частини, ліву для перегляду файлової структури і праву для організації документів у колекції.

Користувачі пересуваються по існуючій ієрархії файлової структури звичайним способом. Вони можуть вибирати файли чи каталоги і перетягати їх у колекцію праворуч. У такий спосіб можна перетягати цілі ієрархії файлів і файли можуть вибиратися звичайним способом. Користувачі також можуть переміщатися по колекції праворуч: встановлювати ієрархії файлів, перетягати елементи, створювати нові під-ієрархії, і в разі потреби, видаляти файли .

Інше джерело документів - Web. GLI має панель Дзеркало (Зеркало), Через неї, користувачі взаємодіють з Web-броузером і вибирають деякі сторінки, чи сайти для відзеркалювання (mirroring). Існує багато опцій: глибина відзеркалювання (mirroring depth); автоматичне завантаження впроваджених об'єктів, наприклад, зображень; тільки дзеркало з того ж самого сайту і т.д. Результуючі файли з'являються як нові папки верхнього рівня.

На Рис. 12 показане інтерактивне дерево файлів, що використовується для перегляду локальної файлової системи. На цьому етапі колекція праворуч порожня; користувач заповнює її, перетягуючи потрібні файли з лівої панелі.

Існуючі колекції представлені підкаталогом ліворуч і називаються “Колекції Greenstone”, його можна відкрити і переглянути так, як і будь-який інший каталог. Однак документи там відрізняються від звичайних файлів, тому що вони вже мають приєднані метадані, котрі GLI зберігає, коли їх додають до нової колекції. Тобто до нової колекції можуть бути додані документи які вже введені в інші колекції ЦБ, але при цьому можуть виникати конфлікти, тому що їх метадані, можливо, були призначені відповідно до різних наборів метаданих, і GLI допомагає користувачу розв’язати цю ситуацію.

Коли вибираються великі набори файлів і додаються до колекції, операція копіювання займе якийсь час — особливо, якщо повинні конвертуватися метадані. GLI указує на процес, показуючи, які файли копіюються і який відсоток файлів вже оброблено. Користувачі можуть переходити на наступний етап, у той час як все ще відбувається копіювання файлів.

Існують спеціальні механізми для роботи з великими наборами файлів. Наприклад, можна фільтрувати дерево файлів, щоб показати тільки визначений тип (Рис. 13).

5.2.2 Додавання метаданих до документів

На наступному кроці побудови колекцій слід збагатити документи, додаючи до них метадані. Це – те місце, де користувачі GLI проводять більшу частину свого часу: поліпшення колекцій здійснюється за допомогою вибора окремих документів і ручного додавання метаданих. Ми вже обговорили дві особливості GLI, що допомагають справитися з цим завданням:

- Документи, що копіюються протягом першого кроку, прибувають з будь-якими придатними приєднаними метаданими.
- Всякий раз, коли це можливо, метадані автоматично витягаються з документів.

Слід пам'ятати ще два важливих моменти, що прискорять ручне призначення метаданих:

- Значення метаданих може призначатися декільком документам відразу, на підставі їх спільного перебування в папці або через множинний вибір.
- Призначені раніше значення метаданих зберігаються і їх легко використовувати багаторазово.

Закладка *Збагатити* (Обогатить) видає інформаційну панель (Рис. 14). Ліворуч - дерево документів, що представляють колекцію, а праворуч можуть додаватися метадані до окремих документів чи груп документів. Часто користувачі хочуть бачити документ, якому вони приписують метадані і, якщо двічі клацнути на документі, то він буде відкритий відповідною програмою перегляду.

В нашому прикладі ми обмежуємося внесенням наступних атрибутів документа типу публікація, що покривається стандартним набором метаданих Дублінське ядро: Назва; Автор; Предмет та ключові слова; Дата; Організація; Мова.

На Рис. 14 користувач вибрав документ і надрукував “Теоретические основы программирования” як його метадані *dc.Предмет і ключові слова*. Опції додавання, заміни і видалення метаданих стануть активними в залежності від того, який вибір було зроблено. Значення раніше призначених метаданих для елемента *Предмет і ключові слова*, показуються у вікні, позначеному “Всі попередні значення”.

Користувачі можуть у будь-який час переглянути всі метадані, що були призначені колекції. Спливаюче вікно на Рис. 15 показує метадані у формі електронної таблиці. Для великих колекцій корисно мати можливість переглянути метадані, що пов'язані тільки з деякими типами документів, і якщо користувач визначив фільтр файлів як було згадано вище, при показі метаданих відображаються тільки обрані документи.

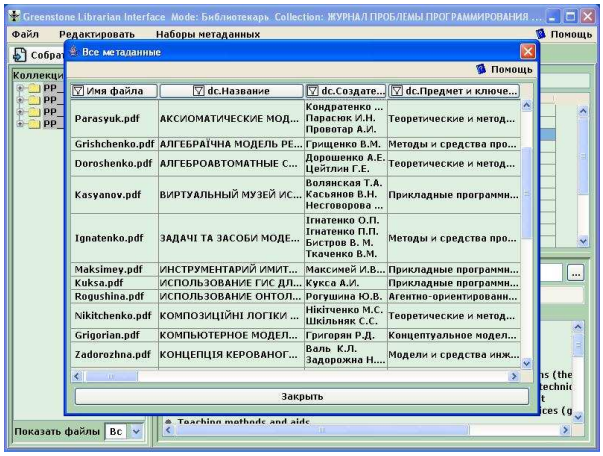

Рис. 15 Перегляд всіх метаданих, що приписані вибраним документам

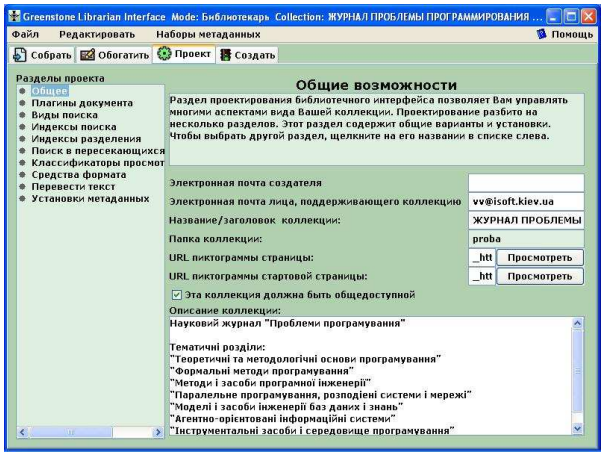

Рис. 16 Проектування колекції

5.2.3 Проектування колекцій

Коли до файлів додані потрібні метадані, далі слід вирішити у якому вигляді колекція Greenstone повинна бути представлена користувачам. Це настроюється на етапі проектування. Результат цього процесу реєструється в файлі конфігурації колекції [5].

Користувачі можуть рецензувати і редагувати метадані рівня колекції наприклад, заголовк, автор і загальнодоступність колекції. Вони мають можливість визначати, які повинні бути побудовані повно-текстові індекси. Вони можуть створювати під-колекції і мати побудовані для них індекси. Вони можуть додавати чи вилучати підтримку визначених мов інтерфейсу. Вони повинні вирішити, які будуть включені формати документа. Greenstone дозволяє обробляти різноманітні формати документів за допомогою плагінів (plug-ins). Кожному плагіну може знадобитися конфігурування, що визначається відповідними аргументами. Проектувальник колекції повинен буде визначити, які в Greenstone будуть створені структури перегляду, вони формуються модулями названими “класифікаторами”, що також мають різні аргументи. Також необхідно визначити форматування різних пунктів (елементів документа) у інтерфейсі користувача колекції. Для всіх цих елементів існують звичайно застосовувані значення за умовчанням.

На Рис.16-20 подається ілюстрація виконання проектування за допомогою панелі *Проект* . Вона має ряд окремих інтерактивних екранів, кожний з який зв'язаний з одним аспектом проектування колекції. Фактично, він служить графічним еквівалентом ручного процесу редагування файлу конфігурації колекції.

На Рис. 16 користувач натиснув на закладку *Проект* і переглядає загальну інформацію про колекцію, що була введена в момент створення нової колекції. Ліва панель

містить список різних властивостей колекції, що користувач може конфігурувати: Плагіни документа, Типи пошуку, Індекси пошуку, Індекси розбивки, Пошук по декількох колекціях, Перегляд класифікаторів, Елементи форматування, Переклад тексту і Набори метаданих. Наприклад, при натисканні кнопки *Плагіни документа* одержимо екран, показаний на Рис. 17, що дозволить додавати, видалити чи конфігурувати плагіни і змінювати порядок, у якому плагіни застосовуються до документів.

І плагіни, і класифікатори мають багато різних аргументів або "опцій", які може застосувати користувач. У діалоговому вікні на Рис. 18 показано список аргументів, що визначає користувач для плагінів. Оскільки Greenstone постійно розвивається і є Open Source–системою, по мірі того як розробники додають у систему нові можливості, число опцій збільшується. Щоб допомогти справитися з цим, Greenstone має сервісну програму "інформація плагінів", що вносить до списку опції, доступні для кожного плагіна, і GLI автоматично викликає цей список, щоб визначити, які опції показати.

На Рис. 19 користувач додає новий індекс повно-текстового пошуку до колекції, у даному прикладі, заснованому на двох елементах метаданих dc.Творець і dc.Назва. Щоб пошук проводився і по іншим колекціям всякий раз, коли вибирається команда пошуку по даній колекції, додається "пошук по кільком колекціям" і визначається список колекцій.

Для підтримки багатомовного інтерфейсу колекції в системі передбачено засіб для перегляду і введення перекладів текстових фрагментів редагуємої колекції (Рис. 20). Будь-які зміни з'являться у колекції відразу.

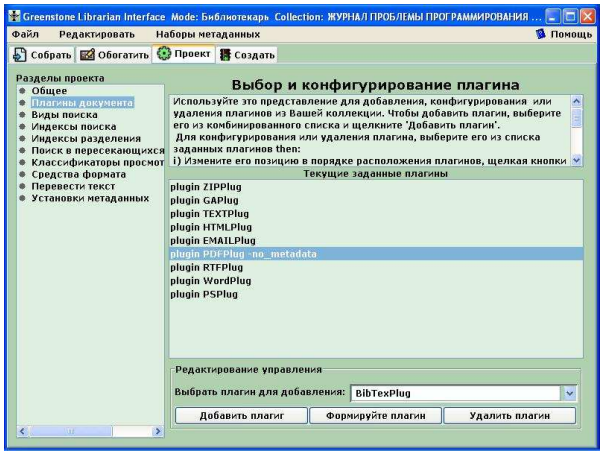

Рис. 17 Визначення використовуємих плагінів

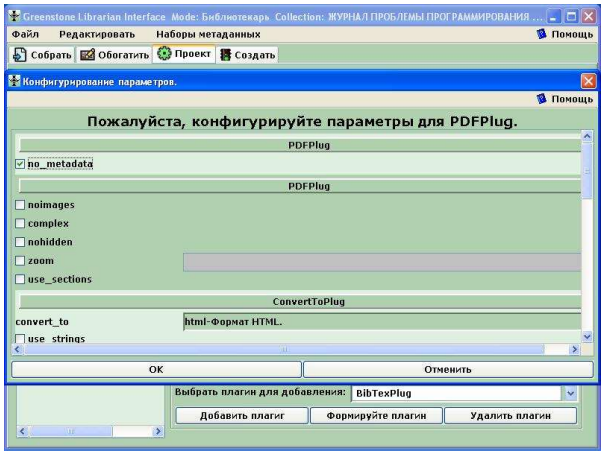

Рис. 18 Конфігурування аргументів плагіну

5.2.4 Побудова колекції

Наступний крок повинен створити колекцію, сформовану документами і призначеними метаданими. Основний тягар цієї роботи падає безпосередньо на Greenstone. Користувачі керують побудовою колекції за допомогою панелі Створити та групи опцій, що можуть застосовуватися протягом процесу створення (Імпорт, Побудова і Реєстрація Повідомлення). Під час побудови колекції Greenstone безупинно виводить інформацію пов'язану з процесом побудови.

Якщо побудова завершується успішно, користувачі мають можливість виконати перегляд створеної колекції (Рис.1-10).

Таким чином, ми описали, як можна створити і настроїти колекцію документів, використовуючи ПЗ Greenstone і інтерфейс бібліотекаря GLI. Ми показали, як можна реалізувати трьохрівневу модель інформаційного ресурсу: наукове видання - випуск – публікація за допомогою ПЗ Greenstone. Наукове видання в нашій системі представлене колекцією й весь або частковий набір атрибутів наукового видання приписуються їй. Випуск для нашого прикладу представляється вхідною папкою колекції, де зібрані всі документи з

однаковою датою. І, нарешті, публікація це документ колекції й атрибути публікації приписуються документові.

Помітний недолік системи – *постійна природа індексних файлів*, що генеруються протягом процесу побудови, збільшує вартість модифікації колекції. Інший недолік – низький рівень користувальницької функціональності, що підтримується системою Greenstone під час виконання, хоча незначні зміни легко підтримуються, більше значні зміни *включають модифікацію й перекомпіляцію початкового коду*. В наступному розділі ми зупинимось на новому проекті Greenstone 3, що обіцяє подолати ці недоліки [12].

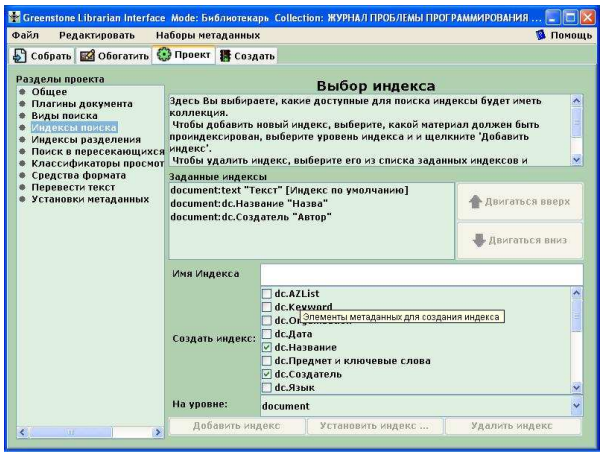

Рис.19 Додавання індексів повно-текстового пошуку

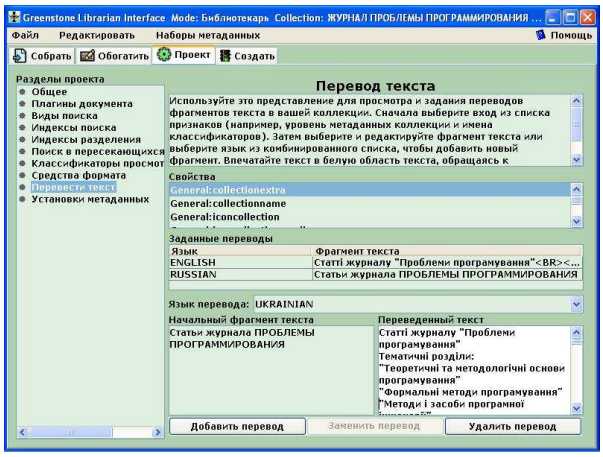

Рис. 20 Переклад фрагментів тексту колекції

6 ПРОЕКТ GREENSTONE3 ЯК ПЕРСПЕКТИВНИЙ СПОСІБ РЕАЛІЗАЦІЇ ЦИФРОВИХ БІБЛІОТЕК

Новий проект Greenstone3 націлений на поліпшення динамічної природи інструментарію, щодо організації змісту і забезпечення сервісами при одночасному зниженні накладних витрат, які несуть розробники колекцій для досягнення такої гнучкості. Проект заснований на сучасних стандартах, таких як XML (і зокрема, XSLT, мові для перетворення структури XML-документа), сучасних реалізаціях методологій: сполучених агентів (communicating agents), сучасних технологіях ПЗ, наприклад, протоколів SOAP (Simple Object Access Protocol), сучасної стратегії крос-платформної розробки (процес розробки, що забезпечує роботу в середовищі різнорідних процесорів і різних операційних систем), сучасних схем для модульного й динамічного відновлення ПЗ. Найбільш важливо, що новий проект враховує досвід попередніх версій Greenstone, проблеми й важкі запитання, з якими зштовхнулися реальні користувачі, реальні розроблювачі колекцій, реальні бібліотекарі.

Далі стисло зупинимось на деяких важливих характеристиках нової системи, що значно поліпшать її властивості.

- Нова розробка має зворотну сумісність, що має додаткові переваги і забезпечує розробників і користувачів легким способом переміщення.
- Для полегшення роботи є різні користувальницькі рівні різних категорій персоналу, що приймають участь у побудові ЦБ, наприклад, розробники змісту, редактори колекцій, проектувальники послідовності виконуваних дій, розробники ПЗ.
- Модульний принцип організації коду - основа будь-якого програмно-інженерного підходу. Цьому сприяє застосування існуючих технологій: систем БД, інструментів індексування й ПЗ візуалізації сторінки й використання стандартів.
- Інший шлях реалізації модульності - базування цифрової бібліотеки на наборі сервісів, у цьому випадку модульного принципу функціонування.

- Багата інфраструктура цифрової бібліотеки підтримується розподіленою архітектурою й відкритим протоколом для інтероперабельності. Запуск цифрової бібліотеки на одній машині - тривіальний випадок.
- Старі колекції можна представити в майбутніх версіях системи.
- Багато аспектів бібліотеки динамічні. Це покриває і динамічний зміст, коли можуть додаватися документи й метадані, що змінюються й видаляються, коли репозитарій перебуває в оперативному режимі й динамічну конфігурацію, що дозволяє впорядкування питань презентації й додавання сервісів під час виконання.
- ПЗ використовує систему інтегрованої документації і т.д.

ВИСНОВКИ

Активний розвиток робіт в багатьох країнах світу по створенню ЦБ і колекцій інформаційних ресурсів безумовно буде сприяти створенню ефективної інфраструктури для підтримки наукових досліджень та інших сфер діяльності і в нашій країні.

Актуальне значення має повноцінне використання можливостей перспективних інформаційних технологій для практичної реалізації ЦБ. Добре відома у світі система Greenstone інтегрує сучасні технології, тому її використання без сумніву буде корисно і науковцям, і розробникам і користувачам ЦБ.

В даній статті зроблено огляд сучасного ПЗ Greenstone, а також перспектив його розвитку. Було побудовано модель інформаційного ресурсу, що складається з трьох рівнів: періодичне видання, випуск, публікація. І на прикладі одного журналу описана процедура створення колекції ЦБ за допомогою інтерфейсу бібліотекаря, що входить до ПЗ Greenstone.

1. Digital Libraries. *E. A. Fox, H. Suleman, D. Madalli, L. Cassel* // Handbook of Internet Computing. ─ CRC Press. ─ 2003.
2. *Когаловский М.Р.* Научные коллекции информационных ресурсов в электронных библиотеках // Первая Всероссийская научная конференция Электронные библиотеки: перспективные методы и технологи. ─ С.-Перербург, Россия. ─ 1999. ─ С.16-31.
3. *Witten, I.H., Bainbridge, D., Boddie, S.J.* Greenstone: open-source DL software // Communications of the ACM. ─ 2001. ─ 44, 5. ─ P.47-57.
4. Witten I.H., Boddie S.J. *Greenstone: User's Guide //* New Zealand Digital Library Project, New Zealand. ─ 2003. (Инструкция для пользователя) ─ 50 p.
5. *Bainbridge, D., MacKay D. Greenstone: Developer's Guide //* New Zealand Digital Library Project, New Zealand. ─ 2003. (Руководство разработчика) ─ 113 p.
6. http://www.udc.org/ Universal Decimal Classification (UDC) Consortium.
7. http://www.acm.org/class/ The ACM Computing Classification System.
8. *Lynch C., Garcia-Molina H.* Interoperability, Scaling, and Digital Library Research Agenda, IITA Digital Libraries Workshop . ─ August,1995. ─ http://www.ccic.gov/pubs/iita-dlw
9. *Witten I.H., Bainbridge D., Boddie S.J.* Power to the people: End-user building of digital library collections // Proc. Joint Conference on Digital Libraries. ─ Roanoke, VA ─ June, 2000.─ P. 94-103.
10. *Witten I. H.* Creating and customizing digital library collections with the Greenstone Librarian Interface // Proc. International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society, DLKC'04. ─ Tsukuba, Ibaraki, Japan, 2004. ─ P. 97-104.
11. http://dublincore.org/usage/terms/dc/current elements/ Using Dublin Core - The Elements.
12. *Don K.J., Bainbridge D., Witten I. H.* The design of Greenstone 3: An agent based dynamic digital library. ─ http://www.sadl.uleth.ca/greenstone3/gs3design.pdf.

Про авторів

Резніченко Валерій Анатолійович,
кандидат технічних наук, сташий науковий співробітник
Проскудіна Галина Юріївна,
науковий співробітник
Овдій Ольга Михайлівна,
молодший науковий співробітник
Дорошенко Анатолій Юхимович,
доктор фіз.-мат. наук, професор, заст. директора по науковій роботі

Місце роботи авторів:
Інститут програмних систем НАН України,
просп. Академіка Глушкова, 40,
Київ-187, 03680, Україна
Тел. (044) 526 5139, 526 60 33, 526 1538
E-mail: reznich@isofts.kiev.ua, gupros@isofts.kiev.ua, olga_ov@kck.com.ua, dor@isofts.kiev.ua